Utrecht University

MBI Master Thesis

Thesis Report

Extracting Components from Privacy Statements with Text Mining

Author:
Lonneke Brakenhoff

First Supervisor:
Dr. Fabiano Dalpiaz
Second Supervisor:
Garm Lucassen

Supervisor:
Pieter Lamens

October 13, 2017

Contents

1 Introduction **1**
1.1 Research Questions . . . 2
1.2 Research Plan . . . 3

2 Literature Study **5**
2.1 Requirements Engineering . . . 5
2.1.1 RE Fundamentals . . . 5
2.1.2 RE Approaches for Privacy Purposes . . . 6
2.1.3 Definitions of Requirement . . . 8
2.2 Linguistic Science for Requirements Engineering . . . 9
2.2.1 Text Mining & Natural Language Processing . . . 10
2.2.2 Text Mining Techniques . . . 10
2.2.3 Text Readability . . . 16
2.2.4 Challenges of Natural Language in Privacy Statements . . . 17
2.3 Privacy Statements . . . 19
2.3.1 Approaches for Analyzing Privacy Requirements . . . 20
2.3.2 Privacy Ontologies . . . 21
2.3.3 Privacy Requirements . . . 22
2.3.4 Legislation & Privacy Statements . . . 30

3 Design Science **33**
3.1 Preliminary Conceptual Model . . . 33
3.2 Revised Conceptual Model . . . 35
3.2.1 Validating the Usefulness and Applicability of the Conceptual Model . . 35
3.2.2 Revised Model . . . 39

4 Method for Extracting Components from Privacy Statements **42**
4.1 Data Understanding . . . 43
4.2 Data Preparation . . . 45
4.2.1 Python Installation and Packages . . . 45
4.2.2 Prepare Training and Test Data . . . 47
4.3 Content Analysis . . . 47
4.3.1 Developing a Test Set . . . 48
4.4 Modeling the Algorithm . . . 51
4.4.1 Part-of-Speech Tagging . . . 51
4.4.2 Dependency Parsing . . . 51
4.4.3 Text Chunking . . . 54
4.5 Testing Scores . . . 55

4.6 Readability Scores . . . 56
4.7 The Final Algorithm . . . 56
4.7.1 Controller . . . 57
4.7.2 Modality . . . 58
4.7.3 Processing Activity . . . 58
4.7.4 Data Subject . . . 59
4.7.5 Personal Data . . . 60
4.7.6 Data Refinement & Restriction . . . 61
4.7.7 Purpose . . . 62

5 Evaluating the Results 63
5.1 Algorithm . . . 63
5.2 Readability . . . 66
5.3 Performance Measurements vs. Readability Indexes . . . 66

6 Conclusion 69

7 Discussion 72
7.1 Further Research . . . 73

Appendix A Used Literature/Data 80

Appendix B Framework Validation 81
B.1 Interview Protocol . . . 81
B.2 Interviews . . . 83

Appendix C Method 89
C.1 Python Anaconda Environment . . . 89
C.2 Privacy Statements . . . 91
C.2.1 Original Statements . . . 91
C.2.2 Preprocessed Statements . . . 110
C.3 Content Analysis Protocol . . . 130
C.4 Text Chunking . . . 132

Appendix D Jupyter Notebooks 133
D.1 Running the Jupyter Notebooks . . . 133
D.2 Preparing the Data Sets . . . 135
D.3 Modeling the Algorithm . . . 140
D.4 Calculating the Observers Agreement with Cohen's Kappa . . . 169
D.5 Calculate Readability Scores . . . 176

Abstract

The effect of the increasing awareness about privacy and new emerging legislation results in an inviting and agitated field of research. This project focuses on an automated approach for analyzing privacy statements. More specifically, text mining is used to first pre-process privacy statements into privacy requirements, followed by extracting components from those privacy requirements. The components, which are aligned with the definitions from the new General Data Protection Regulation (GDPR) that will be enforced 25 May 2018, are identified by a literature study and extracted with text mining techniques. Dependency parsing and text chunking are found to be the best combination of text mining techniques to achieve the goal of this project to extract components from privacy statements.

Keywords. Privacy Requirement Engineering, Requirements Engineering, Linguistic Science, Privacy Statements, Text Mining, CRISP-DM, GDPR.

Chapter 1

Introduction

Privacy is a complex issue that covers data protection and privacy violation. According to the AICPA[1], “the rights and obligations of individuals and organizations with respect to the collection, use, retention, disclosure, and disposal of personal information” (AICPA, 2017). Due to the advances in technology, such as new devices and Web applications, data is collected in various ways (da Silva et al., 2016). Organizations and software product developers, among others, need to protect themselves and their clients from potential breaches of privacy.

A *privacy statement* can be considered as a means of ensuring a certain level of protection. A privacy statement, also referred to as a privacy policy, is a document created by an organization to inform the users of a Website or application about activities performed on the user’s personal data. It defines information and actions concerning data, including the purpose of the data collection, its use, the people or groups who have access to the data, and the period for which the data is stored (Antòn et al., 2002; Karjoth & Schunter, 2002; da Silva et al., 2016).

A privacy statement consists of *privacy requirements* which are requirements that are focused on privacy purposes. The enterprise, the developers, and possible third parties are expected to meet the privacy requirements stated in the privacy statement of their customers to ensure that their privacy is not violated. Organizations have their own privacy statement in which they state what customers or other involved parties can expect to be performed on their data when they agreed on the terms of the privacy statement. Consumers can state their privacy preferences in a statement that requires organizations to act accordingly (Berendt et al., 2005). For example personal settings on a social media website that indicate if the social media organization can use an email address for sending push notifications. A user’s personal settings on a social media account may, for example, indicate that the user’s personal data can be used for specific purposes. A user may indicate that its email address can be used for direct marketing or push notifications functionality. These settings form the personal preferences of the user.

da Silva et al. (2016) mention the need for tools to help organizations align their processes with the privacy statement. The challenges for these organizations is ensuring that the developers understand the privacy requirements described in the statements. Privacy statements can be inconsistent, complex, incorrect or incomplete since they are often written in natural language (da Silva et al., 2016; Karjoth & Schunter, 2002; Sommerville, 2011). Natural language is the interaction between humans. We will focus on the written form of natural language. Karjoth

[1] American Institute of Certified Public Accountants (AICPA) set rules and standards for the Certified Public Accountant (CPA) profession in the US.

& Schunter (2002) emphasize that a clear language for engineering privacy requirements is mandatory to cover all the specifics of the privacy statement. The privacy statements should be concise and coherent and should be written to avoid ambiguous and vague terminology (Reidenberg et al., 2016).

To help in making privacy statements more clear, the main requirements of a policy should be highlighted and extracted from potentially large amounts of text. The goal of this project is to analyze privacy statements of organizations. The analysis will map the behavior of the enterprise into requirements that explain its practices regarding data processing. From a company perspective, a representation of their privacy statement(s) can help optimizing development processes.

1.1 Research Questions

This project is made up of a number of research questions. One of the first topics concerns the quality or structure of a good privacy requirement and is covered in the following question:

Question 1 *What elements does a good privacy requirement consist of?*

A number of steps are needed to answer this first question (1). Starting with modeling the elements that compose a requirement and determining the relation between these elements. A 'good' privacy requirement is defined based on literature, and should be correct, complete, and consistent (da Silva et al., 2016; Pohl, 1994; Berry et al., 2003). This project mainly focuses on one of these three elements: completeness. This project measures completeness on the basis of the presence of required and optional elements. An element is a word or set of words that fulfill a specific meaning in a requirement. For example, Karjoth & Schunter (2002) state that a statement or privacy requirement consists of the following elements: data has a user or owner (principal), the data itself (e.g. personal data), the purpose of the use, the actions taken with the data, the conditions that need to be satisfied, and obligations regarding the use of the data. Next, the elements need to be modeled into a framework. A distinction has to be made between the different elements and their purpose. From now on we refer to the artifacts, relationships, and dependencies as the *components* of the privacy statement.

When we know which elements comprise a good privacy statement and how a privacy statement is modeled. The following challenge is the process of extracting those elements from the text and characterizing them as the right component. This part in particular focuses on the components of one type of requirement in a policy statement: a privacy requirement. Several text mining techniques will be used to process large amounts of text. *Text mining* is described by Feldman & Sanger (2007) as an area of computer science that combines techniques from data mining, machine learning, natural language processing, information retrieval, and knowledge management to automate the process of large amounts of text. The following question is defined with respect the component extraction from privacy requirements:

Question 2 *What text mining techniques can be used to automatically identify the main components of a privacy requirement?*

Furthermore, the found text mining techniques and methods need to be tested on validity and effectiveness. The effectiveness of the text mining techniques will be evaluated on the number of correctly and wrongly identified components. The interpretation of the numbers will be defined during the development of the conceptual framework. The following questions for testing validity are formulated:

Question 3 *Is the method used for categorizing text elements from privacy policies effective for the used cases?*
Question 4 *To what extent is the method generalizable for other cases?*

Besides testing validity and effectiveness, the readability of a privacy policy is also important. The following questions is formulated regarding the readability of a privacy policy. As mentioned before a privacy statement informs users about processes are applied on their personal data. Therefore, it is important that they understand what is said in the policy statements.

Question 5 *What readability measurements can be used to measure readability of privacy statements?*
Question 6 *What combination of readability measurements is most effective on measuring the readability of privacy statements for the used cases?*

1.2 Research Plan

This project will focus on two pillars: (i) *requirements engineering*, and (ii) *linguistic science.* Those two pillars will result in a conceptual model that serves as a framework for modeling and, finally, mapping the components of privacy statements.

First, a literature study is required to search for existing methods for requirement modeling. Different existing modeling languages will be researched and compared to find the best representation for the privacy requirements. The focus will be on the concepts and relationships, instead of the graphical notation. Therewith, proposed requirement structures are compared to determine the components of a good requirement to answer Question 1.

The literature study will be substantiated and followed by putting some found methods or techniques into practice. This practice will imply the start of experimenting with some techniques for processing the text of the privacy statements (see Question 2). We need a text analysis tool that is able to recognize the components of privacy requirements in the privacy statement. A tool that is able to analyze natural language (e.g., human written text) is needed to do that for us. *Natural Language Processing*(NLP), as part of Text Mining, will be used to process the text and to assign words or sets of words to categories that represent the components found in the literature study (Chowdhury, 2003). Hereafter, text mining will be used through out the report as includes NLP.

Before the components are extracted from the text some techniques for processing text are required (Hosseini et al., 2016). These techniques can be, for example, stemming, tokenization, or removing stop words from the text. Also, databases such as WordNet might be useful regarding synonyms (e.g. words that are written differently but have the same meaning) (Fellbaum, 1998).

Text mining will be done in Python with existing packages, such as NLTK and spaCy. With text mining the annotation of words and sentences are analyzed. A (1) process deliverable diagram, or process data diagram, (PDD), (2) a Python implementation, (3) an evaluation, and (4) an interpretation of the results are the deliverables of the project. The deliverables together will form an algorithm that is able to label text with the right categories (components). The PPD is a modeling technique for both processes and its deliverables (Weerd et al., 2005). The second deliverable, the Python implementation, is documented in a Jupyter Notebook that includes an implementation of the algorithm in Python. The notebook will serve as tool for analyzing preprocessed privacy statements (see Section 2.2.2). Additionally, the notebook makes it easy to repeat the research, as it gives an overview of the Python, the input, and output in a clear manner. An evaluation, the third deliverable, is developed to test the generality of the algorithm. This means that the algorithm, which is evolved through a training set, is tested on a different set that is created by three independent observers. Three experts in the field of privacy requirements engineering will be asked to develop a test set that can be used to test the algorithm. The fit and quality of the models output will be measured based on precision and recall scores (Hosseini et al., 2016; Arora et al., 2015). Lastly, the interpretation of the results will answer the research questions.

The rest of the report is structured as follows: the first chapter (Chapter 2) describes the two pillars followed by the translation to privacy statements. Section 2.1, pillar one, describes *requirement engineering* (RE) in general. This section is based on existing literature and will explain models and methods for RE. In the next section (2.2) pillar two is described; we review linguistic science as an umbrella term for text mining. Text mining is a method in both the field of computer science and linguistics (Falessi et al., 2013). Section 2.3 presents literature specifically focused on the content and context of privacy statements. The found literature is combined in Chapter 3.1, design science, where a conceptual model is constructed as a starting point for the implementation of the requirement analysis tool. Additionally, an interview with experts is conducted to refine the conceptual model in Section 3.2. The method is described in Chapter 4 according the *Cross Industry Standard Process for Data Mining*(CRISP-DM) framework. The following three (of the six) steps are explained: data understanding (4.1), data preparation (4.2), and modeling (4.4). The evaluation of the algorithm is described in Result (5) chapter. The results are concluded and discussed in Chapter 6 and Chapter 7.

Chapter 2

Literature Study

2.1 Requirements Engineering

Requirements engineering (RE) is the process of composing and defining requirements and is defined as “the principled application of proven methods and tools to describe the behavior and constraints of a proposed system” (Antòn et al., 2002). Also, RE should cover the purpose, the stakeholders’ needs, and the documentation of requirements (Nuseibeh & Easterbrook, 2000). In the next two sections some fundamentals (Section 2.1.1) and sources (Section 2.1.2) of RE are described to create a basic idea of the discipline.

2.1.1 RE Fundamentals

Requirement Engineering Processes. The process of RE is described in the literature in multiple ways. Sommerville (2011) and Pandey et al. (2010) are both describing a *Requirement Engineering Process* (REP) in four steps. Sommerville (2011) describes the following four steps: 1) *Feasibility study* that determines if changes regarding the current system are needed and if possible changes are affordable; 2) *Requirement elicitation and analysis*, here the potential new requirements are picked and further investigated; 3) *Requirements specification* is the process of translating the ‘raw’ requirement data in a product requirement (och Dag et al., 2005) that can be implemented; 4) *Requirement validation* is the step where the requirement is judged on realism, consistency, and completeness. The four steps described by Pandey et al. (2010) are very similar. However, the first step of Sommerville (2011) is only focused the influence of potential changes in the system where the process Pandey et al. (2010) is directly focused on selecting potential requirements. Additionally, Pandey et al. (2010) added a last step of requirement management and planning. This last step is a continuous process of maintaining the changes made during the REP (Pandey et al., 2010). An overview of both REPs is shown in Table 2.1.

Three Dimensions of RE. A more conceptual approach to RE is proposed by Pohl (1994). He models three dimensions involved in getting the *initial input* into the *desired output*. In order to achieve the desired output each dimension should evolve to an ideal situation. One of the dimensions is *specification* and is about the transformation of a vague understanding of the requirements to a complete system specification. A second dimension is the *representation* dimension and is about the way requirements and knowledge are documented. The degree of

Table 2.1: Overview REP approaches from Sommerville (2011) and Pandey et al. (2010).

Pandey et al. (2010)	**Sommerville (2011)**
	Feasibility Study
Requirement Elicitation and Development	Requirement Elicitation and Analysis
Documentation of Requirements	Requirement Specification
Requirements Verification and Validation	Requirement Validation
Requirement Management and Planning	

formality is important regarding system descriptions since an informal approach of documenting can result in varying interpretations and therefore misconceptions. A formal representation will reduce or prevent misunderstandings, due to the well defined semantics. Important to mention is that representation and specification are independent of each other. This means that it is possible that someone with a low understanding (specification) of the system is able to represent the (mis)understanding in a formal way. The same applies to the opposite. The last dimension in the framework of Pohl (1994) is *agreement.* Agreement is a process of going from the stakeholders personal view towards a 'common system specification'. This agreed-on document is the goal of RE. The three dimensions together with the in- and output are modeled in Figure 2.1.

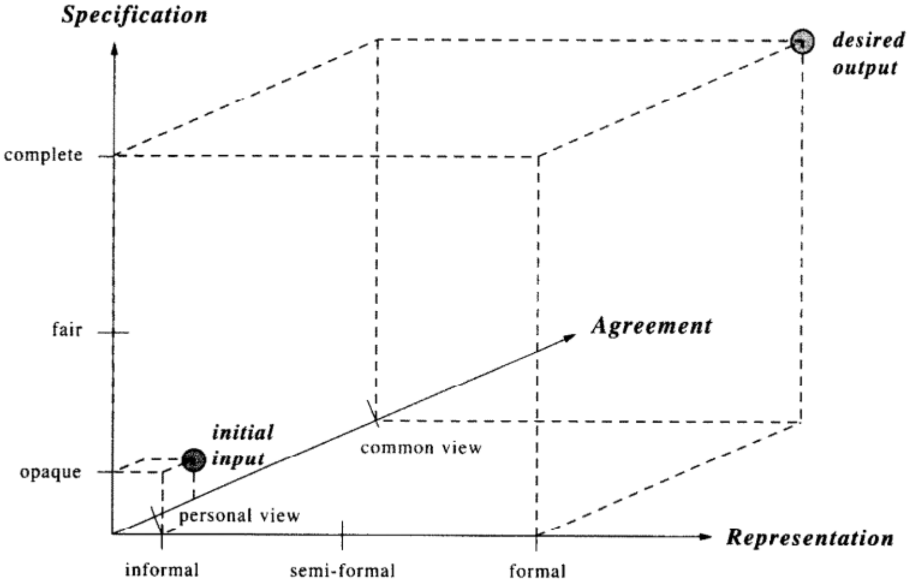


Figure 2.1: Three dimensions of RE. Taken from Pohl (1994).

The three dimensions of Pohl's framework are also applicable to engineering privacy requirements since they can be based on the wishes for many stakeholders, such as consumers, users, development teams, and the organization itself. Therefore, the degree of formality might change depending on the owner of the wish.

2.1.2 RE Approaches for Privacy Purposes

This section summarizes some of the relevant sources of privacy requirements. The first approach is based on system or organizational goals. The second approach is focused on privacy vulnerabilities. And lastly, the third source is legislation and regulations.

Goal-Driven Analysis or Goal Mining

Antòn et al. (2002) use *goals* as a starting point for requirement analysis. The focus on goal instead of specific requirements makes communication with stakeholders more easy since they are familiar with their goals. Since goals are considered to be the objectives for the behavior and output of the system it largely covers the context of the system and therefore the privacy requirements (Antòn et al., 2002; Van Lamsweerde, 2001). Van Lamsweerde (2001) also recognizes goals as the starting point for requirements. He describes a goal as "an objective the system under consideration should achieve" (Van Lamsweerde, 2001). Van Lamsweerde (2001) distinguishes multiple types of goals: step 1) *functional* goals are about what the system is expected to do. On the other hand, step 2) *nonfunctional* goals describe the quality of the system such as security and safety. Van Lamsweerde (2001) also states that goals can have attributes that characterizes the goal, such as their name or specification. Additionally, *priorities* can be attached to goals, so as to determine the 'optionality' of a goal and to help resolve *conflicts*. Conflicts will be further discussed in Section 2.3.3. At last, goals are connected to each other by links that determine the *goal structures* (Van Lamsweerde, 2001).

Anton et al. (2004) used *goal mining* to analyze financial policies that are covered by the Gramm-Leach-Bliley Act (GLBA), the financial privacy legislation in the US since 1999 (Anton et al., 2004). Goal mining is a process which aims at finding goals in data sources by applying goal-based requirement analysis methods (Anton et al., 2004). Anton et al. (2004) refers to those goals as *privacy protection goals*.

Vulnerability-Based RE

Privacy vulnerability is about potential threats to consumer privacy (Anton et al., 2004). There are two types of privacy vulnerabilities: i) *obvious* threats, which are known or soon to be known by the consumer, and ii) *insidious* threats. Insidious threats or privacy invasions are mostly activities that the consumer is not aware of. Activities related to personal data, such as: monitoring, storage, aggregation, and transfer, are often not visible to the consumer. The activities are further discussed in Section 2.3.3.

The difference between privacy vulnerabilities and privacy protection goals is that vulnerabilities are about preventing potential privacy threats where protection goals are about the desired protection of privacy rights (Anton et al., 2004).

Requirements Engineering & Law

Regulations and legislation form a big source of requirements (Otto & Antón, 2007). The combination of requirements engineering and law is shortened to *RELAW*. Requirements needs to be monitored to be complaint with the law and to prevent penalties (Otto & Antón, 2007; European Parliament and the Council of European Union, 2016). Additionally, developing a new system requires the right regulation to align the system requirements with (Siena et al., 2009; Otto & Antón, 2007).

Otto & Antón (2007) searched for a method for identification of the right regulations by classifying the regulations based on the right characteristics (meta-data). A requirement engineering framework was developed by Siena et al. (2009) to identify law-compliant system requirements. The framework can be used during the development or redesign of a system.

The research field of RELAW seems highly related to privacy statement as privacy statements cover both regulation and influences system requirements. Breaux et al. (2006) try to close the gap between law and RE by developing a process of which identifies natural language statements from legal texts. This project is not focused on regulations itself; however, the terminology is derived from the new European Law for data protection and the final deliverable can be used align privacy statements with requirements stated by the law.

2.1.3 Definitions of Requirement

A *requirement* is defined by the IEEE (1990) as follows: "A condition or capability that must be met or possessed by a system or system component to satisfy a contract, standard, specification, or other formally imposed documents". However, according to Sommerville (2011), the term requirement is not used consistently. The definitions vary from a high-level statement of what a system should do to a detailed description of the system's functions (Sommerville, 2011). A requirement is considered to have the following characteristics: unambiguous, complete, verifiable, consistent, modifiable, traceable, and usable. Throughout this research the definition of Robertson & Robertson (2012) will be used and is formulated as follows: "A requirement is something the product must do to support its owner's business, or a quality it must have to make it acceptable and attractive to the owner". Additionally to the definition of Robertson & Robertson (2012), a requirement is not only suitable for products, but can also be applicable to a service provided by an organization. Laplante (2013) describes the difference between goals and requirements as follows: "Goals are the high-level objectives of a business, organization, or system, but a requirement specifies how a goal should be accomplished by a proposed system".

User versus System Requirements. Sommerville (2011) distinguishes two types of requirements: *user requirements* and *system requirements*. User requirements are statements written in natural language that explains what the system offers its users. System requirements are, on the other hand, more detailed and are describing the functions, services and operational constraints. A user requirements is often translated to multiple system requirements (Sommerville, 2011).

We consider privacy requirements as user requirements. On the other hand, when defining privacy requirements based on goals of a system, we can also consider them as system requirements.

Functional versus Non-Functional Requirements. Requirements can also be distinguished as *functional requirements* and *non-functional requirements* according to Sommerville (2011), Robertson & Robertson (2012) and Pohl (1994). Van Lamsweerde (2001) describes functional and non-functional goals in the goal mining method for engineering system requirements (see Section 2.1.2) similar as Sommerville (2011) and Robertson & Robertson (2012) are describing functional and respectively non-functional requirements. First, a functional requirement is in short what the product or service must do. It is about the actions the user or consumer should be able to perform (Robertson & Robertson, 2012) and about the responses of the system on certain input (Sommerville, 2011). Secondly, non-functional requirements are requirements that focus on the constraints of the system (Sommerville, 2011) and explains the quality the system must have (Robertson & Robertson, 2012; Chung & do Prado Leite, 2009). They are called 'non-functional' requirements since they are not about the functions of the system, but about the appearance of the system and how well the tasks are performed (Robertson & Robertson, 2012).

Non-functional requirements can be classified in the following categories: *Look and Feel, Usability and Humanity, Performance, Operational, Maintainability, Security, Cultural and Political*, and *Legal*. In turn, a *Privacy Requirement* is classified in the subcategory of Security, named Privacy (Robertson & Robertson, 2012). The visual presentation of the categories and subcategories are shown in Figure 2.2.

However, the categorization of privacy requirements by Sommerville (2011) and Van Lamsweerde (2001) can be considered to be outdated. Nowadays, we can state that privacy requirements are both non-functional and functional requirements since they define very specific functions or features of systems (i.e. functional requirements). Chung & do Prado Leite (2009) conclude out of summarizing different divergent definitions that a "non-functional requirement is an attribute of or a constraint on a system". However, they stick to the following definition: "non-functional requirements constitute the justifications of design decisions and constrain the way in which the required functionality may be realized" (Chung & do Prado Leite, 2009). This definition encounters the role of quality requirements with respect to defining system functionalities. Additionally, IEC/ISO 25010 (2011) states that quality attributes benefits, among others, the identification of software and system requirements and identifying software and system design objectives (IEC/ISO 25010, 2011).

Privacy requirements will be further explained in Section 2.3.3.

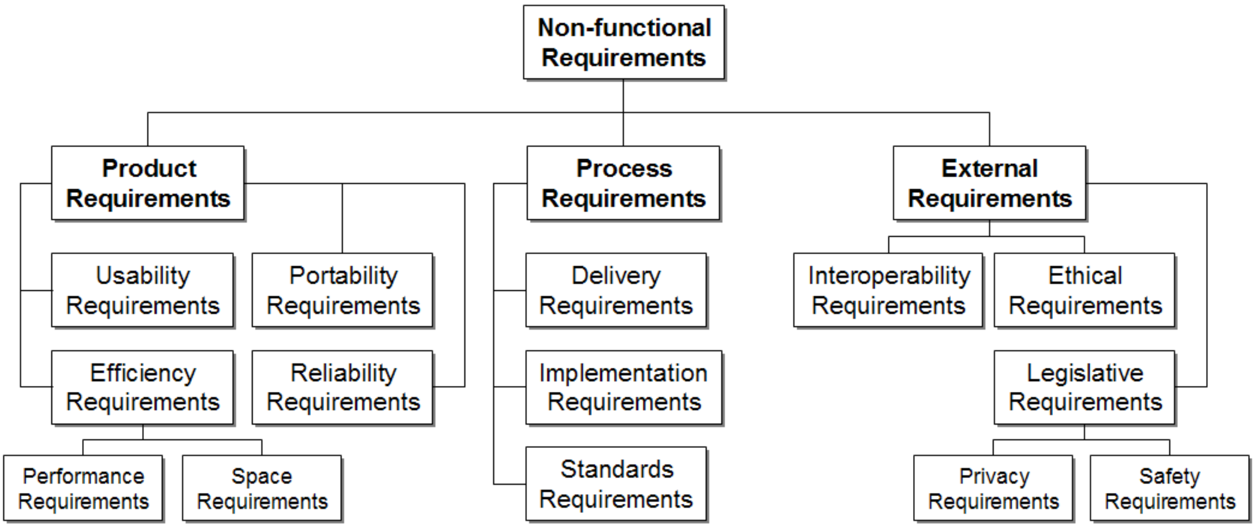


Figure 2.2: Non-functional requirements taxonomy, taken from Sommerville (2011).

2.2 Linguistic Science for Requirements Engineering

Manning & Schütze (1999) defined linguistic science as "a research field that is growing together with the increasing amount of data as a response to the demand for characterization and explanation of conversations, written text and other media".

Nuseibeh & Easterbrook (2000) describe the process of modeling and analyzing requirements as one of the core activities of RE. They also indicate that *linguistics* is important for RE since RE is about communication. Linguistics changed the way in which language is used in specifications. An example of in the involvement of linguistic science in RE is avoiding ambiguity in requirements and improving the understandability (Nuseibeh & Easterbrook, 2000). Another application of linguistics in RE is performed by och Dag et al. (2005) as they are using a linguistic-engineering approach for requirement management. They look at requirement similarity based on vocabulary.

This section describes the field of linguistic science applied to RE.

2.2.1 Text Mining & Natural Language Processing

Text mining, or *natural language processing* (NLP), is used by Falessi et al. (2013) to compare requirements based on their similarity. They compare text mining with *information retrieval* (IR). IR is described as the process of finding unstructured data (e.g. text in documents) that is useful to satisfy a demand for information. As said before, text mining is in both the field of computer science and linguistic science, it is contrary to IR focused on a problem within natural language (Falessi et al., 2013). In short, where IR focuses on a problem, text mining focuses on a solution (Falessi et al., 2013). Text mining is used to process large amounts of text to reduce costs and time (Massey et al., 2013).

Tan et al. (1999) distinguishes text mining into two components: text refinement and knowledge distillation. "*Text refining* transforms unstructured text documents into an intermediate form", and "*knowledge distillation* deducts patterns or knowledge from the intermediate form". In short, text mining consists of two steps to go from written natural language to new subtracted knowledge. The next section (2.2.2) discusses some preprocessing and knowledge subtracting techniques.

2.2.2 Text Mining Techniques

Falessi et al. (2013) looked at available NLP techniques to develop principles for measuring the effectiveness of NLP techniques by comparing them. They classified multiple techniques in the following possibilities: *algebraic models*, *term extraction*, and *weighting schema*. The algebraic models are focused on measuring semantic similarities between words. Term extraction is about how data is preprocessed before any other technique can be applied. Last, term weighting, assigns weights to terms based on how often they occur in the text documents.

The content of this section is based on the above-mentioned types of NLP techniques defined by Falessi et al. (2013). The techniques are modeled in Figure 2.3.

Term Extraction/Preprocessing

Before they started with comparing requirements, och Dag et al. (2005) preprocessed their text documents in four steps. In each step a preprocessing technique was used to prepare the

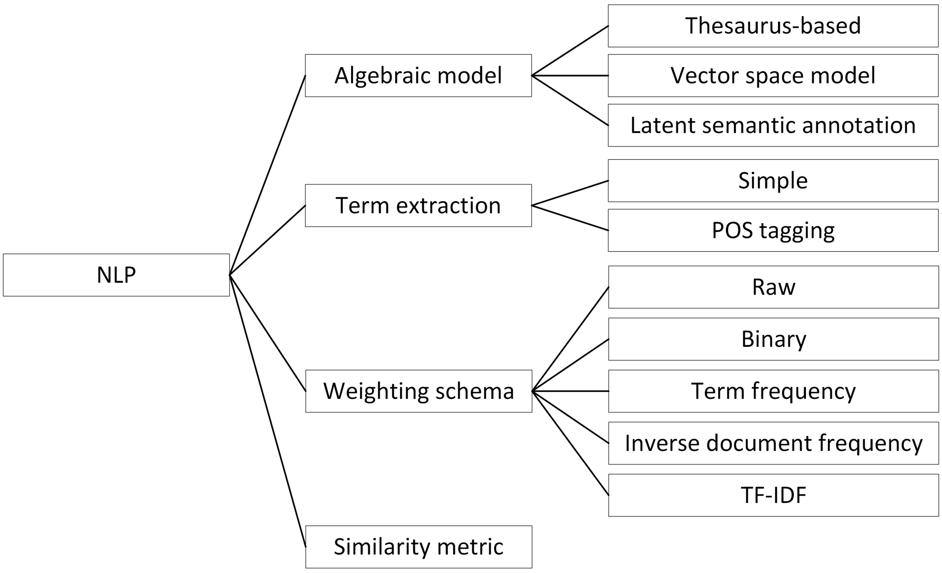


Figure 2.3: A visualization of text mining techniques, based on a figure from Falessi et al. (2013).

natural text into a set of text that is more likely to perform well when applying other (more complex) text mining techniques. The steps were as follows: (1) *flattening*, (2) *tokenization*, (3) *stemming*, and (4) *removing stop words.* Falessi et al. (2013) have two types of what they call *term extraction.* Tokenization and stop-word removal (step 2 and 4) are considered to be the main activities in a preprocessing activity that is called *simple* (Falessi et al., 2013). Additionally, Falessi et al. (2013) considers stemming (step 3), together with *term weighting*, to be one of the methods in *part-of-speech tagging* (POS tagging) which follows up simple. Simple and POS tagging are explained in further detail below.

Basic Text Processing Basic text processing, called *Simple* by Falessi et al. (2013), is the simplest form to process text and is almost always used in order to perform other text mining techniques (Falessi et al., 2013). Flattening is a form of text cleaning by removing the spelling mistakes. Tokenization is one the actions of preprocessing and coverts all words into a token with no capitals. Additionally, punctuation and brackets are removed which only leaves a set of words (tokens). After filtering everything but clean words stop words are removed (Falessi et al., 2013). Stop words are words like ‘and’, ‘a’, or ‘the’. The selection of stop words can be adjusted based on the context and content of the text documents.

Part-of-Speech Tagging As mentioned before, *part-of-speech* (POS) tagging, can be applied after tokenization and stop-word removal and tags each token based on whether it is a noun, verb, adjective, etc. (Falessi et al., 2013).

Listing 2.1 shows an requirement statement that is first tokenized and then cleaned from punctuation and capitals. Lastly, each remaining token is labeled with its word type.

Term Weighting

Term weighing is an umbrella term for multiple techniques of assigning a value to terms based on text analysis (Falessi et al., 2013). A short description is given for the five most occurring weights described by Falessi et al. (2013): raw frequency, binary, term frequency, inverse docu-

```
[(‘we’, ‘PRP’), (‘collect’, ‘VBP’), (‘the’, ‘DT’), (‘content’, ‘NN’),
    (‘and’, ‘CC’), (‘other’, ‘JJ’), (‘information’, ‘NN’), (‘you’,
    ‘PRP’), (‘provide’, ‘VBP’), (‘when’, ‘WRB’), (‘you’, ‘PRP’),
    (‘use’, ‘VBP’), (‘our’, ‘PRP\$’), (‘services’, ‘NNS’)]
```

Listing 2.1: The result of the process of finding a pattern for text chunking.

ment frequency, and combination of term frequency and inverse document frequency (TF-IDF). In Section 2.2.2 a application of the last three term weightings is exampled.

The *raw frequency* is the most simple form of term weighting whereby the weight is based on the number of times the term occurs in the text. *Binary* term weighting is just indicating if a word is present or not by assigning a 0 for not present, and a 1 for present. *Term frequency* (TF) also counts the word frequency, but here the term weighting is represented by the relative frequency of the words with respect to the document length. This approach covers the problem of having high word frequencies without saying much about the document. See Listing 2.4 for an example of a *term frequency vector*. The equation to calculate the *TF* is formulated as follows:

$$TF(x,y) = \frac{n(x,y)}{\sum_k n(k,y)} \tag{2.1}$$

Where $n(x,y)$ is the number of occurrences of a term T in a document D, this is then divided by the total sum of terms in document D.

Inverse document frequency (IDF) assigns a weight depending on the number of documents that contain the term (see Listing 2.8). The equation is written in Equation 2.2.

$$IDF(x) = log\frac{|D|}{|d : T \in d|} \tag{2.2}$$

$|D|$ is the total number of documents divided by the number of documents where term Tx appears.

At last, *TF-IDF* is a combination of both TF and IDF to take advantages of both benefits (see Listing 2.9 for an example). The TF-IDF is calculated by multiplying *TF* with *IDF* as follows:

$$TF\text{-}IDF(x,y) = TF(x,y)IDF(x) \tag{2.3}$$

Text Chunking

Arora et al. (2015) use text mining to process requirements written in natural language. According to Arora et al. (2015), natural language is being sensitive for ambiguity and that no common known restrictions are applied when writing natural language. The goal of the research of Arora et al. (2015) is to create an automated and generalizable tool that is able to check text on the fit on a predefined template. The technique that is used is *text chunking*. Text chunking is a text mining technique that assigns labels to text segments, called *text chunks*. Commonly

```
Tree('S', [('we', 'PRP'), ('collect', 'VBP'), Tree('NP', [('the',
    'DT'), ('content', 'NN')]), ('and', 'CC'), Tree('NP', [('other',
    'JJ'), ('information', 'NN')]), ('you', 'PRP'), ('provide',
    'VBP'), ('when', 'WRB'), ('you', 'PRP'), ('use', 'VBP'), ('our',
    'PRP$'), ('services', 'NNS')])
```

Listing 2.2: An example of a text chunking example tree in raw code. The statement is gathered from Facebook (2016)

occurring text chunks are noun phrases (NPs) and verb phrases (VPs). A NP is a segment that can be the subject or object of a verb, and a VP is a segment that contains a verb (Arora et al., 2015). In Listing 2.2 and Figure 2.4 is respectively an example of a text chunking tree showed with its raw output and its visualization. The text chunking processes is based on statistical methods that create and label the text chunks based on algorithms, this means that applying text chunking on a random selected text might not be with the perfect result.

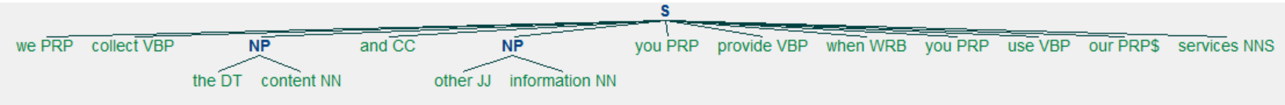


Figure 2.4: The visualization of a text chunking example tree generated from Listing 2.2 visualized by the Python function `.draw()`.

Algebraic Models

Algebraic models focus on the semantic similarity between terms (Falessi et al., 2013).

Vector Space Model The *vector space model* (VSM) is a method which assigns a vector as a representation for a text document in relation to a term (Falessi et al., 2013). In short, “the VSM is a standard way of representing text through the words they comprise” (och Dag et al., 2005). The difference with a *term frequency (TF) vector*, where each word is counted for each document, is that the vectors are not comparable since they are not weighed against the other documents. An example of a term frequency vector with input 2.3 is shown in Listing 2.4. For creating a VSM of the input shown in Listing 2.3 the following steps are needed:

1. A *document term matrix* is created by counting the number of occurrences per word for each document (2.4).
2. A *vocabulary* (2.5) is created as a set of distinct words occurring in all the documents.
3. Next, each document is compared to the vocabulary where the occurrence of each word is counted. The result is a *document term matrix* shown in Listing 2.6. Note that each document is represented as a vector that has the same length, which allows comparison between documents. This is the main advantages compared to the term frequency vector. We can now say that each document is in the same feature space.
4. In order to solve the problem of words being not equally informative within a document, *vector normalization* is applied (see Listing 2.7). In this example the L2Norm is applied, the calculations of which are not being discussed here in detail.

```
[‘We collect information you provide when you use our Services.‘,
 ‘We collect information about the purchase or transaction.‘,
 ‘We use your information to send you marketing communications.‘]
```

Listing 2.3: Input privacy requirements (e.g. documents). Requirements are gathered from Facebook (2016).

```
[(‘information‘, 1), (‘We‘, 1), (‘provide‘, 1), (‘when‘, 1), (‘use‘,
    1), (‘collect‘, 1), (‘Services.‘, 1), (‘our‘, 1), (‘you‘, 2)]
[(‘purchase‘, 1), (‘We‘, 1), (‘information‘, 1), (‘about‘, 1),
    (‘transaction.‘, 1), (‘collect‘, 1), (‘the‘, 1), (‘or‘, 1)]
[(‘information‘, 1), (‘We‘, 1), (‘marketing‘, 1), (‘use‘, 1),
    (‘send‘, 1), (‘to‘, 1), (‘you‘, 1), (‘communications.‘, 1),
    (‘your‘, 1)]
```

Listing 2.4: The term frequency vector of the input given in step 1. Statement is gathered from Facebook (2016).

5. An *inverse document frequency* (IDF) is calculated to compare words across all documents (2.8).
6. Finally, a *TF-IDF* is calculated by multiplying the TF by the IDF to get the weighed word vectors. The result is shown in Listing 2.9.

The examples explained are step by step manually performed in Python using `Counter`, `math`, and `numpy`. However, existing Python packages are developed to calculate the TF-IDF at once, such as: `sklearn`.

Thesaurus-based modeling is, like the VSM, also based on word similarity. A thesaurus is, in this setting, a dictionary that contains synonyms and is used to compare terms (Falessi et al., 2013). An example of a thesaurus is WordNet.

Latent Semantic Analysis *Latent semantic analysis* (LSA), or also called *latent semantic indexing* (LSI), is an expansion of the VSM. VSM did not take synonyms among words into account (Falessi et al., 2013). LSA measures the similarity of words by looking at words that occur together in the same document. The point of view is that if a word occurs more than once in a document it might not be chance (Falessi et al., 2013). With a given input for a LSA model, we expect a thesaurus as output where the similarity is expressed on a ratio scale (Falessi et al., 2013). So instead of working with a predefined thesaurus, the set of synonyms is domain specific. LSA applies *singular value decomposition* (SVD) on the TF-IDF vectors (explained in Section 2.2.2) to reduce the ‘noise‘ in the set of terms.

```
[‘information‘, ‘We‘, ‘your‘, ‘to‘, ‘purchase‘, ‘provide‘,
    ‘marketing‘, ‘send‘, ‘when‘, ‘use‘, ‘transaction.‘, ‘or‘,
    ‘collect‘, ‘about‘, ‘Services.‘, ‘our‘, ‘you‘, ‘communications.‘,
    ‘the‘]
```

Listing 2.5: An example of a vocabulary which is a distinct list of all the words occurring in all the analyzed documents. Statement is gathered from Facebook (2016)

```
[[1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 2, 0, 0],
 [1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1],
 [1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0]]
```

Listing 2.6: The document term matrix. Statement is gathered from Facebook (2016).

```
[[ 0.28867513  0.28867513  0.          0.          0.
    0.28867513
   0.          0.          0.28867513  0.28867513  0.          0.
   0.28867513  0.          0.28867513  0.28867513  0.57735027  0.
               0.        ]
 [ 0.35355339  0.35355339  0.          0.          0.35355339  0.
             0.
   0.          0.          0.          0.35355339  0.35355339
      0.35355339
   0.35355339  0.          0.          0.          0.
      0.35355339]
 [ 0.33333333  0.33333333  0.33333333  0.33333333  0.          0.
   0.33333333  0.33333333  0.          0.33333333  0.          0.
               0.
   0.          0.          0.          0.33333333  0.33333333  0.
             ]]
```

Listing 2.7: An example of the result of vector normalization. Statement is gathered from Facebook (2016).

```
[1.791759, 1.791759, 1.386294, 1.386294, 1.386294, 1.386294,
    1.386294, 1.386294, 1.386294, 1.609438, 1.386294, 1.386294,
    1.609438, 1.386294, 1.386294, 1.386294, 1.609438, 1.386294,
    1.386294]
```

Listing 2.8: An example of the result of vector creating an inverse document frequency. Statement is gathered from Facebook (2016).

```
[[ 0.32905528  0.32905528  0.          0.          0.
   0.25459192
0.          0.          0.25459192  0.29557206  0.          0.
0.29557206  0.          0.25459192  0.25459192  0.59114413  0.
            0.        ]
[ 0.41522937  0.41522937  0.          0.          0.3212653   0.
            0.
0.          0.          0.          0.3212653   0.3212653   0.37297746
0.3212653   0.          0.          0.          0.          0.3212653
   ]
[ 0.38904946  0.38904946  0.30100975  0.30100975  0.          0.
0.30100975  0.30100975  0.          0.3494615   0.          0.
            0.
0.          0.          0.          0.3494615   0.30100975  0.
   ]]
```

Listing 2.9: The result of the application of a VSM: TF-IDF weighed word vectors. Statement is gathered from Facebook (2016).

Latent Dirichlet Allocation (Topic Modeling)

Massey et al. (2013) used text mining to analyze policy documents. With their research they focused on three purposes: (i) valuating the readability of policy documents, (ii) categorizing requirements as either privacy protections or vulnerabilities, and (iii) supporting older work around the subject. They figured that policy readers have a hard time understanding the privacy statements in natural language. However, readers still prefer policies written in natural language compared to other forms when it comes to security. To address the three topics Massey et al. (2013) uses *topic modeling.* Topic modeling is defined by Massey et al. (2013) as “a text mining technique that can discover the themes in massive document collections”. The goals is to find topics and their relation towards the rest of the document (Massey et al., 2013). A specific technique of topic modeling is statistical approach *Latent Dirichlet Allocation* (LDA). LDA assumes that all documents have the same topic, but each document varies regarding how much it discusses the topic. Blei et al. (2003) describe LDA as “a generative probabilistic model for collections of discrete data such as text corpora”. LDA is based on a Bayesian model and the basic idea is that the documents are represented by topics. The topics are determined based on a term weighting technique that takes the relative word frequency into account (Blei et al., 2003).

2.2.3 Text Readability

Five metrics were used by Massey et al. (2013) to measure the readability of the policy documents. The (i) *Flesch Reading Ease* (FRE) which estimates the document on readability with a score between 0 and 100. A measure to understand the level of education needed to understand policies is the (ii) *Flesch Grade Level* (FGL). It measures the years of education needed to be able to understand the policy. The other techniques used by Massey et al. (2013) (FOG, SMOG, and the Automated Readability Index) focus on other aspects of readability and understandability of privacy statements.

Flesch Readability Ease Score The FRE is introduced by Rodolph Flesch in his book “The

Art of Readable Writing" (Talburt, 1986). The index is based on two averages: the (i) average syllables per word, and the (ii) average number of words per sentence.

Anton et al. (2004) used the FRE score (FRES) to evaluate the readability of financial policies. They describe the FRES as a score for more complex texts that is both used for school texts as legal documents. Over the years FRES is developed as commonly used and accepted measurement for the readability of texts (Anton et al., 2004).

The standard Flesch Index (F) is calculated as follows:

$$F = 206.835 - 1.015 \left(\frac{W}{N} \right) - 84.6 \left(\frac{L}{W} \right) \tag{2.4}$$

with W as the total number of words, N as the total number of sentences, and L as the total number of syllables (Talburt, 1986). The static numbers are fixed parameters (Talburt, 1986). The FGL is determined by defining ranges on the scale of 0 to 100. Each range on the scale is assigned to a certain level op education. Mapping the calculated FRES on the scale determines the education since the FRES is also between 0 and 100 (Talburt, 1986).

2.2.4 Challenges of Natural Language in Privacy Statements

In this section some challenges that go paired with the characteristics of natural language are discussed.

Ambiguity

"Ambiguity arises when a statement is incomplete and missing relevant information or when a word or phrase has more than one possible interpretation and the reader is uncertain about which interpretation the author intended“ (Reidenberg et al., 2016). Berry et al. (2003) described ambiguity as real-world phenomenon that occurs in many fields of work, such as writing, linguistics, philosophy, law, and software engineering. They mainly focus on filling the gap in literature about ambiguity in requirement engineering.

Berry et al. (2003) found different definitions from four different sources: dictionary, software engineering, linguistics, and legal definitions. A summary for each source is given below.

Dictionary Definitions Ambiguity is when something is understood in two or more possible senses or ways (Berry et al., 2003). It is also described as "uncertainty".

Software Engineering Definitions No single definition of ambiguity exists within the field of software engineering (SE) (Berry et al., 2003). On the one hand there is said that no *unambiguous* requirements exists. However, there are mature and useful specifications that are understood by experts in the field of SE. Ambiguity is described as the situation where an essential part is left out, undefined, or stated in a way that results in confusion or misunderstanding. It is caused by missing information and communication errors (Berry et al., 2003). On the other hand, unambiguous is when multiple readers (with similar domains) have the same or similar interpretation of the requirement (Berry et al., 2003).

Reidenberg et al. (2016) researched ambiguity in privacy statements. They defined four categories of vague terms. They categories and some examples are shown in Table 2.2.

Berry et al. (2003) distinguish two types of ambiguities. First the ambiguities that can be recognized by any reader, and second, SE ambiguities where only people with domain knowledge can recognize ambiguities.

Linguistic Definition The linguistics definition of ambiguity is split into multiple categories (Berry et al., 2003):

- Lexical ambiguity;
- Syntactic ambiguity;
- Semantic ambiguity;
- Vagueness and generality;
- Language error.

Lexical ambiguity is when a word has several meanings. *Syntactic ambiguity* is when a particular combination of words can have multiple grammatical structures with each a different meaning. When a sentence can be read in multiple ways with the same structure it is called *semantic ambiguity*. *Vagueness and generality* are related to ambiguity, but not the same (Berry et al., 2003). Vague words are words that fit the sentence properly on semantic and syntactic level, but the words itself are generalizing what is meant. For example the word `cousin`. Both a male or female can be meant i.e. it is generalizing gender, but the general meaning covers all the possibilities. Where the word `bank` can mean a whole different context, such as money bank and river bank. The last type of a linguistic ambiguity is the *language error*. A language error is a written mistake in the text which leads to wrong interpretations.

Legal Definition In the definitions given form the legal perspective there is less attention given regarding the source an the type of ambiguity, but more important is how to handle it. The Black's Legal Dictionary provides information on the treatment of an ambiguity (Berry et al., 2003).

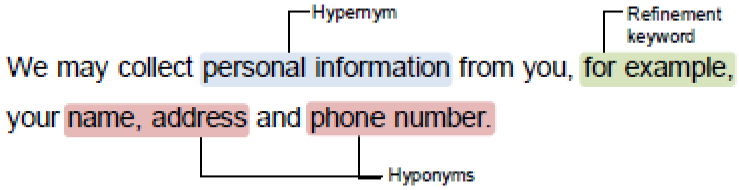
Table 2.2: Overview of categorized vague terms used in privacy statements. Taken from: Reidenberg et al. (2016).

Category	**Description**	**Examples**
Condition	Action(s) to be performed are dependent on a variable or unclear trigger.	Depending, necessary, appropriate, inappropriate, as needed, as applicable, otherwise reasonably, sometimes, from time to time.
Generalization	Action(s) or information types are vaguely abstracted with unclear conditions.	Generally, mostly, widely, general, commonly, usually, normally, typically, largely, often, primarily, among other things.
Modality	Vague likelihood of action(s) or ambiguous possibility of action or event.	May, might, can, could, would, likely, possible, possibly.
Numeric quantifier	Vague quantifier of action or information type.	Anyone, certain, everyone, numerous, some, most, few, much, many, various, including but not limited to.

Hyponymous & Hyperonymous

Anton et al. (2004) defines the goal refinement process where similar goals are simplified to one goal. In this process the information in the goals can be either abstracted to a higher level (e.g. ‘email-address‘ turns into ‘personal information‘), or refined (e.g. ‘personal information‘ becomes ‘email-address‘ and ‘name‘). Abstracting goals have similarities with role abstraction (Breaux et al., 2014) and when narrowing down a meaning the result is called a *hyponym* (Anton et al., 2004; Evans et al., 2017). The same applies to refining a goal or role. In this case a refined goals can be called a *hypernym* (Anton et al., 2004; Evans et al., 2017).

Evans et al. (2017) research information type *hyponymy* in privacy statements. One of the phases in their research is identifying hyponymy relationships which can be either a hyponym or a hypernym. In order to identify those relationships Tregex patterns are used which are based on regular expressions. With the patterns hypernym phrases, hyponym phrases, and keywords (i.e. word(s) that indicates a hyponymy relationship, such as ‘for example’) are identified (Evans et al., 2017). An example of a hyponymy relationship is shown in Figure 2.5 followed by an example of how the Tregex pattern is applied on a constituency parse tree in Figure 2.6.



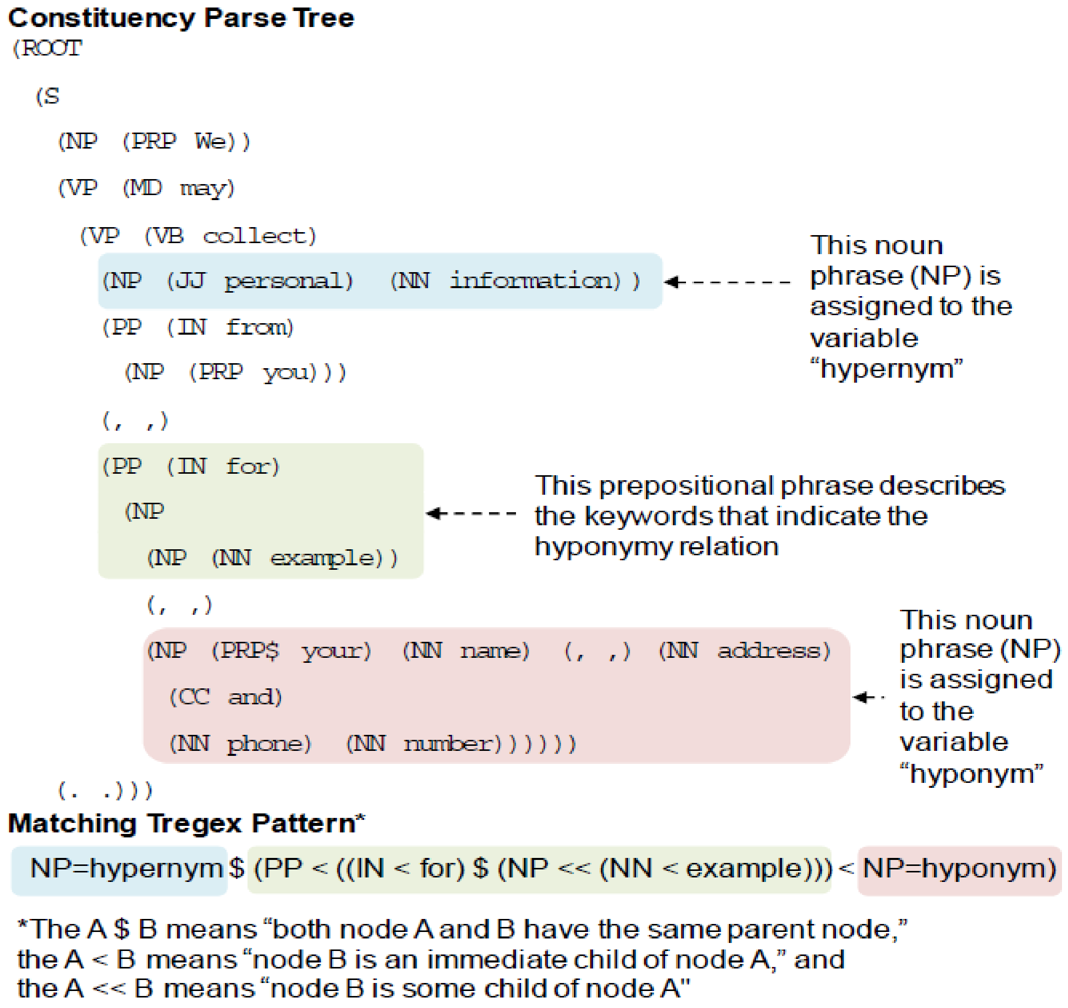
We may collect personal information from you, for example, your name, address and phone number.

Figure 2.5: An example of a requirement with a hyponymy relationship. Taken from: Evans et al. (2017).

2.3 Privacy Statements

In this section, we focus on the notion of *privacy statements.* A question prior to a more detailed explanation of PR is what the meaning of privacy is. According to Reidenberg (1994) developed *privacy* its-self over the last few years. Reidenberg (1994) mentions *privacy* as “maintaining the integrity of personal information and fairness to individuals about whom the data relates”. More specifically, privacy is mostly about “the collection, storage, use, and disclosure of personal information” (Reidenberg, 1994).

The owner of the privacy statement is in the Regulation (EU) no 2016/269 described as the *controller* (European Parliament and the Council of European Union, 2016). They define the controller as an actor “which, alone, or jointly with others, determines the purposes and means of the processing of personal data”. According to Gharib et al. (2016) the concept of transparency is important regarding data protection. The European Parliament and the Council of European Union (2016) state that as part of being compliant with the legislation around data protection a controller needs to be transparent by providing the data subjects with a privacy statement. Written privacy statements will be analyzed in this project as they can be freely accessed and are of great relevance to the context of the case study. Since organizational external privacy statements are used we adjust the terminology on the definitions given by European Parliament and the Council of European Union (2016).



Constituency Parse Tree

```
(ROOT
 (S
  (NP (PRP We))
  (VP (MD may)
   (VP (VB collect)
    (NP (JJ personal) (NN information))
    (PP (IN from)
     (NP (PRP you)))
    (, ,)
    (PP (IN for)
     (NP
      (NP (NN example))
      (, ,)
      (NP (PRP$ your) (NN name) (, ,) (NN address)
       (CC and)
       (NN phone) (NN number))))))
  (. .)))
```

This noun phrase (NP) is assigned to the variable “hypernym”
This prepositional phrase describes the keywords that indicate the hyponymy relation
This noun phrase (NP) is assigned to the variable “hyponym”

Matching Tregex Pattern*

NP=hypernym $ (PP < ((IN < for) $ (NP << (NN < example))) < NP=hyponym)

*The A $ B means “both node A and B have the same parent node,” the A < B means “node B is an immediate child of node A,” and the A << B means “node B is some child of node A"

Figure 2.6: An example the application of a Tregex pattern on a constituency parse tree of a requirement. Taken from: Evans et al. (2017).

This section is structured as follows: first an existing approach of extracting privacy requirements from a privacy statement (Section 2.3.1) and an existing privacy ontology (Section 2.3.2) are explained. In Section 2.3.3, the components found in the literature are summarized and explained. At last, Section 2.3.4, privacy related regarding legislation is covered.

2.3.1 Approaches for Analyzing Privacy Requirements

Breaux et al. (2014) specified a language to formalize privacy requirements with description logic, called Eddy. Eddy‘s purpose is to create an understandable language which will help developers, among others, to check consistency of privacy requirements (Breaux et al., 2014). The proposed method for extracting requirements from privacy statements consists of six steps. In step 1) each sentence of the statement is classified as a privacy statement, non-data requirement, or data requirement. In the following step, step 2) , an action is assigned to the requirement. The following actions are available according to Breaux et al. (2014): collect, use, retain, and transfer. These actions will be further explained in Section 2.3.3 and described in Table 2.3. Next, step 3) , each requirement is dissected into roles. This means that the statement is split into sections with a different role. Breaux et al. (2014) defined the following six roles: modality, subject, datum, purpose, sources, and target. The target role is only applicable when the earlier mentioned transfer action is performed. Again, these roles will be further explained in Section [ref:roles] and in Table [ref:table]. In step 4) , the given roles are further analyzed. The roles that are found in step 3 are generalized or refined. For example, 'information' is a generic form of 'a person's email address' (see Figure 2.8). In step 5) and step 6) , the requirements are respectively encode to the Eddy specification language and complied into a description logic such as OWL (see Figure A.1). The steps are visualized in Figure 2.7.

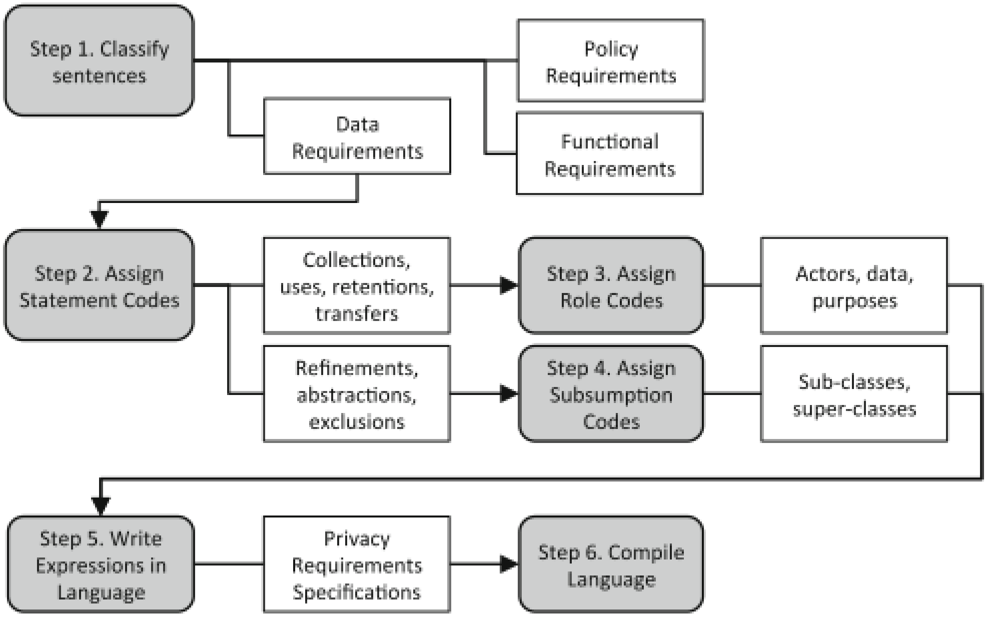

Step 1. Classify sentences
Policy Requirements
Data Requirements
Functional Requirements
Step 2. Assign Statement Codes
Collections, uses, retentions, transfers
Step 3. Assign Role Codes
Actors, data, purposes
Refinements, abstractions, exclusions
Step 4. Assign Subsumption Codes
Sub-classes, super-classes
Step 5. Write Expressions in Language
Privacy Requirements Specifications
Step 6. Compile Language

Figure 2.7: The six-step requirement extracting process, taken from Breaux et al. (2014).

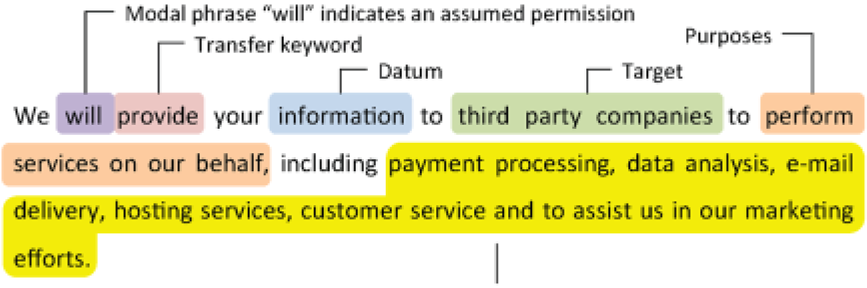

Modal phrase "will" indicates an assumed permission
Transfer keyword
Purposes
Datum
Target
We will provide your information to third party companies to perform services on our behalf, including payment processing, data analysis, e-mail delivery, hosting services, customer service and to assist us in our marketing efforts.

Figure 2.8: Example of dissecting a requirement statement with a permission about an information transfer in step 3 of the requirement extracting process. Taken from Breaux et al. (2014).

2.3.2 Privacy Ontologies

The proposed privacy ontology of Gharib et al. (2016) (see Figure 2.9) is used as a base framework. The framework will be adapted and possibly extended to fit it to the scope of this project.

Gharib et al. (2016) distinguishes four dimensions. The (i) *organizational dimension* describes entities that are interacting with the system and the dependencies, interactions, and relation-

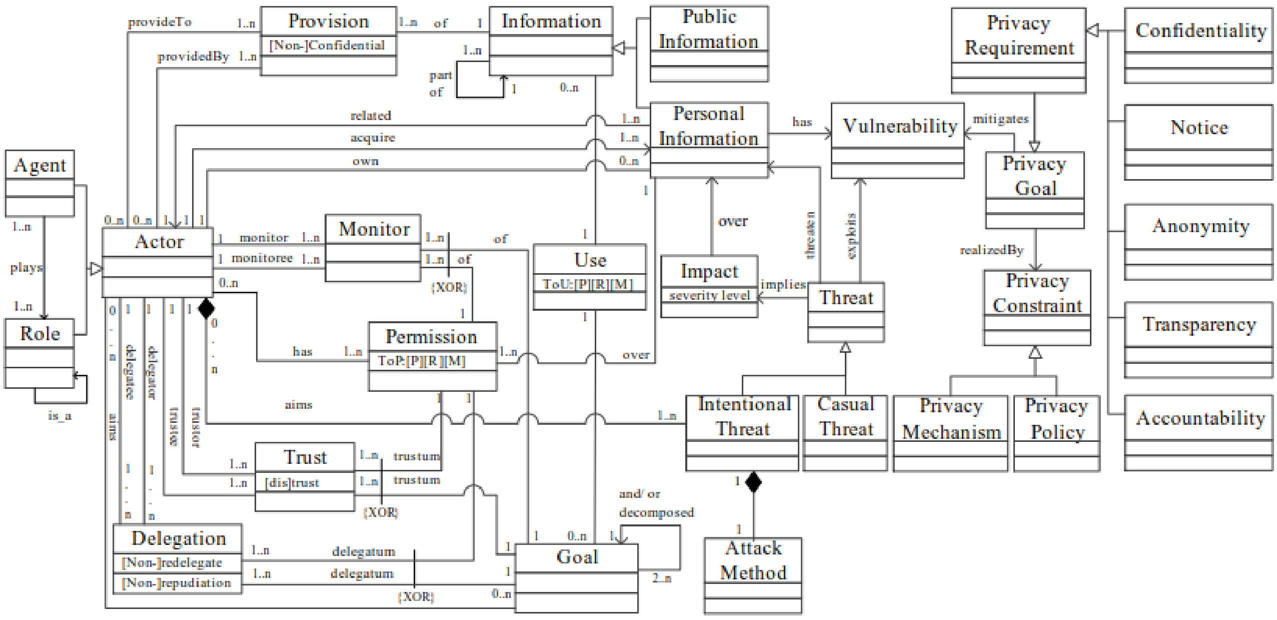


Figure 2.9: The proposed privacy ontology by Gharib et al. (2016). Taken from Gharib et al. (2016).

ships between them. (ii) *Risk dimension* is the second dimension defined by Gharib et al. (2016). A risk is defined as “an event that has a negative impact on the system” (Gharib et al., 2016) and it covers concepts as threats, vulnerabilities, and attacks. The third dimension is the (iii) *treatment dimension*, and consists of concepts that try to prevent the risk mentioned above. At last, the (iv) *privacy dimension* is about the privacy requirements that comprise the users‘ privacy needs regarding personal information, including confidentiality, notice, anonymity, transparency, and accountability.

Breaux et al. (2006, 2014) and Breaux & Antòn (2005) specified a requirement from the activity perspective. In 2005 and 2006, Breaux et al. distinguish six components that form an activity: subject, action, object, source, purpose, and target. The components and their relations are modeled in Figure 2.10.

2.3.3 Privacy Requirements

Privacy requirements (PRs) are requirements that focus on privacy purposes. As mentioned in the previous section (2.1.3), a PR is a nonfunctional requirement which means that they focus on the quality of the system instead of its functions (Sommerville, 2011). They are often used to describe the objectives and targets to be achieved by the system, or the so called goals (Antòn et al., 2002).

In order for a privacy requirement to be consistent, simple, and complete (da Silva et al., 2016; Karjoth & Schunter, 2002; Sommerville, 2011) we searched the literature for the components that are required or optional in a privacy requirement. The found components are described below.

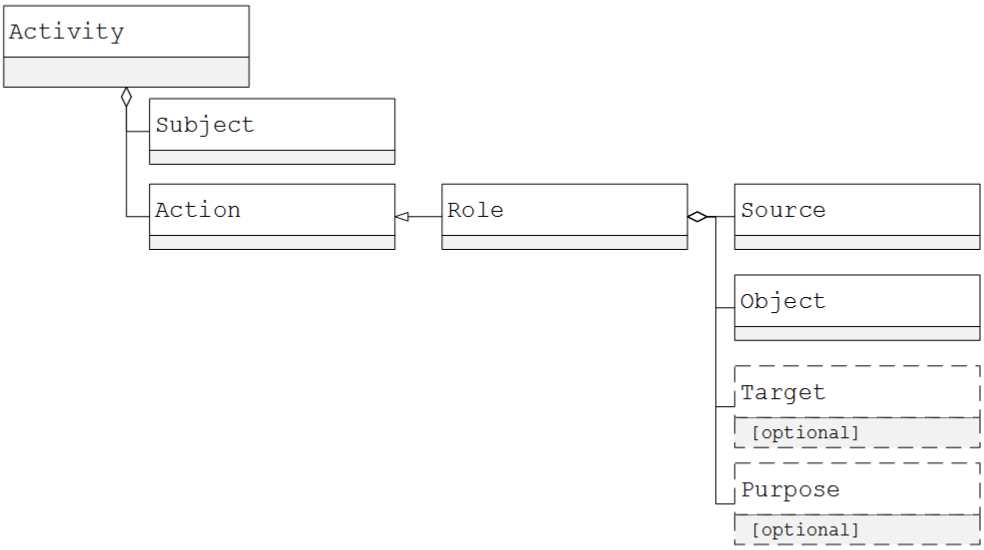


Figure 2.10: Integration of both models of Breaux et al. (2006, 2014).

Modalities in PRs

This section describes the possible forms of privacy requirements. We call these forms *modalities* as they characterize different types of requirements which each have a different meaning regarding the stakeholders based on a modal verb. Breaux et al. (2014) created a knowledge base with *description logic* that contained information about concepts and roles (referred to as terminology) and information about properties, objects, and individuals (summarized as assertions) (Breaux et al., 2014).

Singh (2013) introduces *norms* as the basis of a socio-technical system. Even so, socio-technical systems are based on norms. He defines a norm as a 'normal' interaction between stakeholders within a system that indicates their common intentions or goals. In order to align the expectations of all involved stakeholders the interactions (which are actions that involve or can influence other stakeholders) are important for *governance* (Singh, 2013). Governance describes how individual stakeholders manage themselves. The goal of norms is to form a basis for 'coherence', which means that it should result in stakeholders complying with the rules. Singh (2013) explains a model based on a participant in an Org (an organization that involves two or more roles). Each participant (or stakeholder) has assigned a set of elements that define their role. This set consist of three components. At first, *qualification*, a prerequisite or eligibility requirements that determines the participant's specified role. Secondly, a *privilege* that indicates the liberty applicable to the role. At last, a stakeholder acting in a specified role must behave or perform some actions according to that role. Singh (2013) calls this *liability*.

Following the theory of Singh (2013), a norm can be one of five defined types. For all five types an *antecedent* (condition), and a *consequent* (result or effect that follows) is involved. Firstly, a norm can be a (i) *commitment*, where the stakeholder (subject) is expected to be committed to the other stakeholder (object). A second type is (ii) *authorization*, where the subject authorizes or permits the object to do something. The opposite of authorization is (iii) *prohibition*. Here the subject forbids the object some action. A (iv) *sanction* is when the subject punishes the object by delivering the consequent after the antecedent is met. Lastly, (v) when the norm type is *power*, the object empowers the subject with the consequent when the antecedent holds.

Schaad & Moffett (2002) mentions that in order to delegate an obligation to an object, the subject that delegates the obligations needs the right to do so. Therefore, such a delegation should be preceded by an authorization that specifies rights among the activities regarding particular information.

Permission A *right* (or *permission* (Gharib et al., 2016)) is a “claim legitimately ascribed to a right bearer with respect to an implicit or explicit other, the counter-party“ (Breaux et al., 2006). In short, the stakeholder holds the right to perform some action (Gharib et al., 2016; Breaux et al., 2014). Actions can be reading, modifying, or writing information (Gharib et al., 2016). Textual indications in written natural language for rights can be found in the Appendix A.1.

Demsetz (1974) describes a transaction of data as a set of rights that is exchanged. He also states the rights, compared to the product or service that is exchanged in the transaction, is the real value (Demsetz, 1974).

Prohibition The opposite of a right is an *anti-right* (Breaux et al., 2006) or a *prohibition* (Breaux et al., 2014). A prohibition explicitly indicates an activity that is not allowed by a specific stakeholder (Breaux et al., 2006).

Obligations *Obligations* are described by Breaux et al. (2006) as a “duty bound to an obligated party that must be complied with“, this goes often with measures to guarantee the compliance. Referring to an obligation, if a party has a right it can be turned to saying that a counter-party has an implicit obligation (Breaux et al., 2006). In other words, if a user has the right to access their personal data, then the system (or system administrator) must provide this personal data (the obligation). Indications in written natural language for obligations can be found in the Appendix A.1.

Privacy Requirement Concepts

Where delegation is about prohibiting or granting rights for certain actions, provision is about providing data.

Delegation A right or obligation can be assigned to a stakeholder by the stakeholder who owns the data (Gharib et al., 2016). Breaux et al. (2006) call this a *delegation* and can be described as relation between a stakeholder and an expression. For example, a stakeholder can grant another stakeholder a permissions, which allows the permission receiver to perform an action it could not before.

Giorgini et al. (2005) and Schaad & Moffett (2002) distinguishes two types of delegations. At first, the *at-most delegation* which indicates that the subject wants the object to at most, but not mandatory, to fulfill an action (Giorgini et al., 2005). This type is referred to as the *delegation of permission*. The intent of the delegation is to grant the receiver (object) the right to perform an activity (Schaad & Moffett, 2002). The second delegation type is the *at-least delegation*. This type is also called *delegation of execution* (Giorgini et al., 2005). Here, the subject wants the object to at least fulfill an action, and therefore requires an action. According to Schaad & Moffett (2002) is demanding an object to perform a set of activities delegating obligations. Therefore, we will use *delegation of obligation* instead of delegation of execution to keep consistency in terms throughout the paper.

Provision *Provision* is the action of supplying something for use. In the case of information, it means that a stakeholder is providing another stakeholder with information (Gharib et al., 2016). Therewith, "provision is a relation between the provider and the receiver" (Gharib et al., 2016). Provision is directly applicable to the information, where a delegation gives access to a particular use. The difference between provision and delegation is the on which the action is performed. A provision performs an action on information or data, where a delegation is an action on a modality.

Controller

The entity that performs one or more actions described in a privacy statement is called an *actor* (Breaux & Antòn, 2005), a *stakeholder* (Breaux et al., 2006), *subject* (Breaux et al., 2014), or a *controller* (European Parliament and the Council of European Union, 2016). Gharib et al. (2016) divides the entity actor into two different concepts: a *role* and an *agent*. Each user or actor in a statement plays a certain role that characterizes the behavior of the actor (Gharib et al., 2016). An agent is an entity with a specific meaning that is accompanied with its role (Gharib et al., 2016). European Parliament and the Council of European Union (2016) also define a *processor* and describe the processor as follows: an actor "which processes personal data on behalf of the controller". The controller is described above (Chapter 2.3) as the privacy statement owner. The difference is that "a controller is the entity that determines the purposes, conditions and means of the processing of personal data, while the processor is an entity which processes personal data on behalf of the controller" (European Parliament and the Council of European Union, 2016). Despite of the distinction between the controller and the processor we will refer to the controller when talking about the actor that performs activities or processes regarding personal data.

Personal Data

A privacy policy statement always contains an *object*. The object is "the information on which the action is performed" (Breaux et al., 2014). This object is often a piece of information, such as: data, resource, asset, etc. (Gharib et al., 2016). Gharib et al. (2016) made a distinction between personal information and public information. With the first type the information can be directly related to an identifiable object, where as this is not possible with the second. Breaux et al. (2014) referred to the object with the term *datum*. In the scope of this project we will only refer to *personal data* as any information related to an actor (European Parliament and the Council of European Union, 2016).

Data Subject The *data subject* is the 'natural person' (i.e. actor) that is related to the personal data described above (European Parliament and the Council of European Union, 2016). Giorgini et al. (2005) calls this data *ownership*. Ownership is described by Giorgini et al. (2005) as a representation for the legitimate ownership of an actor of a service. Gharib et al. (2016) describes the concept of owning as a relation between information and its legal entity (owner). This ownership is described as a set of statements about the usages of this information.

Van Alstyne et al. (1995) researched how information sharing and database value could increase regarding costs and benefits by incentive principles. One of the important aspects included in the research: *data ownership*. According to Van Alstyne et al. (1995) lots of information system projects failed due to ignoring ownership. Additionally, data ownership will result in self-interest since owners of the a specific dataset will be more careful with the information than the actors who are not the owner (Van Alstyne et al., 1995). Important is the difference between usages rights and ownership (Van Alstyne et al., 1995). Usage rights (given privileges) indicates the ability to perform an action with the data object, where ownership is about controlling the rights regarding the data. Van Alstyne et al. (1995) defines the task of ownership as “determining the privileges for others”. Important to mention is that the GDPR (European Parliament and the Council of European Union, 2016) use the word ownership.

Another definition was found to describe the relationship between the data subject and its personal data: *data custodian*. Data custodian a role of managing data for an organization (Cheong & Chang, 2007). Carnegie Mellon University describe in their governance policy that data custodians, which is someone working at the University, has a administrative and operational role over institutional data (Carnegie Mellon University, 2009).

Data Refinement As described above in Section 2.3.3, the *object* represents the data on which an action is performed. But differences become visible in how detailed this data is described when reading privacy statements. In the fourth step of the six-step approach of Breaux et al. (2014) for extracting privacy requirements from text is this called refinement. In the second step the type of activity is assigned to each requirement. Next, in step 4, the roles in the requirements are further specified. Three types of role specification are possible: abstraction, refinement, and exclusion (Breaux et al., 2014). *Role abstraction* takes place when one piece of information is more general than the other (e.g. ‘information‘ is more general than ‘date of birth‘). Role abstractions can be indicated by the word ‘other‘ (Breaux et al., 2014). When a concept is explained in more detail in the same sentence or document, it is called *role refinement*. Common keywords that indicate a refinement are ‘including‘, ‘such as‘, and ‘for example‘ (Breaux et al., 2014). The last possibility is *role exclusion*. This happens when a piece of information is not considered to be part of a bigger whole (e.g. ‘IP address‘ is not part of ‘personal information‘).

Facebook (2016) uses a lot of examples to explain the object in further detail (see Example 2.3.1 and Example 2.3.2).

Example 2.3.1 *We also use information we have to provide shortcuts and suggestions to you. For example, we are able to suggest that your friend tag you in a picture by comparing your friend‘s pictures to information we've put together from your profile pictures and the other photos in which you've been tagged.*

Example 2.3.2 *We collect information about the purchase or transaction. This includes your payment information, such as your credit or debit card number and other card information, and other account and authentication information, as well as billing, shipping and contact details.*

Examples 2.11: Examples are taken from the Facebook Data Policy (Facebook, 2016). Marked is the refinement keyword and the abstraction keyword for indicating a more detailed description of the initial statement.

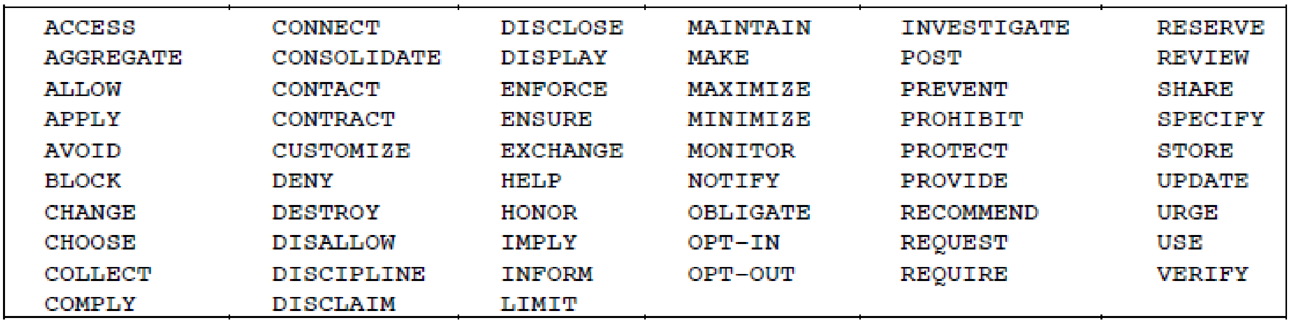
A more ethical questions that comes in mind is which information is given in the examples and which information is not.

Processing Activities

Gharib et al. (2016) describe actions as the relationship between information and a specific goal of the user, and refers to the actions as a type of *use*. Therewith, an use action can be *produce*, *read* or *modify* information. However, Breaux et al. (2014) define four types of actions (operations), including ‘use‘, that can be described in a privacy requirement: *collect*, *use*, *retain*, and *transfer*. Another set of basic actions is CRUD. The actions are derived from standard SQL statements for database management and HTTP-requests. CRUD stands for *create*, *read*, *update*, and *delete*. European Parliament and the Council of European Union (2016) describe *processing* as “any operation or set of operations which is performed on personal data”.

Antòn et al. (2002) use a goal-based approach for privacy requirement engineering. They describe that system goals can be based on privacy vulnerability. Also, they point out the importance to make a distinction between the different action types. (i) *Information monitoring* is, from a marketing perspective, about tracking information to benefit the consumer (e.g. remembering the last seen products in a Webshop) or for the purpose for optimization (statistical analysis). (ii) *Transferring information* is the a contradictory action regarding privacy since it is normally information should not be transfered. However, the goal is to perform the transaction when allowed and controlled. Combining data from different sources is called (iii) *information aggregation*. This action can be a privacy vulnerability goal since combining data can result in new information that is more personal or valuable than the uncombined data. Note that prior to a data aggregation at least two information transfers should have been performed, since the action is only allowed when an actor has permission to use both data sets. Another goal is the (iv) *information storage* goal. This goal includes how and what data is stored. (v) *Information collection* can be of two types. A collection can be direct when the user itself enters information or indirect when the system owner collects data of the user‘s behavior on, for example, a website. The last type mentioned by Antòn et al. (2002) is (vi) *information personalization*. Here, personal information is used to personalize the system environment for the user.

The Privacy Goal Management Tool (PGMT) maintains a goal repository which consist of a list of 57 keywords that occur in privacy statements often (Anton et al., 2004). Each keyword is an expression of a goal to standardize the goals in privacy statements. In the scope of this projects the keywords will be recognized as processing activities. The keywords are shown in Figure 2.12.



ACCESS
AGGREGATE
ALLOW
APPLY
AVOID
BLOCK
CHANGE
CHOOSE
COLLECT
COMPLY

CONNECT
CONSOLIDATE
CONTACT
CONTRACT
CUSTOMIZE
DENY
DESTROY
DISALLOW
DISCIPLINE
DISCLAIM

DISCLOSE
DISPLAY
ENFORCE
ENSURE
EXCHANGE
HELP
HONOR
IMPLY
INFORM
LIMIT

MAINTAIN
MAKE
MAXIMIZE
MINIMIZE
MONITOR
NOTIFY
OBLIGATE
OPT-IN
OPT-OUT

INVESTIGATE
POST
PREVENT
PROHIBIT
PROTECT
PROVIDE
RECOMMEND
REQUEST
REQUIRE

RESERVE
REVIEW
SHARE
SPECIFY
STORE
UPDATE
URGE
USE
VERIFY

Figure 2.12: Privacy statement keywords, taken from Anton et al. (2004).

An overview of the actions described in the literature is given in Table 2.3.

Transfer Source & Recipient When transferring information two new actors are involved.

Table 2.3: Overview of actions regarding information or data.

Action	Description	Source	Activity Type
Produce	Creating information.	Gharib et al. (2016)	use
Read	Consuming information.		use
Modify	Modifying information.		use
Collect	Access, collect, obtain, receive data from another party.	Breaux et al. (2014)	transfer
Use	Using data for own purpose.		use
Retain	Retain data for a period of time or location.		transfer
Transfer	Transfer, move, send, or relocate data to another party.		transfer
Create	Create, PUT (HTTP), INSERT (SQL)	Aho et al. (1986)	use
Read	Read, GET (HTTP), SELECT (SQL)		use
Update	Update, POST (HTTP), UPDATE (SQL)		use
Delete	DELETE (HTTP and SQL)		use
Monitor	Collecting user information.	Antòn et al. (2002)	use
Aggregate	Combining or associating data to new data.		use
Store	Storing specific information in a specific place.		transfer
Transfer	Transfer, disclose, sell, share, and provide information.		transfer
Collection	Collecting information from direct or indirect sources.		transfer
Personalization	Recognize valuable information to customize or tailor the system environment.		use

First, the information is transferred from a *source* (Breaux et al., 2014). The source can be the data subject, but also a third party. Examples for both cases are highlighted in green in Examples 2.3.3 and 2.3.4 respectively. The second actor involved in a transfer activity is the entity that receives the information, called the *target* (Breaux et al., 2014), or *recipient* (European Parliament and the Council of European Union, 2016) . A recipient can be the controller (the performer of the activity) as shown in Examples 2.3.3 and 2.3.4 (marked pink), but it can also be a third party that is receiving the personal data (see Example 2.3.5).

Condition

A privacy statement may require predefined conditions to be satisfied before the request for permission can be granted (He et al., 2003). Examples 2.3.8 and 2.3.9 shows two examples of conditions.

Constraint *Constraints* indicate a precondition (for example: except if, if not, unless) (Breaux

Example 2.3.3 *[Facebook] collects information from you when [. . .].*
Example 2.3.4 *[Facebook] receives information about you from companies [. . .].*
Example 2.3.5 *When you use third-party services that use, or are integrated with, our Services, they may receive information about what you post or share. [. . .].*

Examples 2.13: Examples are based on the Facebook Data Policy indicating the source and the recipient (Facebook, 2016).

Example 2.3.6 *We work hard to protect your account using teams of engineers, automated systems, and advanced technology such as encryption and machine learning. .*
Example 2.3.7 *We also offer easy-to-use security tools that add an extra layer of security to your account.*

Examples 2.14: Examples are taken from the Facebook Data Policy. The activities are highlighted (Facebook, 2016).

et al., 2006). This might mean that the controller is only allowed to process personal data when either the controller or data subject meet a precondition.

Example 2.3.8 *We collect information from or about the computers, phones, or other devices where you install or access our Services, depending on the permissions you've granted.*
Example 2.3.9 *When we have location information, we use it to tailor our Services for you and others [. . .].*

Examples 2.15: Examples are taken from the Facebook Data Policy. The conditions are highlighted (Facebook, 2016).

Purpose

The *purpose* is the reason the process activity in the privacy requirement is performed. It is important since it explains why the information is used or transfered (Breaux et al., 2014). He et al. (2003) distinguish two types of purposes: the consumer purpose and the business purpose. A consumer purpose is when a consumer has agreed on a request from a data collector (He et al., 2003). The purposes also explain how the data will be used. An example of a consumer purpose is shown in pink in Example 2.3.10. Business purposes describe a certain business task wherefore consumer information is needed (see the green highlighted part of Example 2.3.10). Consumer information is requested for getting access or to perform operations.

Conflicts & Priorities

When two or more contrary modalities are applicable to one element, e.g. an actor is both permitted and prohibited to perform an action, there is a *conflict* (Breaux et al., 2014; Nis-

Example 2.3.10 *We use the information we have to improve our advertising and measurement systems so we can show you relevant ads on and off our Services and measure the effectiveness and reach of ads and services.*

Examples 2.16: Examples are taken from the Facebook Data Policy. The consumer (pink) and business (green) purposes are highlighted (Facebook, 2016).

senbaum, 2011) (see Figure 2.17). Matteucci et al. (2013) split conflicts into three different categories where each category includes a conflict between a 'deny' and an 'allow' state. A conflict can be a (i) *contradiction*. This means that two privacy requirements are contradictory to each other and fit the definition given by Breaux et al. (2014) above. The policies state the same, but with a different contrary effect (Matteucci et al., 2013). A contradictory effect can also be a result of a sub-requirement that specifies an extra conflicting condition, a so called (ii) *exception*. For example, when only one person (the exception) from a group is allowed to access particular information where the rest is prohibited. At last, (iii) , a *correlation* also has an extra condition, but this condition is applicable to the an attribute in the environment (Matteucci et al., 2013). For example, when a time limit is added to the requirement. So, in general a specified group is not allowed to access information, but they are temporarily allowed to access the information until the end of the month.

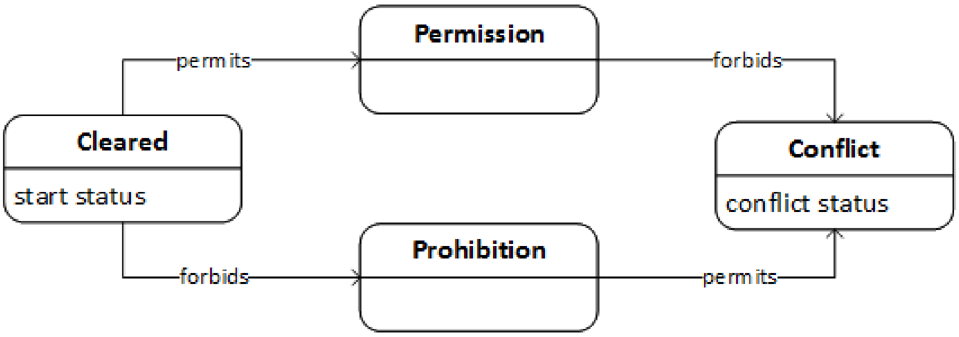


Figure 2.17: The relation between permission and prohibition.

2.3.4 Legislation & Privacy Statements

Privacy statements are often based on legislation (Laplante, 2013). This section focuses on the influence of law on requirement engineering.

In the 1960s the upcoming information technologies and innovations in the digital world have been viewed as a great risk to privacy violation. This because these innovations are paired with a growing amount of data, including personal information and growing organizational databases (Nissenbaum, 2011).

The American organization for protecting consumer privacy in Information and Communication Technology (ICT) systems is the Federal Trade Commission (FTC) (ISO/IEC 29100:2011(en), 2011). The FTC is enforcing the law to prevent violation of privacy (Laplante, 2013; ISO/IEC 29100:2011(en), 2011). Initially, the FTC came up with the *Fair Information Practice Principles* (FIPPs) in the late 1970s and begin 1980s. In the period of 1990s and 2000s the FTC reduced the FIPPs to five principles. These principles were said to be the pillars for privacy protection and were named as follows: 1) *Notice/Awareness*, 2) *Choice/Consent*, 3) *Access/-Participation*, 4) *Integrity/Security*, 5) *Enforcement/Redress* (Cate, 2006).

In 1994 stated Reidenberg that the US government was avoiding the need of legal rules for the last 20 years, however at that time the urgency for a legislation regarding privacy peaked. The more general definition of privacy, “the allocation of rights to personal information“, did not cover the ‘new‘ context of privacy anymore (Reidenberg, 1994). The FIPs alone were not sufficient to protect us from privacy risk, however, they are still relevant (Richards & King, 2014).

The FTC started to use enforcements regarding privacy violations for companies that did not align their business processes with their privacy statement (Breaux et al., 2014). In addition, the US introduced new domain specific law, such as Fair Credit Reporting Act (FCRA), the Gramm–Leach–Bliley Act (GLBA), the Children Online Privacy Protection Act (COPPA), and the Health Information Portability and Accountability Act (HIPAA) (Breaux et al., 2014).

In May 2018 the *General Data Protection Regulation* (GDPR) will be launched as the data protection regulation on European level (Tankard, 2016). The GDPR is a result of the lack of consistency within European countries as all countries had their own legislation which resulted into discrepancy for companies operating abroad. Note, the previous legislation (Directive 95/46/EC) is dated from 1995 when only 1% of the people in the world used the Internet, whereas nowadays almost everyone uses the Internet (Tankard, 2016).

Table 2.4 shows an overview of the evolution of privacy in legislation over time for both US as European laws. However, there is a difference between the two nation regarding privacy legislations. The main difference is that Europe handles an omnibus law system where each sector is covered by the same law with some adjustments per sector. On the other hand, the United States have for each sector their own law, sectoral laws. Privacy is considered to be a Human Right in Europe where in the United States persons have some constitutional rights to privacy. Additionally, no processing of personal data is the default in Europe, but in the United States processing of personal data for commercial use is acceptable.

Table 2.4: Time line of the evolution of Privacy (Requirements)

1948	Declaration of Human Rights, United Nations
1950	Convention of Human Rights, Council of Europe
1970s - 1980s	The FTC introduces the FIPPs (Cate, 2006), United States
1980	Organization for Economic Cooperation and Development (OECD) Guidelines
1981	Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, Council of Europe
1995	Directive 95/46/EC of the European Parliament and the Council of 24 October on the protection of individuals with regard to the protection of personal data
1996	Health Insurance Portability and Accountability Act (HIPAA), protects health care coverage in the United States
1990s - 2000s	FIPPs are reduced to five principles and translated to a national law (Cate, 2006).
1999	Gramm-Leach-Bliley Act (GLBA), the most extensive financial privacy legislation in the US history (Anton et al., 2004).
2000	Children‘s Online Privacy Protection Act (COPPA), covers the information protection of children under 13 years old, United States
2000	Charter of Fundamental Rights, Nice, Europe
2004	Asia-Pacific Economic Cooperation (APEC) Privacy Framework for uses and integrity of personal information, security safeguards, and notice, among others
2007	Treaty of Lisbon, Europe
2018	General Data Protection Regulation (GDPR) is a new privacy regulation that will cover privacy protection on European level

Chapter 3

Design Science

In this chapter the literature is summarized in a conceptual model. An extensive conceptual model is created as a representation of the found literature in Chapter 2(3.1. By holding interviews the conceptual model is validated on usefulness by Privacy Expert, which results in a revised conceptual modal (3.2) that is used as a guideline for the text mining techniques in the next chapter (4).

3.1 Preliminary Conceptual Model

In software engineering a *conceptual model* is the expression of a system concept that models the interactions or relationships between components and its functionalities (Sokolowski & Banks, 2010, p. 149). We use the broader definition of Mylopoulos (1992) that describes a conceptual model as an "abstract specification of [...] a knowledge representation".

The terms and definitions of privacy ontology of Gharib et al. (2016), explained in Section 2.3.2 of this report, are used as starting point for finding and modifying new and existing components. Based on the literature, multiple components were found. The components and the relationship between the components are modeled in a class diagram based on the *Unified Modeling Language* (UML).

During the modeling five actors were defined: statement owner, controller, data subject, source, and recipient. An actor is described by the European Parliament and the Council of European Union (2016) as a "natural or legal person, public authority, agency or other body". The first three actors are always present in a privacy statement, where the source and target actors are only there when linked to a specific type of activity.

This section first introduces the framework components that have a direct relation between with the written statement, this includes: conflict, privacy statement type, statement owner, and then the privacy requirements written in the privacy statement. Next, the separate components of a privacy statement are explained in more depth.

Privacy Statement

The *privacy statement* is the policy statement document that contains the privacy requirements that are analyzed in this project.

Privacy Statement Type. Antòn et al. (2002) mention that privacy statements are hard to compare with each other without considering the domain, we selected a set of privacy statements that have common features to narrow down the scope of the analysis. This means that a set of policies with similar characteristics will be selected for building the tool for modeling, such as: owners, topics, or content.

Statement Owner. The *statement owner* is an actor that owns the privacy statement (see Section).

Conflict. A *conflict* is a contradiction between two or more privacy requirements. In other words, when two privacy requirements contradict each other, by one granting a right to perform an activity and the other prohibiting the same activity. Three types of conflicts are identified by Matteucci et al. (2013): contradiction, exception, and correlation.

Privacy Requirement

As explained in Section 2.3.3, *privacy requirement* is one requirement, written in natural language, in a privacy statement document. This statement consists of different, mandatory and optional, components that together form the statement. The components are further explained and set out below.

Modalities. Modalities define the type of the privacy statement as described in Section 2.3.3. The type is determined by the combination of the operations (or activities) performed on an object (He et al., 2003). We defined the three most common and representative modalities that we found. In the framework a privacy statement can indicate a permission, prohibition, or obligation. A permission, also called a right, is a transition of the ability to perform an action on a specified data set (Gharib et al., 2016; Breaux et al., 2006, 2014). Permission includes an authorization activity. Contrary to a permission, one is also able to remove the rights to perform an action on a data set. This is referred to as a prohibition. The last modality, obligation, refers to a duty of an obligated party to perform a certain task (Breaux et al., 2006).

Privacy Requirement Types. We found two types of privacy statements in literature on this subject (see Section 2.3.3). However, we assume that each privacy statement concerns delegation, which means that is assumed that in each statement a right or obligation is assigned from one actor to another, and therefore the statement type is not separately modeled in the framework.

Controller. The *controller*, as described in 2.3.3 also called actor or stakeholder, is the actor that performs the activity in the privacy statement. Since multiple actors (or stakeholders) are involved in the framework, we use the terminology of the European Parliament and the Council of European Union (2016) to indicate the entity that performs an action as the *controller* of the privacy statement.

Personal Data. As in Section 2.3.3, taking from literature we defined the entity *personal data* as (a set of) information on which an activity is performed. The actor that owns this data object is referred to as the *data subject*. Specifying the personal data is out of the scope of this project, and therefore object refinement will not be modeled in the framework.

Processing Activity. The *activity* component represents the action that is performed on a data object by the controller. The activity describes what happens with the personal data. An overview of activities is given in Figure 2.12, Table 2.3, and described in Section 2.3.3. A distinction is made between a *use activity* and a *transfer activity*, since two more actors are involved when transferring information from a sender to a receiver. These actors are called the *source* and the *recipient* of the activity. Therefore two actors are added to the framework to model the information transition from the source (the sender) to the recipient (the receiver).

Condition. Optionally, a privacy statement can contain a *condition* (see Section 2.3.3). A condition is a prerequisite that needs to be met before the described activity can be performed.

Breaux et al. (2006) refers to a constraint as a part of a statement that describes a precondition. A constraint will not be included in the framework since we can assume that the exception is applied to a particular person or group, and therefore a constraint does not need to be modeled separately.

Purpose. Section 2.3.3 describes that the *purpose* is the reason the activity is performed. It explains the 'why' of the activity. The purpose is not always included in a statement and therefore optional.

Data Refinement. As an extension to the personal data component *data refinement*, described in Section 2.3.3, is added to the framework as it provides important information about personal data (Breaux et al., 2014).

3.2 Revised Conceptual Model

This conceptual model is evaluated by experts in the field of risk advisory from the host company where this research is being conducted. Instead of focusing on the validity of the conceptual model itself, the focus is on the application of the model. In short, the purpose of the model is validated by conducting a semi-structured interview, and is therefore a qualitative method.

A validation task tests the conceptual model to check how well it represents the real world. In this case, we strive to validate how the conceptual model fits a useful purpose for the final deliverable (Sargent, 2005). Sargent (2005) mentions the *face validity* method, among other methods, where people's knowledge is used to assess the behavior or the results of a conceptual model. The interviews conducted and the model revision accordingly are explained in the next sections.

3.2.1 Validating the Usefulness and Applicability of the Conceptual Model

The usefulness and a potential application is tested by holding interviews. The interviews were conducted with four Privacy advisors from the Deloitte Privacy Services team. Questions were

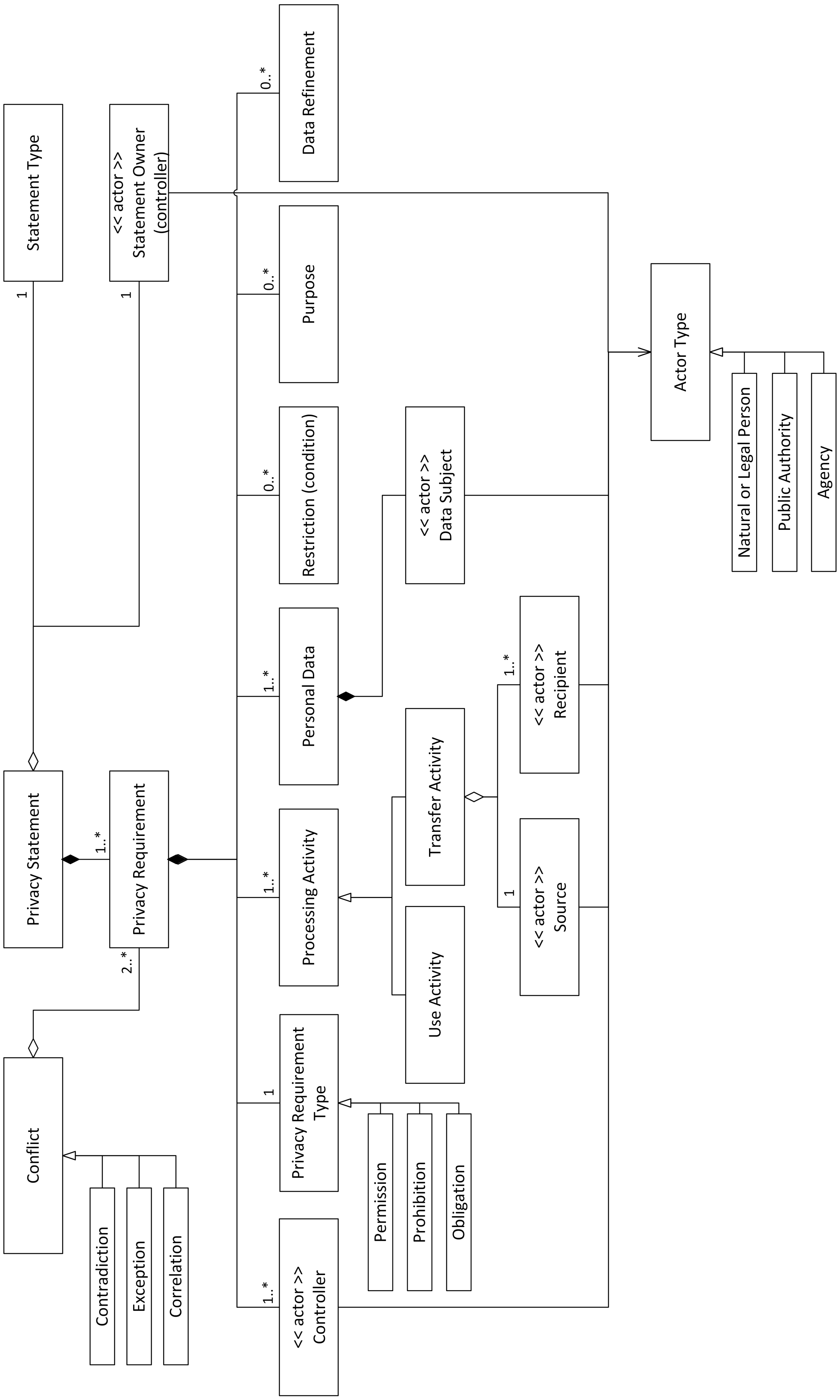


Figure 3.1: Privacy requirements framework

asked are about their work activities concerning privacy statements. The first questions focus mainly on the communication towards clients and any possible problems that may occur. After these first questions, the conceptual model is shown. Additional questions are asked about how the model could serve as a solution for their stated problems. The interview protocol is included in Appendix B.1.

The four interviews (Deloitte Privacy Services, 2017) took between 23 and 27 minutes and were conducted in Dutch. All interviews were recorded with consent of the interviewee and transcribed afterwards. A summary of the answers has been translated into English and included in Appendix B.2.

Each participant had some type of experience with privacy statements. Some had written a policy themselves, but mostly the involvement with a privacy statement was as part of a bigger project. Their contribution consists mainly of providing guidelines (verbally and on paper) to their clients regarding privacy statements. The creation of a privacy statement can be iterative as multiple feedback sessions with the client can be involved. In the evaluation and feedback sessions with the client mostly structure and topics are discussed. The communication with the client about privacy statements covers four steps (Deloitte Privacy Services, 2017). These steps are, the advisors check 1) if there is a privacy statement. If a privacy statement does exist: 2) the date of the last revision, and 3) the alignment with the processing activities. Lastly, 4) a list of possible updates is created.

The main points of communications are:

- The balance between being vague about data processes, and thus flexible, and being detailed enough to comply with regulations;
- The approach for addressing consumers (the addressees of the privacy statement) (i.e. the text should be understandable for its readers);
- The use of layered statements for breaking up large amounts of text for legibility;
- The selection of the topics that need to be covered by the statement (the chapters of Article 29[1] can be used as guideline.).

One of the respondents also mentioned that a privacy statement is not necessarily mandatory (Respondent 3). Instead, an informative text can be used to explain to the users what happens with their personal data.

A communication approach is adjusted depending on the knowledge and experience of the client, and the relationship between the advisor and the client. If a client has no or limited knowledge about privacy statements, the steps of creating one need more support from the privacy advisors. Support can consist of frameworks, templates, checklists, or Excel-sheets with requirements or topics, but may also consist of advice about how the consumers need to be addressed from the perspective of what is important for the client.

Main Identified Problems

The main problems that the advisors found while communicating with clients are communicating about the level of detail that the privacy statement requires. The challenge is to create a

[1] Article 29 Data Protection Working Party (Art. 29 WP) is part of European Data Protection Regulation.

privacy statement that is both detailed and flexible to allow for changes to be applied to certain processes. With flexibility we mean the ability to revise the privacy statement (and data inventory) in a limited amount of time. When a privacy statement is detailed, all the details should change when a processing activity concerning those details changes. On the other hand, a vague description, with less detail, does not have to change when the same change in activity applies. Important to note is the lack of knowledge of organizations regarding the processing activities they have. When organizations are not aware of all the processing activities, they could not explain the activities in their statements. Additionally, the GDPR requires that each organization, that processes personal data, should have a *Data Processing Inventory* (DPI)(European Parliament and the Council of European Union, 2016)(Respondent 4). According to Article 30 of the Council Regulation (EU) no 2016/269 the "controller or processor [. . .] is required to 'maintain a record of processing activities under its responsibility' and 'shall make the record available to the supervisory authority on request' " (European Parliament and the Council of European Union, 2016). This means that the data on processing activities should be available at all times.

Besides the difficulties in finding the balance between details and flexibility, another problem is the lack of a specified format and therewith a lack of knowledge of what the statement should contain. Moreover, the organizations often have the wrong approach regarding the purpose of privacy statements. They think of a privacy statement as a 'tick-the-box' exercise. The pitfall of that approach is that it is often fast and legally written.

The number of stakeholders (and interests) involved might slow down the iterative process of giving and processing feedback. Also, the purpose of the processing activities is rarely defined.

Most problems can be solved by providing examples and including enough feedback loops. The interviewees suggested to think of a strategy first. An organization should only process data if it fits the organization's strategy. If they store data just for the sake of having it for future purposes, they do not comply with the GDPR. Personal data may only be processed when it has a purpose, as described in Article 5 of the Council Regulation (EU) no 2016/269: collected data must be "adequate, relevant and limited to what is necessary in relation to the purposes for which it is processed (*data minimization*)"(European Parliament and the Council of European Union, 2016). In short, if the activity process does not support the organization's strategy, it should not be preformed and therefore not described in the privacy statement.

Proposed Solution Based on the Conceptual Model

After showing the framework, two types of solutions were mentioned. First, solutions to problems that focus on the writing process of privacy statements, and second solutions that are focused on problems with existing statements.

A proposed solution for writing statements was to create a tool that can create a privacy statement automatically based on human input. The tool will determine the structure of the privacy statement and the privacy requirements. The conceptual model can also be used as framework or guideline for the creation of a privacy statement.

A solution to existing privacy statements that are difficult to read is the use of *layered statements*. Those are statements that consist of multiple layers, where each layer has a different level of detail. The highest level is the level that is initially visible to the consumers. It is a high-level summary of the content of the statement. However, when a user wants more in-depth information about a certain piece of personal data he or she is able access an second layer where more details are provided for that piece of information. This layer might even contain legal explanation or support. A layered statement can also be presented in the form of a chat-bot. As a consumer you can ask a chat-bot a question concerning your personal data and the chat-bot will answer your question based on the data filtered from one or more privacy statements.

Another application of the conceptual model can be a tool that filters all the processing activities and the purposes from the textual privacy statement and creates a list of it which can be compared to the DPI. In this way, the list based on the privacy statement and the data inventory can be compared regarding completeness. But organizations can also compare multiple statements with each other.

3.2.2 Revised Model

Based on the validation interviews we will focus on the high-level content of a privacy requirement. We will filter the following components from the privacy statement requirements: controller, requirement type (modality), processing activity, personal data, data subject, purpose, and depending on the performance of the algorithm possible restrictions.

Privacy Statement, Statement Owner, and Privacy Requirements

Privacy statement, as the root of the conceptual model, is the source of privacy requirements and therefore stays in the revised conceptual model. Naturally, the privacy requirement block also stays in the conceptual model as it contains the components we look for during the analysis phase. As mentioned before, the statement owner is, in this project, assumed to be the controller as well as the processor. However, the European Parliament and the Council of European Union (2016) makes a clear distinction between the two.

Privacy Requirement Components

Besides the controller, some other components are kept in the revised conceptual model based on the interviews. As mentioned by some of the participants and stated in the GDPR (European Parliament and the Council of European Union, 2016), organizations are required to keep track of all processing activities that involve personal data. Additionally, organizations are also required to provide a purpose for storing and using data. Therefore, the decision was made to focus on the processing activity and personal data (including the data subject and data refinement). Another item mentioned was that, from a consumer perspective as well as a controller perspective, it might be useful to search for a specific piece of personal data and the processing activity involved.

Lastly, the modality (requirement type) and the restriction are included in the revised framework. The modality is an interesting case since it can be only one of the predetermined values: permission, prohibition, or obligation. So, in stead of extracting a specific set of words from the privacy requirement, an indicator needs to be found first before choosing one of the three modalities. The condition is considered to be an important part of a privacy requirement, and is therefore included.

The revised model is shown in Figure 3.2.

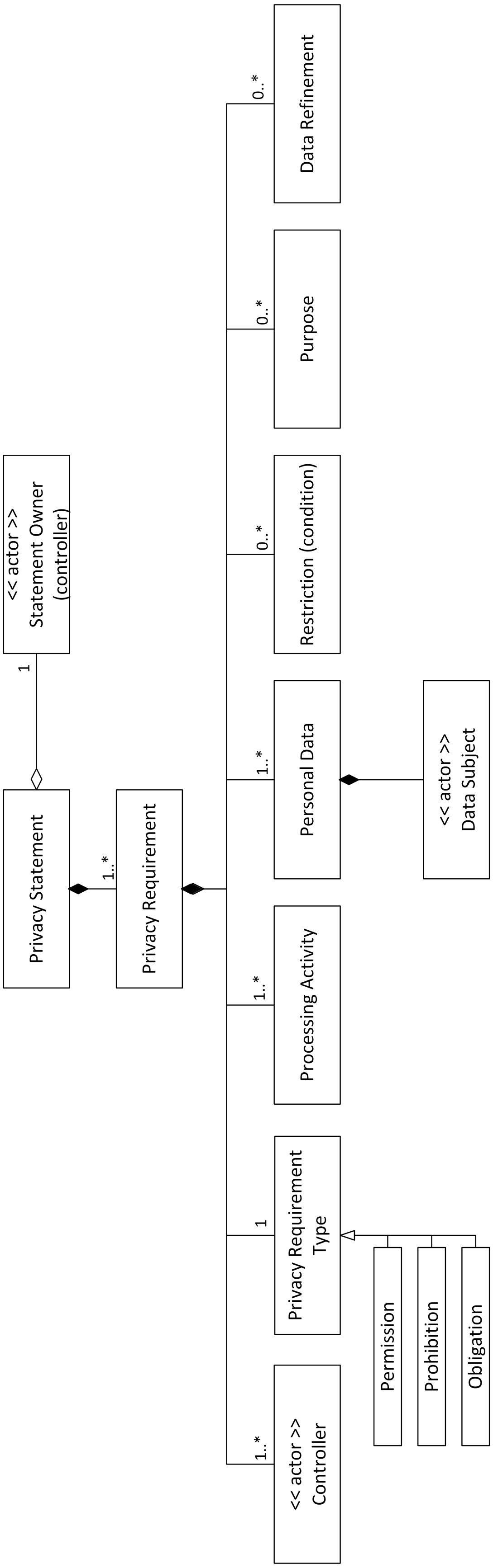


Figure 3.2: Revised privacy requirements framework

Chapter 4

Method for Extracting Components from Privacy Statements

The processes described in the upcoming sections are explained according to the *Cross Industry Standard Process for Data Mining*, or so called CRISP-DM model. CRISP-DM is a framework created as a standard approach for data mining tasks (Wirth & Hipp, 2000). Figure 4.1 shows the CRISP-DM framework that is used to define the phases of this project. The solution that consists of the combination text mining techniques that perform best on the task of extracting components from privacy statements is referred to as the *algorithm*.

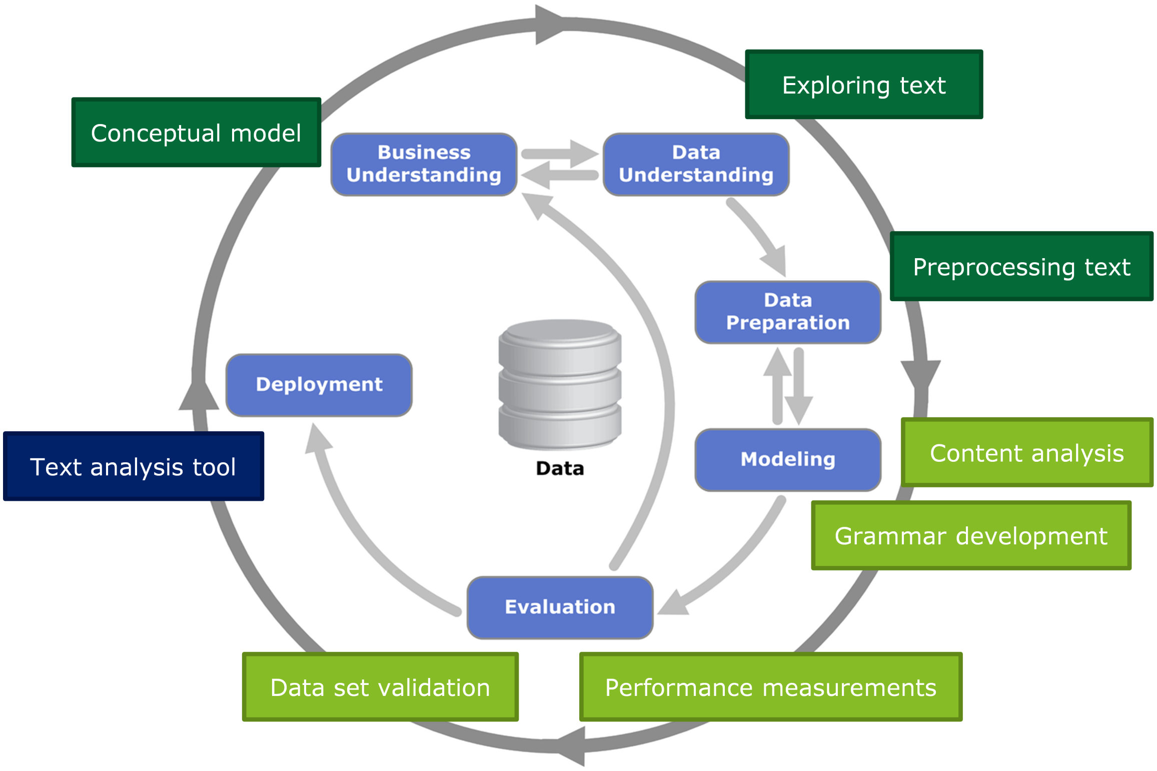


Figure 4.1: CRISP-DM framework as a guideline for the exploratory research on text mining techniques. Own creation based on Wirth & Hipp (2000) and Wikipedia (2017).

Following Figure 4.1, the first section (4.1) will explain the data that is used for the analysis. Section 4.2 describes the preparation process, followed by the explanation of the content analysis and the development of the algorithm in Section 4.3 and 4.4. Last, the evaluation process of the algorithm, the readability of the privacy statements, and the final results are described in Section 4.5, 4.6, and 4.7.

The whole process from the preprocessing steps to the final output is shown in a Process Deliverable Diagram (PDD) which shows both the activities as well as the products of the activities. The PDD is a technique for modeling meta-processes and meta-data (Weerd et al., 2005). On the left-hand side of the diagram the meta-process describe the activities that are performed during the development of the algorithm. The meta-data model, on the right hand side, describes the deliverables, or data objects, that are a output result of the activities. The deliverables are modeled using the Unified Modeling Language (UML) (Weerd et al., 2005). The PDD is shown in Figure 4.2 and is explained in the sections below.

4.1 Data Understanding

Three privacy statements of websites were used and prepared as a training set and two additional privacy statements were treated in the same way to serve as test set. All privacy statements were written in English and fit two main categories: social media and Webshops. Those categories were chosen based on the common way of gathering personal data: people visiting the websites consciously opt to provide the organization with personal data. For example, on social platforms, people use personal data in the form of a name and email address to sign in. In addition they provide personal data such as a location when leaving a message on a timeline. They leave similar information when buying goods via a Webshop. People who buy an airplane ticket or a book are aware that giving personal data at the end of the buying process (check out) is necessary to have the product delivered to the right address (home address or email address). The used privacy statements are shown in Table 4.1.

Table 4.1: Overview of the used privacy statements and their use.

Name	**Data of last revision**	**Number of requirements**	**Number of words**	**Category**	**Training / test**
Facebook Data Policy	September 29, 2016	46	1250	Social media	training
KLM Privacy Policy	December 1, 2016	53	1117	Webshop	training
Google Privacy Policy	April 17, 2017	50	1200	Social media	training
Instagram Privacy Policy	January 19, 2013	31	975	Social media	test
Lufthansa Information on our Privacy Statement	April 17, 2017	14	287	Webshop	test

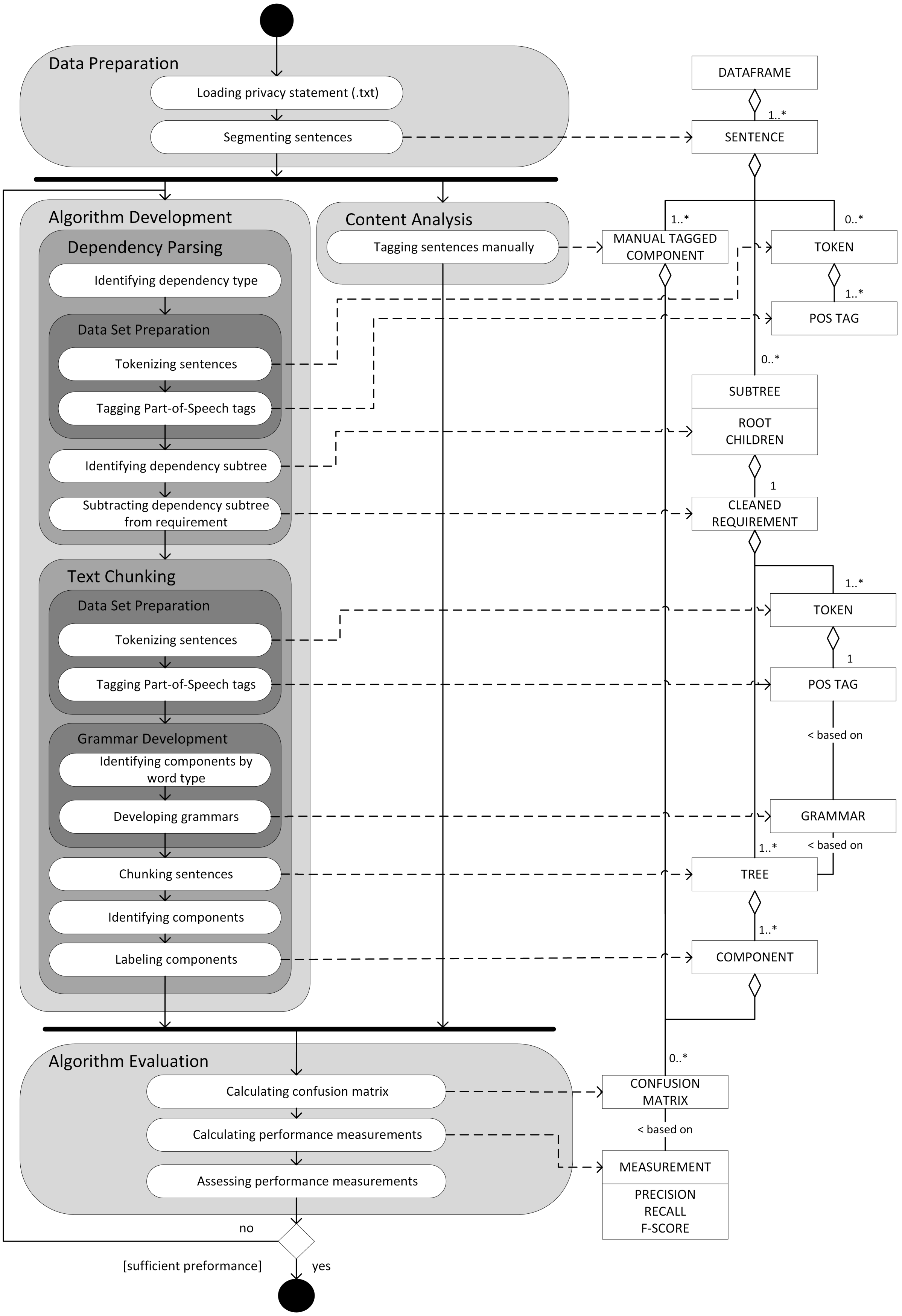


Figure 4.2: PDD visualizing the process of the used (preprocessing) text mining techniques applied to the textual privacy statements.

4.2 Data Preparation

Before any text mining techniques can be applied, the programming language Python 2.7 and the necessary packages needs to be installed to be able to apply various text mining techniques (see the first section of Figure 4.2). This sections explains how Python is used to segment the privacy statements into separate sentences which then form the requirements of the training and test sets.

4.2.1 Python Installation and Packages

All data processing is done in Python. Including data cleaning and preprocessing, text mining, and performance measuring.

The processes are performed with Python in an interactive iPython environment, a `Jupyter Notebook` (hereafter: notebook), which supports existing packages, such as `nltk`, `spaCy`, `pandas`, `scikit-learn`, and `matplotlib`. Most Python packages were installed through *Anaconda*, a package manager containing a wide range of pre-installed open source packages[1]. The main reason for using Anaconda is the predefined internal and external dependency for various Python packages. All packages used were installed in a virtual environment. The environment allows sharing the written code more easily by just installing the environment and the code will run on other devices. See Appendix C.1 for the complete list of packages that are installed in the used environment. A README containing the installation guidelines for the Anaconda environment can be found in Appendix D.1.

NLTK

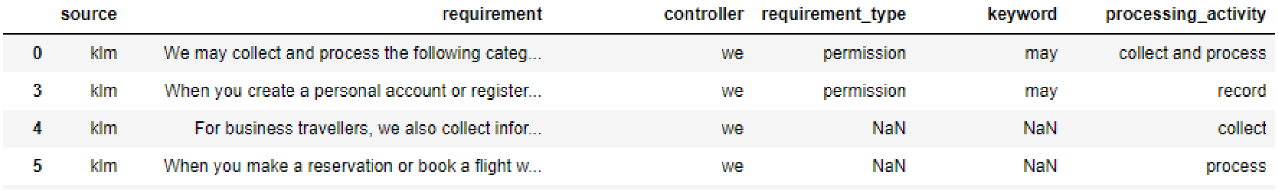
A common Python package for natural text processing is the *Natural Language Toolkit* (NLTK)[2]. NLTK is an open source program that contains modules that support text mining techniques (Bird, 2006; Loper & Bird, 2002). A number of models that were used this research are the following (Loper & Bird, 2002):

- *Parsing modules* can be used to represent the structure of texts by creating trees, but also as a parser that identifies linguistic groups (e.g. nouns, verbs, etc.);
- *Tagging modules* have taggers that label words with information;
- *Visualization modules* define an interface for visualizing data structure, such as trees and plots.

Simple processing tasks can be performed with NLTK, such as tokenization and stemming, and tagging. Tokenization is the separation of words in a sentence on whitespace (Bird, 2006). For tagging, a regular expression approach can be used to assign a tag to a token according to the pattern (Bird, 2006). Chunking with NLTK is a technique for analyzing text that is first tagged, again a regular expression is used to determine the relationship between the tags (Bird, 2006).

[1] https://docs.continuum.io

[2] Used version: 3.2.3 | http://www.nltk.org



	source	requirement	controller	requirement_type	keyword	processing_activity
0	klm	We may collect and process the following categ...	we	permission	may	collect and process
3	klm	When you create a personal account or register...	we	permission	may	record
4	klm	For business travellers, we also collect infor...	we	NaN	NaN	collect
5	klm	When you make a reservation or book a flight w...	we	NaN	NaN	process

Figure 4.3: The *head* (i.e, the first five rows) of a part of the dataframe as visualized in a notebook.

spaCy

`spaCy`[3] has a large scale of text mining modules based on *Cython*. Cython is a C-extension for Python and that allows the creation of C extension for Python. spaCy is considered to be fast and accurate. The average speed for processing words is 13.96 (WPS) and the accuracy is 91.8. With these scores spaCy is one of the fastest syntactic parsers and has an average accuracy compared to other parsers (spaCy, 2017). SpaCy has a text parser that is able to create a dependency tree of natural language.

pandas

`pandas`[4] is an open source library that is used for creating and editing data structures, and as a data analysis tool for Python. The pandas `dataframes` (tables) are used to store the data in a structured manner. The benefit of using a dataframe is the visualization in the notebooks (see Figure 4.3). It allows for quick feedback. The pandas dataframes can be queried and multiple tables can be merged.

scikit-learn

`scikit-learn`[5] is an open source package in Python that provides tools for data mining and data analysis. The scikit-learn toolkit has an extensive number of machine learning algorithms (Pedregosa et al., 2011). The toolkit is only dependent on two other Python packages: `numpy` and `scipy` (Pedregosa et al., 2011). The metrics functions of scikit-learn will be used to calculate the performance of the algorithm that is created in iterative steps. The functions that are used from the metrics module include precision, recall, and the f-score.

Matplotlib

For the visualizations of the data and the results a 2D graphics packages was used: `Matplotlib`[6] (Hunter, 2007). Matplotlib is build on `pylab` and `numpy` and is able to visualize data with various types of charts and graphs (Hunter, 2007). A benefit of the Matplotlib is the ability to interpret a pandas dataframe as input for a visualization. Additionally, the output (e.g. a

[3] Used version: 1.9.0 | https://spacy.io

[4] Used version: 0.20.1 | http://pandas.pydata.org/index.html

[5] Used version: 0.18.1 | http://scikit-learn.org/stable/index.html

[6] Used version: 2.0.2 | https://matplotlib.org

chart) can be directly shown in a notebook, which supports the iterative process of improving the algorithm by direct feedback.

4.2.2 Prepare Training and Test Data

The privacy statements were copied manually from the website and pasted as plain text into a text editor. Next, each statement was separately stored as a text file (.txt). For preprocessing the privacy statements a Jupyter Notebook is used. A Jupyter Notebook allows for Python code to be run in code blocks that show the output of the code immediately under the code block. The text files (.txt) are interpreted by Python and cleaned according to the following rules that are written in Python code:

1. Each requirement is extracted from the textual privacy statement by:
   (a) Replacing the website name that includes dots with a name without the dots ('Amazon.com' and 'Amazon, Inc.' turn into 'Amazon');
   (b) Replacing abbreviations with dots such as "e.g." with similar wording that does not have dots, such as 'for example';
   (c) Replacing the character ":" with a blank space (" ");
   (d) Replacing the character "?" with a dot (".");
   (e) Replacing the newlines ("\n") with a dot (".");
   (f) Splitting the text on all the dots.
2. Each requirement must be at least 31 characters long (30 >) to avoid that section indications within a privacy statement are in the list of requirements.

Note that the order of cleaning rules is important since the dots that do not indicate the end of a sentence should be removed before the text is split into separate requirements based on dots. The preprocessing steps implemented in Python shown in Notebook D.2 and the result of preprocessing the privacy requirements is shown in Appendix C.2.2.

4.3 Content Analysis

To create the training and test set, explained in Section 4.1, *content analysis* is applied. Content analysis is defined by Stemler (2001) as, among others, a technique for identifying characteristics of text. Here, components are considered to be the content analysis categories that "are a group of words with similar meaning".

The content analysis is performed following the guidelines and heuristics of the conceptual model, defined in Section 3.1 based on the literature described in Section 2.3.3, for annotating the right labels to the right word or set of words. Stemler (2001) describes six steps in performing content analysis:

1. Which data are analyzed?
2. How is the data defined?

3. What is the population from which the data is drawn?
4. What is the context relative to which the data are analyzed?
5. What are the boundaries of the analysis?
6. What is the target of the inferences?

We performed content analysis on six privacy statements where components are manually identified and categorized for each privacy statement. As described in the previous section, the privacy statements (samples) were chosen based on deliberately providing personal data by the consumers (data subjects) to an organization (controller). To narrow the scope, privacy statements of two types of organizations were chosen that meet the selecting criteria: social media and Webshop privacy statements. The goal is to create a training and a test set for developing the algorithm and testing the algorithm on a dataset that is not only tagged by the author, respectively.

A table was created by applying the preprocessing steps explained in the previous section (4.2.2) containing a list of requirements (sentences) together with a ID (i.e. the number of the requirement in the list) and the source (i.e. the source/owner of the original privacy statement). The following columns were added afterwards: valid (i.e. if the tagger considers the sentences to be a requirement, based on the presence of mandatory components) and a column for each components that needs to be tagged. Guidelines are composed and applied during the manually tagging process to keep consistency for both authors. Table 4.2 shows the guidelines in a table with the column names and a description.

4.3.1 Developing a Test Set

To test the algorithm on a test set that was not developed by the authors, a new test set is created. The new test set is established by asking three participants to manually tag the same data set consisting of 48 privacy requirement that are composed of two privacy statements: Instagram as a social media organization, and Lufthansa as Webshop.

The participants were all familiar with the term *privacy requirement* as they were either privacy experts from Deloitte (i.e. one senior privacy advisor, and one work student in privacy services), or a student in the Master Business Informatics at the Utrecht University. The participants were asked to perform the content analysis following the content analysis protocol shown Appendix C.3.

All columns except for the purpose and data refinement were tagged, due to the limited time of the participants. After the three content analyses the test sets are compared where after a *golden set* is created. The golden set is the combination of the three test sets done by choosing the best option for each tag. In other words, all input from the participants is compared, and the tags that differ in one of the test sets are discussed and the disagreement is solved.

Cohen's Kappa

The three data sets are compared using Cohen's Kappa, hereafter referred to as kappa. Cohen (1960) developed kappa (κ) as an "index of agreement for use with nominal scales" (Cohen, 1968, 1960). Equation 4.1 shows the calculation of κ "where p_o is the empirical probability of

Table 4.2: Explanation and manual for the content analysis procedure.

Column name	**Status**	**Type**	**Description**	**Desired input**
ID	required	integer	Identifier of the requirement.	automatically
source	required	string	Owner of the privacy statement (controller).	pre-filled value
valid	optional	integer	0 indicates a sentence that does not have the mandatory components; 1 indicates a useful privacy requirement.	0 or 1 (note: 1 should be filled in when the requirement needs to be included)
requirement	required	string	the requirement subtracted from the textual privacy statement.	pre-filled value
controller	required	string	The performer of the (processing) activity. Often 'we' or the name of the organization.	manually
requirement type	optional	string	Indicate the type of requirement (permission/prohibition/obligation).	manually (permission/prohibition/obligation)
keyword	optional	string	The modal verb that gives an indication for the requirement type. Only select the keyword if it is before the processing activity. For example: 'may', 'will', 'will not'.	manually
processing activity	optional	string	The activity performed on the personal data.	manually
personal data	optional	string	The data object consisting of nouns, it is the object that the activity is performed on.	manually
data subject	optional	string	The owner of the personal data. Often 'you' or 'your'.	manually (if applicable include 'the')
restriction	optional	string	A condition or prerequisite that influences the processing activity (indicated by 'when', 'if', etc.). Select all the words that are explaining the restriction.	manually (can be a 'big' part of the sentence)
purpose	optional	string	The reason for the processing activity. Select all the words that are explaining the purpose.	manually (can be a 'big' part of the sentence)
data refinement	optional	string	A refinement or specification of the 'personal data' object (indicated by 'such as..', 'for example..', 'including..', etc.). Make sure that the data refinement contains hypernyms of the personal data. For example: contact detail (personal data) can be name and email address (data refinement). Select all the words that are part of the data refinement.	manually (can be a 'big' part of the sentence)

agreement on the label (the observed agreement ratio), and p_e is the expected agreement when both annotators assign labels randomly".

Kappa κ is used to calculate the measurement of agreement between two observers. The calculations in Python are shown in Notebook D.4.

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{4.1}$$

κ is calculated separately for each component. The results of the comparison is shown in Figure 4.4 and in Table 4.3.

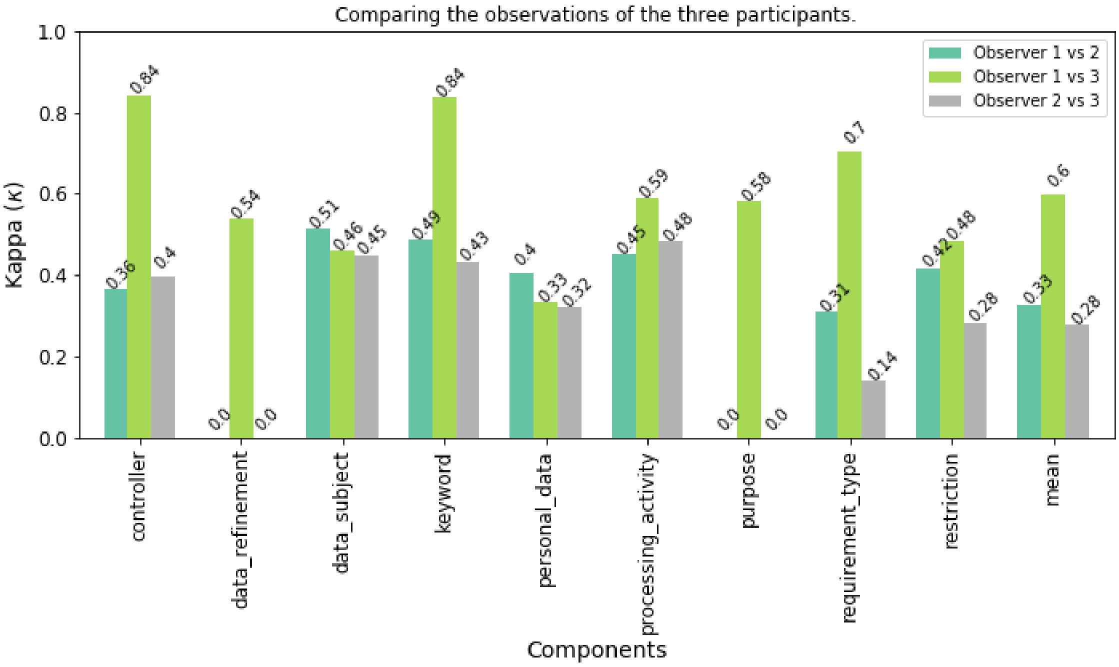


Figure 4.4: The comparison of the three taggers that participated in the content analysis for creating the test set.

Table 4.3: Kappa scores on the agreement between the three observers.

	Observer 1 vs 2	**Observer 1 vs 3**	**Observer 2 vs 3**	**mean**
controller	0.364839	0.842001	0.396985	0.534609
data_refinement	0.000000	0.538462	0.000000	0.179487
data_subject	0.513514	0.457627	0.448276	0.473139
keyword	0.486326	0.837563	0.431953	0.585281
personal_data	0.404035	0.331787	0.320644	0.352155
processing_activity	0.450046	0.588262	0.481690	0.506666
purpose	0.000000	0.579685	0.000000	0.193228
requirement_type	0.308108	0.704918	0.138810	0.383945
restriction	0.416667	0.484909	0.283582	0.395053
mean	0.327059	0.596135	0.277993	0.400396

The Golden Test Set

The pair of observers with the highest Kappa score per component will serve as a base for the golden test set. For example, for the controller component the kappa score of observer 1 and observer 3 is higher than another combination of two observers. However, when disagreements occur between the observers from the highest scoring pair, the input of the author will be used to decide the final answer. The observations of observer 2 will not be taking into account as it might lead to random deviations in the results.

4.4 Modeling the Algorithm

As shown in Figure 4.2 the development of the algorithm has two steps besides the content analysis: (1) reducing the noise in a requirement, and (2) applying text chunking on the requirement to subtract components.

Both steps are explained in general in this section. The specified approaches for each component will be explained in Section 4.7. However, first the preprocessing step is explained, which both have in common: POS-tagging.

4.4.1 Part-of-Speech Tagging

After the manual tagging in the previous section (4.3), the requirements need to be prepared for the text mining techniques that will be applied to label the components of the requirements automatically. We need to know some features of the textual requirements to determine the composition of the requirements. Part-of-Speech (POS) tagging is applied to label words with its type, such as: noun, verb, personal pronoun, etc. (Falessi et al., 2013). See also Section 2.2.2 for more explanation on POS tagging.

Before the words can be labeled they first need to be separated from each other, this is step is called *tokenization* and is explained in Section 2.2.2. Tokenization splits the words from a sentence in a set of separate words. The build-in NLTK `word_tokenize` function is used. This function splits a string into separate words based on spaces or punctuation (NLTK, 2017).

Next, the `pos_tag` function of NLTK takes the input of the tokenizer and tags each token with a *part-of-speech tag* (NLTK, 2017) (Section 2.2.2). See Listing 2.1 with an example output of the POS tagging function of NLTK.

The result of the POS-tagger is used as input for both of the following two steps.

4.4.2 Dependency Parsing

During the exploration of text mining techniques, text chunking in specific, a problem arises regarding the selection of the right words for the right components. The main requirements are often enclosed in phrases that deduce the main information of the requirement. For our own convenience, the controller, modality, processing activity, and data subject are referred to as the *main components*. The phrases that did not contain main components turned out to be, in most cases, the restriction, data refinement, or purpose that were also described in the privacy requirement.

In Example 4.4.1 the correct controller is “we”, the data subject is “you”, and the personal data will be “sign in details”.

Example 4.4.1 *When you create a personal account, we may also record your sign in details.*

To help the text chunking grammar select the main components from the main requirement phrase, a new text mining technique is applied as a cleaning step before text chunking: *dependency parsing*. This cleaning step means the subtraction of the phrases, like the restriction and the data refinement, that do not contain the main components. As an example (4.4.1) text chunking might select “a personal account” as personal data. However, the phrase “When you create a personal account” is part of the restriction component. Text chunking and component identification is explained in more detail in the next section (2.2.2).

Dependency parsing was found as a text mining technique that identifies the relationship between words and groups of words. It turns out that the restriction and the data refinement can be identified by labels given by the spaCy dependency parser. SpaCy is a Python package that indicates dependencies in written sentences. According to Choi et al. (2015) the spaCy dependency parser is the fastest parser from the ten parsers they tested. According to a comparison research of different tools for dependency parsing conducted by Choi et al. (2015) spaCy performs on average regarding accuracy.

To solve the problem of having too much to choose for the text chunking grammar, the restriction and data refinement component are removed from the requirement after identification by the dependency parser. Now, the text chunking grammar has less words or phrases to choose from regarding the selection of the main components (see Appendix C.2.2 for the cleaned privacy requirements).

After loading a language processing pipeline, specified for English text in the case of this project, each requirement is converted from a string to a `doc` object that enables accessing linguistic annotations for the object. SpaCy uses *head* and *child* to indicate the head of a (sub)tree and the children within that (sub)tree. To stay with the NLTK terminology, the term *parent* will be used in the report instead of the term head. With spaCy we are able to define the POS-tag (`.pos_`), the dependency type (`.dep_`), and the parent (head) of a word (`.head`).

Figure 4.5 shows a *dependency parse tree* that is created by the spaCy dependency parser.

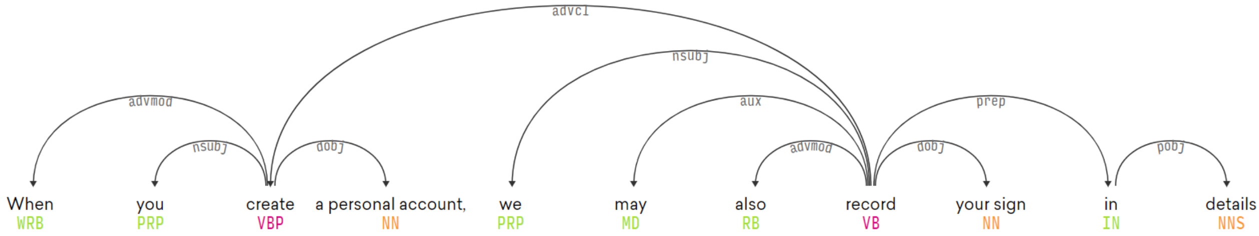


Figure 4.5: A dependency tree of a requirement (see Example 4.4.1) created by spaCy and visualized by displaCy.

Two types of dependencies are found that identify the restriction and the data refinement: adverbial clause modifier and prepositional modifier, respectively. Both labels are explained in the following sections.

Adverbial Clause Modifier

De Marneffe & Manning (2008) describe an *adverbial clause modifier* in the Stanford type dependencies manual as a clause that modifies a verb. It can indicate a temporal, conditional, or purpose clause or a consequence. In the scope of this project it is used as a tag to indicate the restrictions (or conditions) in privacy requirements.

The goal of identifying the adverbial clause modifier is to filter out the noise that is created by the restriction phrase of the sentence (i.e. the nouns and verbs that also indicate controllers, data subjects or activities), and therefore improve the accuracy of our algorithm.

SpaCy labels the adverbial clause modifier with the `advcl` tag, which is assumed to indicate the restriction in the privacy requirements. By removing the subtree that is indicated by spaCy as the restriction thereby excluding the the words that are part of the restriction phrase we have a smaller and more clean set of the sentences left.

The result of removing the adverbial clause modifier from Example 4.4.1 is shown in Figure 4.6.

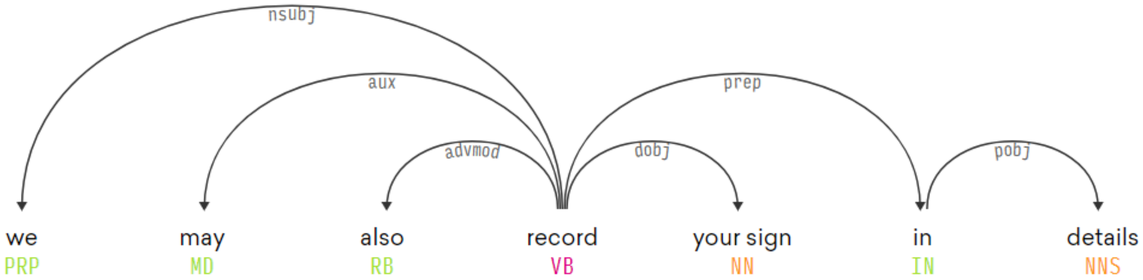


Figure 4.6: The dependency tree of the requirement (see Example 4.4.1) without the adverbial clause modifier dependency created by spaCy and visualized by displaCy.

Prepositional Modifier

To reduce even more noise the data refinement phrase needs to be subtracted from the original requirement. The prepositional modifier is assumed to be the data refinement phrase. Again, spaCy is used to create a dependency parse tree and to remove the subtree of the requirement labeled with: `prep`. As a result that the requirement is reduced even further.

Example 4.4.2 is used to show the effect of identifying and removing the prepositional modifier form a requirement with a data refinement.
Example 4.4.2 *We may also record your sign in details such as your name and email address.*

Figure 4.7 shows the dependency parse tree of Example 4.4.2. Note that the parse tree contains two `prep` dependency annotations. Figure 4.8 shows the result of removing the prepositional modifiers from the initial requirement. A new problem arises here since the dependency parser removed too much from the requirement. Ideally, only the part “such as your name and email address” would be removed, however, “in details” is also removed. This way the text chunking grammar is not able to identify “sing in details” as personal data. We will discuss our final approach in Section 4.7.6.

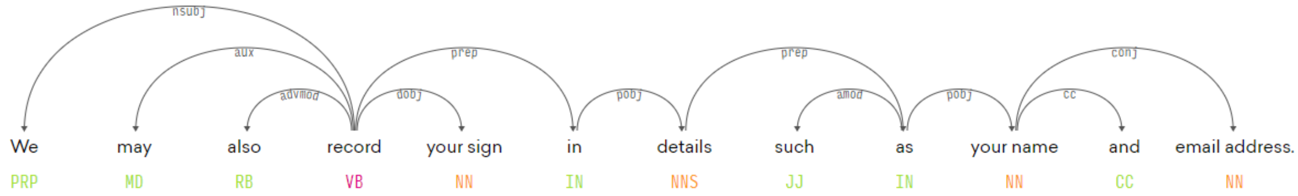


Figure 4.7: A dependency tree of a requirement (see Example 4.4.2) created by spaCy and visualized by displaCy.

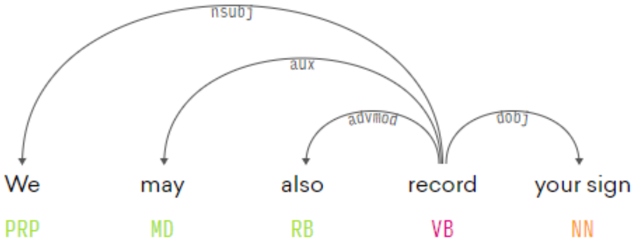


Figure 4.8: A dependency tree of a requirement (see Example 4.4.2) without the prepositional modifier dependency created by spaCy and visualized by displaCy.

4.4.3 Text Chunking

Text chunking is the process of labeling non-overlapping text segments (Arora et al., 2015) based on omitting whitespace (Bird et al., 2009). These segments are called *chunks* and are a set of words that are grouped based on their POS-tag (Bird et al., 2009). For the correct labeling of the chunks a pattern is created. This pattern is called a *chunk grammar* and is a set of "rules that indicate how sentences should be chunked" (Bird et al., 2009). This means that after some preprocessing steps a pattern is created by simply trying to improve the pattern until the algorithm performs sufficiently. The `RegexpParser` chunker of the NLTK packages is used. This means that the grammars created are based on regular expression rules (Bird et al., 2009). The output of text chunking is a *constituency parse tree.*

The pattern is composed by starting with a basic pattern that only extracts a basic Noun Phrase (NP) and the controller from the sample requirement. Each adjustment on the pattern is tested by running a sample requirement through a POS-tagger (function) followed by applying the pattern to the requirement in the text chunker (function). This is an iterative process (see Figure 4.9) where the first created pattern is tested and adjusted accordingly.

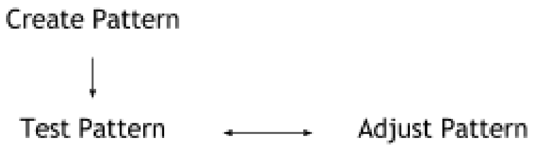


Figure 4.9: Iterative process of creating the most optimal pattern.

Figure 4.10 shows an example for the constituency parse tree of Example 4.4.1 and Figure 4.11 for the tree that is first cleaned by the spaCy dependency parser explained in the previous section.

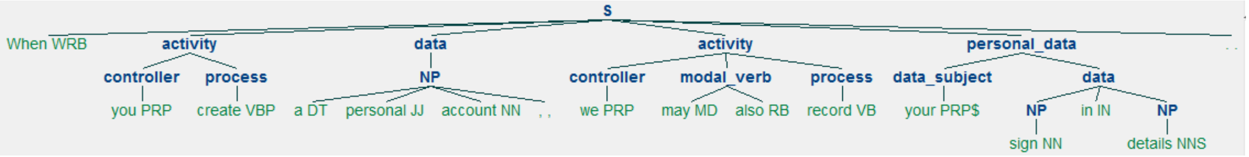


Figure 4.10: A constituency parse tree drawn for the requirement in Example 4.4.1.

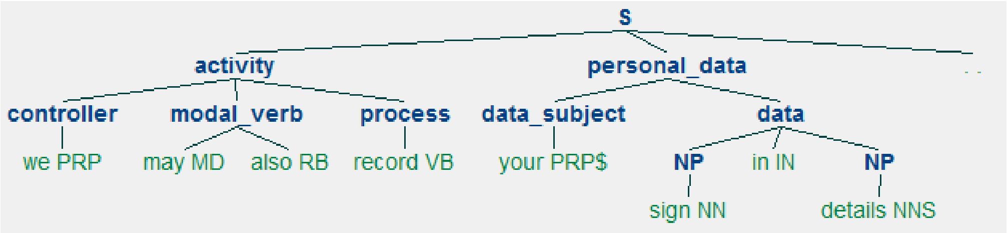


Figure 4.11: A constituency parse tree drawn for the requirement with reduced noise in Example 4.4.1.

4.5 Testing Scores

The goal of the iterative process is finding the best performing pattern, which means creating a pattern that is able to cover as much of the requirement as possible without making errors. Errors are wrongly tagged or grouped text chunks. Optimal, or making no errors, means that the algorithm can identify as much text chunks as possible in the right way. This is tested by counting the number of true positives and false positives.

In order to determine the performance, three metrics are used: *precision*, *recall*, and *f-score*. The precision is defined as the number of predicted positives that actually have the predicated positive condition (T_p) divided by the total number of positive predicted values ($T_p + F_p$). This also includes the predicted values that in real life where not positive (see Equation 4.2).

$$precision = \frac{TP}{TP + FP} \tag{4.2}$$

In the context of mining components from privacy requirements, we can explain precision as the ratio of the correctly identified components over the total number of identified components. This means that we can assert how well the algorithm predicts the components. However, it does not say anything about the total number of cases that were predicted as negatives. *Recall* looks at the number of correctly identified components and compares that number to the total number of predicted elements (see Equation 4.3).

$$recall = \frac{TP}{TP + FN} \tag{4.3}$$

An algorithm with a high recall returns more results than an algorithm with a higher precision. Additionally, together with precision and recall, the f-score (or f-measure) is calculated. It test the accuracy of the algorithm. The F_β score is used as f-measure and is calculated with the following Equation (4.4):

$$F_\beta = \frac{(1+\beta^2) \times recall \times precision}{recall + \beta^2 \times precision} \tag{4.4}$$

Here, β is set on 1 as it weights recall more than precision by the factor of β and *precision* and *recall* are calculated as shown in Equation 4.3 and 4.3.

The function of the `sklearn` Python package `precision_recall_fscore_support` calculates the *precision*, *recall*, and F_β. It calculates the precision, recall, and f-score for each class. The function takes the manually tagged components (true values), the components labeled by the algorithm (predicted values), and a parameter *average* that determines the type of averaging performed on the data. The average parameter is set on `weighted`, which is the default setting. It takes the average weight based on the support (the number of true instances for each label).

From the interviews (3.2.1) we learned that a useful application of the conceptual model could be a model that can be used as starting point for a data inventory. Also, if the data inventory does exists it can be used to align the privacy statement with the data inventory or the other way around. With this usefulness and applicability in mind the result of the model needs to be as accurate as possible. Since the data inventory is used to account for the *Dutch Data Protection Authority* (Dutch DPA) it needs to be maintained without errors. Therefore, from a Privacy Expert point of view, we want the results based on a high precision score.

4.6 Readability Scores

As mentioned before, this project analyses privacy statements that are written in natural language. Since writing privacy statements is not based on rules, a difference in readability can be noticed when reading different statements. To test if the perceived readability by human readers has an influence on the performance of the algorithm, existing readability indexes are used to measure the readability of the used privacy statements. The following readability indexes are used:

- Flesch Reading Ease (FRE) (see Section 2.2.3),
- Automated Readability Index (ARI),
- Dale Chall Readability Index (DCRI), and
- Flesch Kincaid Grade Level (FKGL).

The indexes are calculated based on the total number of both words and unique words, number of syllables per word, number of characters, number of sentences, and the number of words per sentence. The formulas are shown in Section 2.2.3 and/or in the notebook in Notebook D.5.

4.7 The Final Algorithm

Each component is extracted or subtracted in a different way. This section explains the extraction method for each component.

```
grammar8 = r"""
        # controller
        controller: {<PRP>}

        # processing activity
        modal_verb: {<MD><RB>?<RB>?}
        process:
           {<VB.*><RP>?<CC|,|:>?<VB.*>?<RP>?<,|:>?<CC>?<VB.*>?<RP>?}
        activity: {<controller><modal_verb>?<process><RB>?}

        # data
        data_subject: {<PRP$>}
        NP: {<DT|JJ|NN.*|,>+}
        personal_data: {<NP>?<CC>?<NP><IN>?<NP>?<CC>?<NP>?}
        data: {<data_subject><personal_data>}

        # purpose
        purpose:
           {<TO><process><controller|modal_verb|process|activity|
                NP|data_subject|personal_data|data|,|:|IN|RB>+}
        """
```

Listing 4.1: The result of the process of finding a pattern for text chunking.

Figure 4.12 shows the structure of the grammar: each component consists of one or more POS tags, also the leaves. The result of an applied pattern (i.e. grammar) is a sentence that is turned into a list with sublists with tuples. Figure 4.10 shows the result of applying the pattern from Pattern 4.1 on the requirement from Example 4.4.1.

4.7.1 Controller

To find the controller in a requirement we simply start looking for the personal pronouns in the requirement. A personal pronoun is indicated by POS-tagging with ‘PRP’.

According to templates found in literature the controller is often followed by the activity that is performed on the personal data (Breaux & Antòn, 2005; Breaux et al., 2014). Therefore, instead of only looking for the PRPs, the newly defined grammar searches for combinations of different POS-tags to determine the difference between the controller and the data subject since the data subject is also tagged by the POS-tagger as a PRP. In other words, text chunking creates a tree based on a combination of words that will form the controller, modality, and processing activity. From now on this combination is referred to as the activity subtree.

The combination is important since multiple personal pronouns can be in the requirement (e.g. ‘you’ that might indicate the data subject). In short, the grammar extracts this combination from the requirement, hereafter the controller is extracted from this subtree.

However, it is possible to find multiple activity subtrees in one requirement. Rules are set to pick the most likely controller from the various activity subtrees. While looping through the tagged subtrees, the following rules are applied for choosing the right controller:

1. The controller is chosen from the subtree if no controller is chosen yet (i.e. the controller

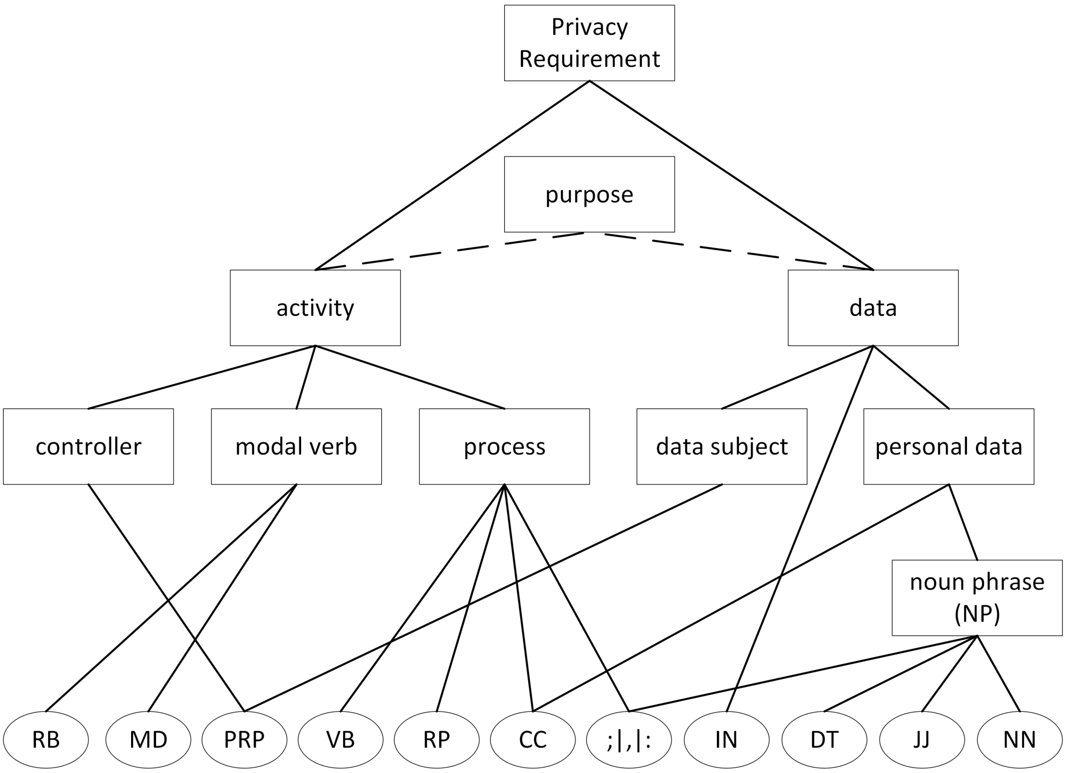


Figure 4.12: The grammar consist of a defined hierarchical structure with the POS tags as lowest level.

variable/object is empty);

2. The controller is chosen from the subtree if the current chosen controller is equals 'you', since we assume that the controller is most likely referred to by the word "we". This assumption is based of the number of occurrences of the word "we" ($n(we) = 138$) compared to the word "you" ($n(you) = 15$) on the total number of privacy requirements in both training and test set ($N = 194$)(see Figure 4.13).

The applied rules can be found in Notebook D.3.

4.7.2 Modality

An indicator of the modality is first extracted by focusing on one POS-tag, the modal (MD). When the modal, a type of verb, is extracted from the activity subtree, the modal verb will be compared to a list of modals that are linked by Demsetz (1974) to specific modalities (see Table 4.4. The modality that fits the modal verb will be assigned to the modality component of that requirement. If no modal verb is found by the algorithm, assumed is that the modality is a permission.

4.7.3 Processing Activity

The last component that is extracted based on the activity subtree is the processing activity. For the processing activity the algorithm searches for a combination of POS-tags that is defined by

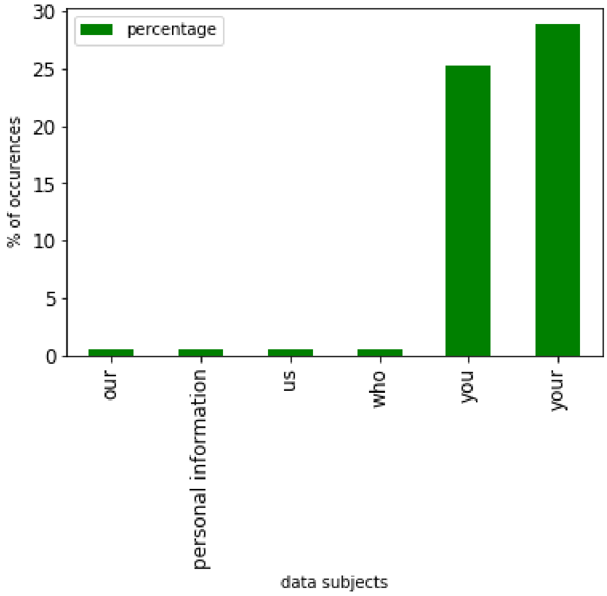


Figure 4.13: The distribution of the counted unique data subject found during content analysis

the text chunking pattern. The combination of POS-tags is compiled to cover various patterns. The following examples are covered by the grammar (see Listing 4.1, line 7):

- Single verb ("collect"),
- Single verb with a particle ("Looking for"), and
- "And", ",", and ";" (when multiple activities are applied, such as "collect; process, and access").

As with the modality, it is possible to find multiple activity subtrees in one requirement. Rules are set to pick the most likely processing activity. While looping through the tagged subtrees, the following rules are applied for choosing the right processing activity:

1. The activity process is chosen from the subtree if no activity process is chosen yet (i.e. the activity process variable/object is empty),
2. The word length of the activity process must be at least 3 characters long.

This means that short verbs such as "do" and "be" will not be assigned to the processing activity component. The applied rules can be found in Notebook D.3.

4.7.4 Data Subject

Besides the activity subtree another subtree is defined to extract the data subject among others. This subtree is called data. Data combines POS-tags indicating the data subject and the personal data.

The grammar pattern `data_subject:{<PRP$>}` was part of subtree:
`data:{<data_subject><personal_data>}` (see the first line of Example 4.1).

We assume that all the data objects are referred to as 'you'. This means that the *y_pred* is always 'you', and is tested towards the actual value of the data subject (*y_true*) which is, when

Table 4.4: Overview of modal verbs linked to a modality, based on Demsetz (1974)

modal verbs	**modality**
does not have a right to	prohibition
does not	prohibition
do not	prohibition
has a right to	permission
is not required to	anti-obligation
may	permission
may deny	permission
may not	obligation
may not require	obligation
may require	permission
must	obligation
must deny	obligation
must permit	obligation
must request	obligation
retains the right to	permission
could	permission
can	permission
will not	prohibition

labeled with ‘your’, also changed to ‘you’. This assumption is based on the results from the content analysis where $n(you) = 49$ and $n(your) = 56$ from the total of 194 requirements (see Figure 4.14.

Different options are applied to find the best performing combination of techniques.

4.7.5 Personal Data

Personal data seems to be the biggest challenges since it covers a variety of words with different annotations. The challenge seems to be the choice between two approaches: (1) cleaning the privacy requirement by removing the data refinement phrase, or (2) not removing the data refinement phrase and optimizing the text chunking pattern. When removing the data refinement phrase (option 1), as described in the previous section, the dependency parser does not only remove the phrase indicated as the data refinement by the literature. In addition, some other sets of words are removed that do have the `prep` dependency annotation (see Section 4.4.2 for example 4.7 and 4.8). What happens is that the dependency parser removes too much text, and therefore the personal data phrase is not always complete in the cleaned requirement.

However, option 2, did not improve the performance of the algorithm and is therefore not chosen. Additionally, removing the `prep` sub-sentence from the privacy requirement, did improve the algorithm with regard to the tagging of the data subject.

To conclude, the prepositional modifier was removed from the privacy requirement before text chunking was applied to identify the personal data. The personal data component was then identified by the following text chunking grammar `personal_data:{<NP>?<CC>?<NP><IN>?<NP>?<CC>?<NP>?<process>?}` (4.1, line 13) as part of the data subtree.

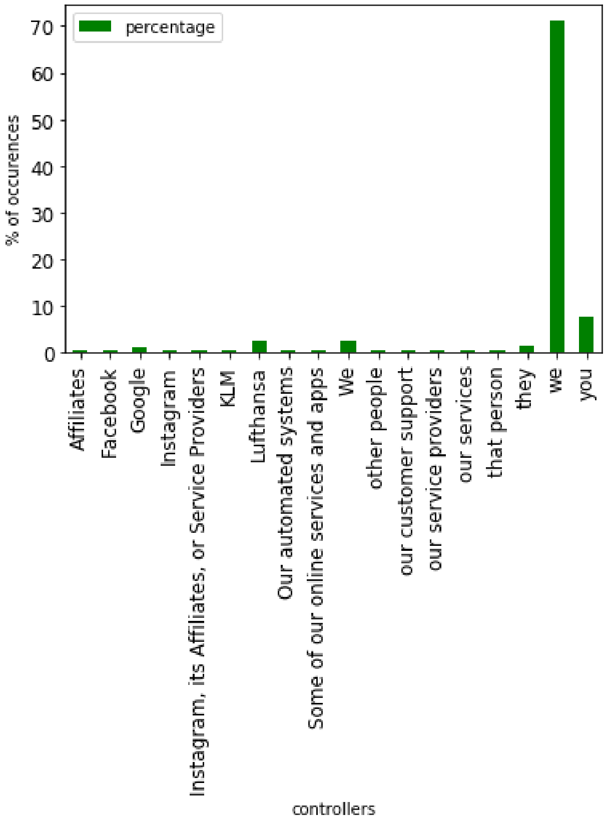


Figure 4.14: The distribution of the counted unique data subject found during content analysis

4.7.6 Data Refinement & Restriction

As explained in Section 4.4.2, dependency parsing is used to clean the privacy requirements from noise. We also saw the dependency types that were considered to be the noise: adverbial clause modifiers (`advcl`) and prepositional modifiers (`prep`). The spaCy dependency parser is used to remove both dependencies from the privacy requirement.

The algorithm applies the dependency parser according to the following steps:

1. A lookup object is created to map all the separate words from the privacy requirements together with their dependency type and the index of the parent (see Figure 4.15 for Example 4.7.1).

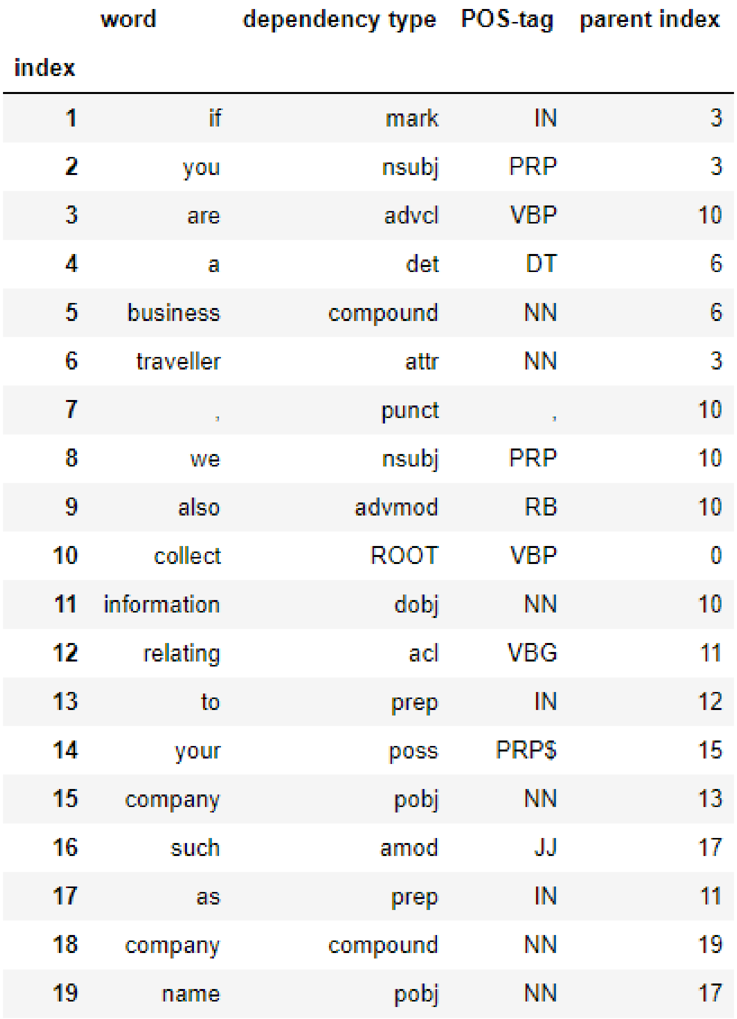
2. All parents with the dependency type that we want to remove from the initial requirement (`advcl` or `prep`) are now selected. In the case of our example the index of dependency parent for `advcl` is 3 ('are') and the indexes for `prep` are 13 ('to') and 17 ('as').

3. Now that the parents are established, its children need to be found since the whole sub-sentence needs to be removed. All words correspondent to a parent index belong to this parent and are considered its children. This step is repeated within the parent index until all the children of the dependency subtrees are identified.

4. The last step is to remove the sub-sentences that that were identified as a dependency subtrees.

Example 4.7.1 *If you are a business traveler, we also collect information relating to your* *company such as company name.*

The result of the privacy requirement in Example 4.7.1 will be: “we also collect information relating”. Note that the dependency parser removed too much text as “to your company” is also subtracted. However, this side effect is accepted since no other modification of the dependency parser was found that did improve the performance of the algorithm as much as the previous explained application.



	word	dependency type	POS-tag	parent index
index				
1	if	mark	IN	3
2	you	nsubj	PRP	3
3	are	advcl	VBP	10
4	a	det	DT	6
5	business	compound	NN	6
6	traveller	attr	NN	3
7	,	punct	,	10
8	we	nsubj	PRP	10
9	also	advmod	RB	10
10	collect	ROOT	VBP	0
11	information	dobj	NN	10
12	relating	acl	VBG	11
13	to	prep	IN	12
14	your	poss	PRP$	15
15	company	pobj	NN	13
16	such	amod	JJ	17
17	as	prep	IN	11
18	company	compound	NN	19
19	name	pobj	NN	17

Figure 4.15: An example of a lookup table for one privacy requirement.

4.7.7 Purpose

At first, it seems very difficult to find a suitable text mining technique for extracting the purpose from a privacy requirement. However, after some analysis it seemed that the purpose is often indicated by the word ‘to’ followed by a verb. After the combination of the word ‘to’ and a verb, almost all combinations of word annotations can be found. Since the a verb was already defined as a ‘process’, the ‘process’ label was included. The following text chunking pattern was created: `purpose:{<TO><process> <controller|modal_verb|process| activity|NP|data_subject| personal_data|data|,|:|IN|RB>+}` (see row 17 from Listing 4.1). The whole set of words retrieved by the text chunking patterns is considered to be part of the purpose.

Chapter 5

Evaluating the Results

5.1 Algorithm

In this result section the used text mining techniques are evaluated by assessing their output based on measurement scores explained in Section 4.5 To test the generalization of the text mining techniques the algorithm is tested on a test set created by three observers (see Section 4.3). The test is applying the algorithm on the test that is created by three observers that are known with the field of Privacy or Requirement Engineering. The test had the same preprocessing steps as the training set. The results of both the training and the test set are explained in this section.

As explained in Section 4.5 the precision score is used as a leading factor for evaluating the found combinations of text mining techniques. Figure 5.1(a) shows the precision score for each each component per privacy statement. Additionally, all the scores are shown in Table 5.2. The f-score is calculated with $\beta = 1$ which means that precision and recall are equally weighted in the calculation of the f-score.

The Pearson Correlation Coefficient (r) is calculated with the `pandas.DataFrame.corr` function to determine the relationship between the performance of the text mining techniques on the training set and the test set. The Pearson correlation is $r = 0.892$ for the performance of the algorithm on the training and test.

Table 5.1 shows the standard deviation and mean for each component in both the training and the test set. The standard deviation is calculated by the following function from the pandas package in Python: `pandas.DataFrame.std`. The smaller the standard deviation, the less variation between the scores and the average score (i.e. the mean). An outlier the data subject in the test set where the standard deviation ($SD = 0.182$) is significantly higher than the standard deviation of the other components. This means that, on average, the results of the text mining techniques applied on the data subject component are relatively far away of the mean.

The controller is the component that is predicted by, the text mining technique, text chunking ($\mathit{precision} = 0.697, \mathit{recall} = 0.752, \mathit{f\text{-}score} = 0.712$). It is the stakeholder that is responsible for the processing activity in the privacy requirement.

The modality is the component that is best predicted. Text chunking is used to extract modal verbs as indicators for the modality. Keywords are words like: ‘will’, ‘may’, and ‘must not’.

Table 5.1: The standard deviations .

set	component	mean	std
training	controller	0.920510	0.003490
	keyword	0.892623	0.012174
	modality	0.980840	0.000853
	processing_activity	0.841145	0.011881
	data_subject	0.794554	0.011918
	personal_data	0.319826	0.020894
	data_refinement	0.688537	0.082012
	restriction	0.611191	0.007469
	purpose	0.472053	0.054303
test	controller	0.866938	0.003690
	keyword	0.946966	0.009326
	modality	0.945253	0.010182
	processing_activity	0.670546	0.004097
	data_subject	0.606346	**0.182000**
	personal_data	0.210097	0.027364
	data_refinement	0.751973	0.071942
	restriction	0.551986	0.008582
	purpose	0.542110	0.064221

The text chunking grammar of the modality keywords performs as second best ($precision = 0.808, recall = 0.821, f\text{-}score = 0.809$). Based on Table 4.4 of Demsetz (1974) the keywords are linked to one of the three modalities (permission, prohibition, or obligation). When no modality is assigned to the privacy requirement, due to no detection of the modality keyword, we assume that the modality of the requirement is a 'permission'. Based on the identification of the keywords with text chunking and the assignation the matching modality the $precision = 0.897, recall = 0.935, f\text{-}score = 0.912$. The difference in performance between the modality keyword and the modality itself can be explained by the assumption of the privacy requirement being a permission when no keyword is detected.

The processing activity explains what happens with the personal data in the privacy requirement. Anton et al. (2004) used the term *goal* to indicate the processing activity, as the requirement describes a goal that can be achieved when the permission is given. The processing activity is extracted through text chunking ($precision = 0.628, recall = 0.631, f\text{-}score = 0.627$).

Personal data is the component where, of all components, the text mining techniques perform the least. From the 194 privacy requirements that are analyzed (both training and test set) only 52 are extracted correctly by the text mining techniques ($precision = 0.274, recall = 0.223, f\text{-}score = 0.231$). As mentioned before in Section[dependency_parsing]. The dependency parser identifies too much words as part of a data refinement phrase. This results in personal data objects that are not complete (see the Examples given in Section 4.4.2).

The owner (Van Alstyne et al., 1995) or custodian (Cheong & Chang, 2007) of the personal data is referred to as the data subject. The data subject is extracted with text chunking in combination with the personal data component ($precision = 0.749, recall = 0.681, f\text{-}score = 0.671$).

The purpose is considered to be an important part of a privacy requirement by the interviewees

Table 5.2: Performance of the final combination of text mining techniques per source for each component.

	source	**KLM**	**Facebook**	**Google**	**Lufthansa**	**Instagram**
	component					
precision	controller	**0.918239**	0.868530	0.866000	0.210204	0.621701
	keyword	0.881551	**0.937888**	0.847500	0.696429	0.676853
	modality	0.981509	0.936715	0.830638	0.738095	**1.000000**
	processing_activity	**0.853774**	0.671739	0.690667	0.500000	0.425806
	data_subject	0.807383	0.816496	0.626445	**0.885714**	0.609791
	personal_data	0.342767	0.184783	**0.570000**	0.035714	0.236559
	data_refinement	0.609942	0.682420	0.471837	**1.000000**	0.662442
	restriction	0.618711	0.560641	**0.735556**	0.630952	0.694789
	purpose	0.419932	0.480549	0.505128	**0.785714**	0.501792
recall	controller	**0.924528**	0.869565	0.900000	0.357143	0.709677
	keyword	0.905660	**0.956522**	0.820000	0.714286	0.709677
	modality	0.981132	0.956522	0.880000	0.857143	**1.000000**
	processing_activity	**0.830189**	0.673913	0.700000	0.500000	0.451613
	data_subject	0.792453	0.500000	0.640000	**0.857143**	0.612903
	personal_data	0.301887	0.239130	**0.440000**	0.071429	0.064516
	data_refinement	0.773585	0.826087	0.680000	**1.000000**	0.741935
	restriction	**0.603774**	0.543478	0.540000	0.571429	0.516129
	purpose	0.528302	0.608696	0.620000	**0.714286**	0.451613
f-score	controller	**0.918762**	0.862719	0.877143	0.250000	0.652842
	keyword	0.890658	**0.946488**	0.826374	0.692331	0.688377
	modality	0.979880	0.942523	0.846142	0.792208	**1.000000**
	processing_activity	**0.839473**	0.665985	0.691683	0.500000	0.437276
	data_subject	0.783827	0.502541	0.610478	**0.850794**	0.608768
	personal_data	0.314825	0.206377	**0.484000**	0.047619	0.101382
	data_refinement	0.682086	0.747412	0.557108	**1.000000**	0.699939
	restriction	0.611090	0.551839	**0.614894**	0.598901	0.588710
	purpose	0.467925	0.537084	0.556000	**0.748299**	0.475382

as it is the reason for processing the personal data. It seems a complex component as it consists of various word annotations assigned by the POS tagger. The best performing text chunking grammar scores as follows: $\mathit{precision} = 0.539, \mathit{recall} = 0.584, \mathit{f\text{-}score} = 0.557$.

The phrase of the privacy requirement indicating additional information about the personal data is covered in the data refinement component. The data refinement component is associated with the prepositional modifier that can identified with dependency parsing ($\mathit{precision} = 0.685, \mathit{recall} = 0.804, \mathit{f\text{-}score} = 0.737$).

The restriction is a clause with a condition that needs to be fulfilled for the privacy requirement to be applicable. The restriction is identified with adverbial clause modifier annotation of the spaCy dependency parser. Dependency parsing performs as follows for the restriction: $\mathit{precision} = 0.648, \mathit{recall} = 0.555, \mathit{f\text{-}score} = 0.593$.

5.2 Readability

To assess the readability of the privacy statements three scores are used: ARI, FKGL, and the DCRI. These scores are used because they have in common that they map the readability of a written text by (US) grade level. The range of scores (i.e. the grade levels) varied per readability index and therefore the scores could not be normalized.

The readability scores of each source are shown in Table 5.3. The Instagram privacy requirements are considered most difficult to read as for all three scores Instagram scores the highest grade level.

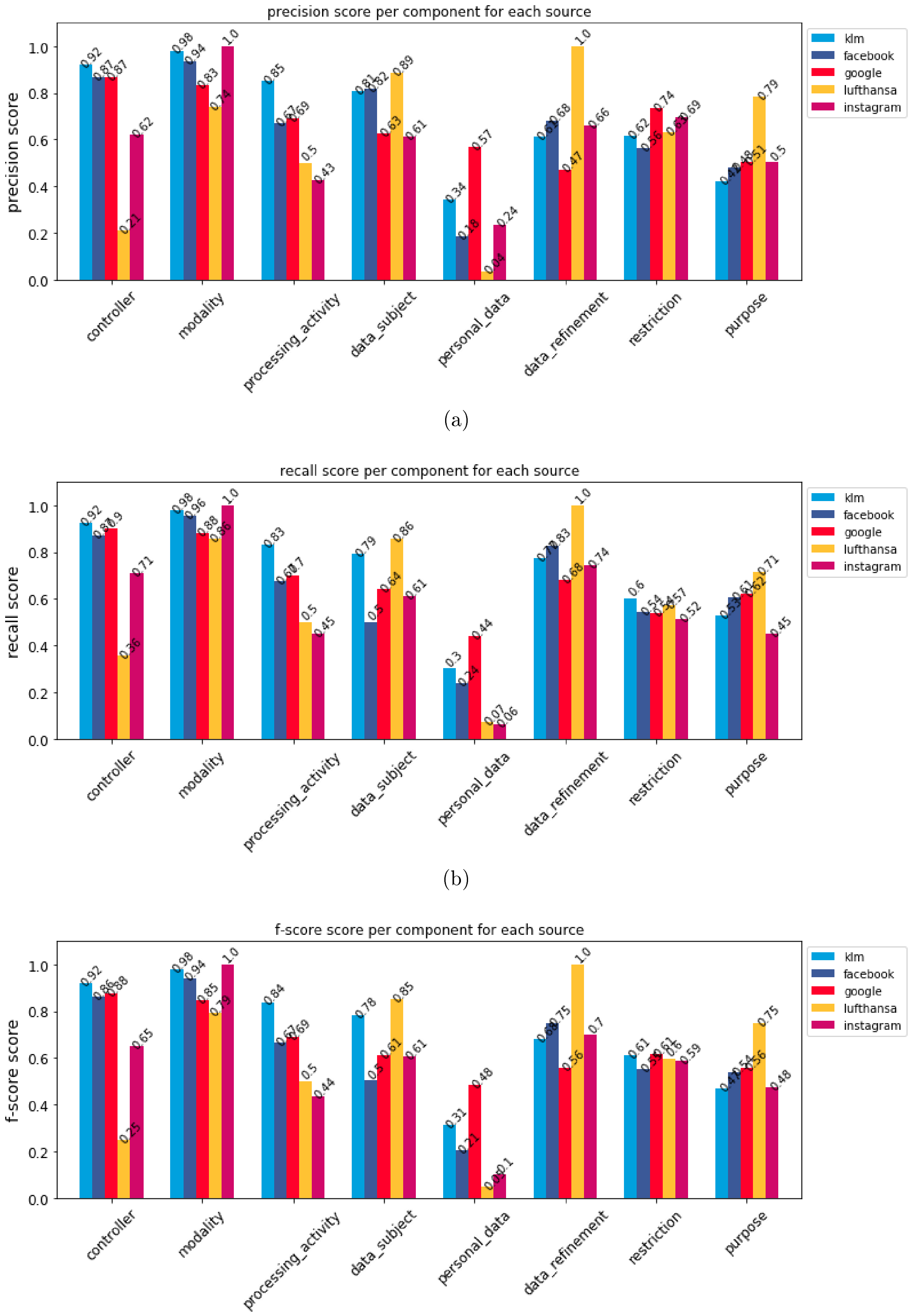
Table 5.3: Readability scores for each source.

source	**Facebook**	**KLM**	**Google**	**Instagram**	**Lufthansa**
ARI	15.6200	7.9100	14.1200	**17.6200**	12.120000
FKGL	6.7400	4.4000	5.5700	**8.3000**	4.010000
DCRI	1.3392	1.0416	1.1904	**1.5376**	0.992000

Figure 5.2 shows the correlation between the readability indexes. All indexes have a good relationship, where $r = 1$ for the relationship between DCRI and FKGL, and $r = 0.86$ for the relationship between ARI and FKGL, and ARI and DCRI.

5.3 Performance Measurements vs. Readability Indexes

Figure 5.3 shows the relationship between the performance measurements used to assess the performance of the text mining techniques and the readability indexes that are used to assess the readability of the sources used. From the three readability indexed used, it seems that ARI predicts the performance of the algorithm best as the ARI score seems to be low when the performance measurement scores are high ($r = -0.71$ for recall and the f-score, and $r = -0.56$ for recall).



(a)

(b)

(c)

Figure 5.1: Performance measurements for each source per components: (a) precision, (b) recall, and (c) f-score.

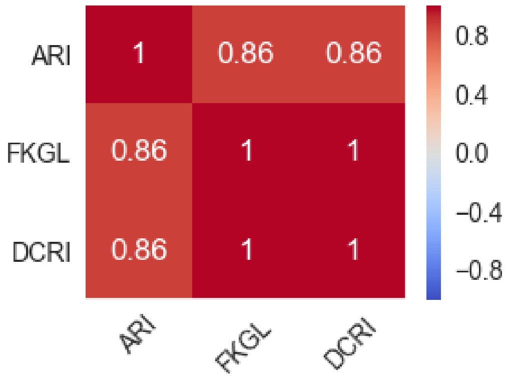


Figure 5.2: Correlation of readability indexes used to asses the readability of the used privacy statements.

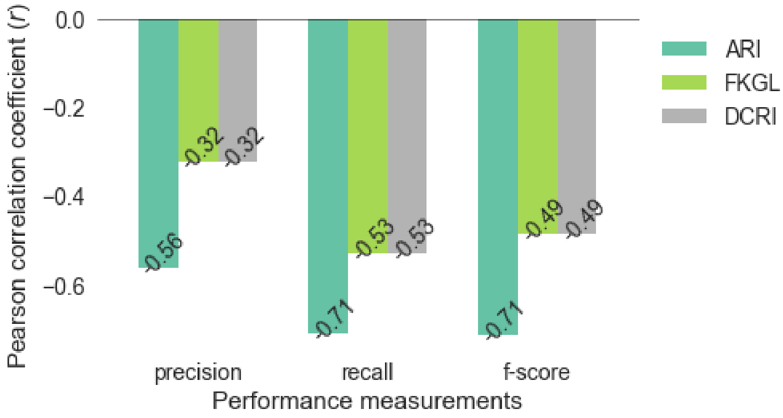


Figure 5.3: Pearson's r for determining the relationship between the performance measurements and the readability indexes.

Chapter 6

Conclusion

In this chapter the research questions, defined in Section 1.1, are answered based on the literature study conducted in Chapter 2, and the exploitative research and its results described in Chapter 4 and 5.

Question 1 *What elements does a good privacy requirement consist of?*

To answer the first research Question 1 a literature study was conducted in the first part of the project (Chapter 2). Here, the elements were identified by combining the field of requirements engineering and linguistic science extended with the field privacy requirement engineering. The components found in privacy statements according to the literature were mapped in a conceptual model (3.1). Afterwards the conceptual model was revised based on interviews conducted with four Privacy experts from Deloitte Privacy Services (see Figure 3.2 in Section 3.2).

The components of the revised model were the following: controller, modality (requirement type), processing activity, personal data, data subject, purpose, data refinement, and restriction. Out of those eight components, based on the literature and expert opinion, the mandatory components of a privacy requirement are processing activity, personal data, and a purpose for the privacy requirement to be complete.

Question 2 *What text mining technique can be used to automatically identify the main components of a privacy requirement?*

A combination of dependency parsing and text chunking seems the combination of text mining techniques that works best for the used cases (a combination of dependency parse trees and constituency parse trees) with respect to the techniques that are tested.

With dependency parsing the two components were extracted and subtracted from the privacy requirements. The spaCy dependency parser was able to identify two types of dependency annotations that were linked to the data refinement (as prepositional modifier) and restriction (as adverbial clause modifier) component (4.4.2).

The identification of the word annotations was followed by removing both phrases from the initial requirement. For both training and test set, on average, the dependency parser reduces length of a requirement with 39.98%. Reducing the length of the initial privacy requirement has a positive influence on the identification of most components. Only the scores for the controller and the purpose slightly decreases. The modality is not influenced. For the remainder the score increases. Overall, without removing the restriction and the data refinement the text mining techniques score: $precision = 0.649, recall = 0.634, f\text{-}score = 0.631$. When the phrases are removed the average score over all the components is: $precision = 0.658, recall = 0.664, f\text{-}score = 0.649$. To conclude, the subtraction of the restriction and data refinement phrase does increase the overall performance of text mining techniques.

After cleaning the requirement from what can considered to be noise by the rest of the components, text chunking is applied. Text chunking uses grammars, based word annotations, to identify words or groups of words as a category.

The iterative process of testing and improving the text chunking grammars have lead to a final grammar that covers: controller, modality (requirement type), processing activity, personal data, data subject, purpose components.

Question 3 *Is the method used for categorizing text elements from privacy policies effective for the used cases?*

The dependency parser performed for data refinement as follows: $precision = 0.685, recall = 0.804, f\text{-}score = 0.737$, were the restriction was identified with the following scores: $precision = 0.648, recall = 0.555, f\text{-}score = 0.593$.

As explained before for the main components and the purpose the text mining technique text chunking is used. The average score for all the components for the data set is $precision = 0.658, recall = 0.665, f\text{-}score = 0.650$.

The results per component are shown in Table 6.1. Not every component is easy to recognize with the used techniques. The component with the lowest measurement scores is the personal data ($precision = 0.274, recall = 0.223, f\text{-}score = 0.231$). However, personal data is considered to be one of the most important components in a privacy requirement according to the interviews. Since, the privacy requirement is written in the privacy statement for the organization to give themselves the right of processing the personal data.

Table 6.1: Performance of the text mining techniques for each component in descending order based on the precision.

	precision	recall	f-score
component			
requirement_type	0.897392	0.934959	0.912151
keyword	0.808044	0.821229	0.808846
data_subject	0.749166	0.680500	0.671281
controller	0.696935	0.752183	0.712293
data_refinement	0.685328	0.804321	0.737309
restriction	0.648130	0.554962	0.593087
processing_activity	0.628397	0.631143	0.626883
purpose	0.538623	0.584579	0.556938
personal_data	0.273965	0.223392	0.230841

Question 4 *Assuming that the method used for categorizing text elements from privacy policies works well for the used cases, to what extent is the method generalizable for other cases?*

To test if the text mining techniques performing similar on a new data set a test set was developed. The method was applied to both training and test set. Based on the Pearson correlation coefficient ($r = 0.892$) we can conclude that the performance of the used combination of text mining techniques on the training set and the test set have a strong relationship. Therefore, we can conclude that the combination of text mining techniques are generalizable on other privacy statements with the same preprocessing steps as the training and test set.

An outstanding difference, visualized in the graphs in Figure 5.1, is seen in the performance measurements of the various sources for the controller component. The text mining techniques are performing significantly lower for the controller component for the Lufthansa privacy statement. As explained in Section 4.7.1, we assume that the if no controller is found or when "you" is identified as being the controller is "we". However, in the case of Lufthansa the controller in the privacy requirements is often "Lufthansa". Both the text chunking grammar and the just described rule are not taking the source name into account.

Question 5 *What readability measurements can be used to measure readability of privacy statements?*

The Automated Readability Index (ARI), Flesch Kincaid Grade Level (FKGL), and Dale-Chall Readability Index (DCRI) are the three indexes used to assess the readability of the privacy statements.

Question 6 *What combination of readability measurements is most effective on measuring the readability of privacy statements for the used cases?*

The readability indexes are compared with the performance measurements to find a relationship between readability and the performance of the text mining techniques. The relationships are assessed with the Pearson correlation coefficient and visualized in Figure 5.3. It seems that ARI predicts the performance of the algorithm best as the ARI score seems to be low when the measurement scores are high. This is against the initial hypothesis that privacy statements which are easy to read are also easy for the text mining techniques to process. The negative relationship between the readability indexes and the performances measurements are against our expectation.

We therefore can conclude that the used readability indexes are not effective on assessing the the ease of identifying components with text mining.

Chapter 7

Discussion

During the search of the best combination of text mining techniques some point came up for discussion. This section explains those points together with the limitation of this project.

The conceptual model can be used as a guideline for writing new privacy statements. It focuses on the completeness of the privacy requirements within in the privacy statements. Note that the conceptual model is not a guideline for the content of the privacy statement. This means that the topics that needs to be covered according to the legislation are not included.

During the preprocessing of the privacy statements in privacy requirements, we assumed that each privacy requirement is at least 36 characters long. The goal was to prevent sentences that were not a privacy requirement were listed as privacy requirement. For example, the headers of sections or subsections were excluded in this process. The drawback of this method is that short privacy requirements could be excluded.

In the next step, where POS-tagging is applied as pre-processing step for dependency parsing and text chunking, we rely on the performance of the NLTK package. Since both text mining techniques (dependency parsing and text chunking) are based on POS tagging, a badly trained POS tagger influences the results greatly. An example of a commonly occurring issue is the annotation of the word “share”. In the case of privacy requirements “share” often indicates a processing activity (i.e. verb). However, the POS tagger labels the word as a noun since, statistically, the word “share” is more often used as a noun according to the training set that is used to train the NLTK POS tagger.. Therefore, “share” is not recognized as a processing activity, but as the personal data component.

Besides the reliance on the NLTK POS tagger, we also assume that the privacy statements do not contain spellings or grammar errors. However, one of the challenges of mining natural language is the variation of writing words and sentences, which can include typos in words. When a word is misspelled the POS tagger might not recognize the word which results in a words without or a wrong annotation and is therefore correctly identified by a parse tree of text chunking grammar.

Another example of a challenge regarding natural language, the different ways of writing, is the following: “[. . .] send to Google” compared to “we collect [. . .]”. Here, Google will not be recognized by the text chunking grammar as the controller of the privacy requirement, where in the second example “we” will be identified as the controller. For the combination of text mining techniques found in this project a problem arises regarding the use of the source name (e.g. Google and Lufthansa) in the privacy requirement. Namely, the text chunking pattern does not take the source name into account when looking for the controller. Therefore, during new efforts of optimizing the combination of text mining techniques a focus can be on the identification of the source name as possible controller. A technique named *entity recognition*[1] could be researched and potentially used to indicate names, such as Google and Lufthansa, as an organizational entity.

The drawback of using dependency parsing, to remove the restriction and data refinement component from the privacy requirement, is the possibility of removing too much words of the original requirement. The dependency parser is told to remove the prepositional modifier from each requirement that is assumed to be part of the restriction or purpose. However, the spaCy dependency parser will possibly assign more phrases as prepositional modifiers. What might happen when running the algorithm is the process of subtracting too much of the 'prep' dependencies (see Section 4.4.2). A solution might be to state additional conditions on subtracting a subtree. For example, a text chunking grammar could be used to assess the subtree on word types. Also, if multiple prepositional modification subtrees are identified, the characteristics of the parent of the subtree can be taken into account (e.g. if the parent is the root or not).

Additional content analyses should be performed by more experts in the field of Requirement Engineering or Privacy to have better say in generalization. For this research three observers were used. However, one of the three observer had a low agreement with both other observers. To evaluate the quality of the observations of the participants more observations should be made. Therewith, the content analysis protocol should be evaluated and possibly improved.

Overall, when changing something to increase performance the algorithm could perform better for the component, but it could also have a bad effect on other components and therefore decrease the overall performance of the algorithm.

Some issues arose during the assessment of the readability of the privacy statements. As explained the three readability indexes (i.e. ARI, FKGL, and DCRI) were based on US grade levels. However, not all scores had an upper bound. This means that the scores could not be normalized and compared with each other. Additionally, there was mentioned that readability indexes are out dated and do not take the meaning of words into account (Hartley, 2016).

7.1 Further Research

Dependencies between different statements are not taken into account. When reading the privacy statements a lot of enumerations are used. Here, a problem arises when splitting sentences based on dots and newlines. The enumerations are recognized as separate sentences, and are therefore not considered to be dependent on other sentences.

Overall, personal data seemed really hard regarding extraction from privacy requirements with the chosen text mining techniques. For further research the focus regarding personal data and data refinement may be only on the nouns (or separate words). The content analysis should be adjusted as only the most specific words need to be selected, like: 'email address', 'location

[1] https://spacy.io/docs/usage/entity-recognition

data', and 'information'. Here, we would make use of the hypernym/hyponym relationships to have better results on the personal data component (Evans et al., 2017).

Some additional topics for further research that can be implemented in the current research could be:

- A time frame (or retention period) could be extracted from the privacy requirement by using the 'NMOD' label of the dependency parser.
- The list of keywords given by Anton et al. (2004) could be used to recognize goals in privacy requirements. In this project, the keywords can be considered as processing activities (3.2). However, focusing on the goals described in privacy requirements could be of great interest for both analysis an existing statement as well as writing a statement. When privacy statement writers focus on the goal of the requirement, they have to take into account the purpose of the processing activity. Since a purpose is mandatory to provide according to the legislations, a framework or guideline focusing on goals will support the thought of how to achieve the final objective.
- Make a distinction between use and transfer processing activities as shown in the initial conceptual model (3.1) and Section 2.3.3.

Additionally, with in mind that the new European Data Protection Regulation is going to be of great influence on privacy statements an interesting new research subject could be relating to topic modeling. As mentioned in the interview the GDPR can be used as guideline for the content for privacy statements. With topic model ling and document similarity a privacy statement can be assessed on having the mandatory topics of the GDPR described.

References

Aho, A. V., Sethi, R., & Ullman, J. D. (1986). *Compilers, principles, techniques.* Addison wesley Boston.

AICPA. (2017). *Security and privacy.* Association of International Certified Professional Accountants. Retrieved 2017/09/22, from `http://www.aicpa.org/InterestAreas/InformationTechnology/Resources/Privacy/Pages/default.aspx`

Anton, A. I., Earp, J. B., He, Q., Stufflebeam, W., Bolchini, D., & Jensen, C. (2004). Financial privacy policies and the need for standardization. *IEEE Security & privacy*, *2*(2), 36–45.

Antòn, A. I., Earp, J. B., & Reese, A. (2002). Analyzing website privacy requirements using a privacy goal taxonomy. In *Requirements engineering, 2002. proceedings. ieee joint international conference on* (pp. 23–31).

Arora, C., Sabetzadeh, M., Briand, L., & Zimmer, F. (2015). Automated checking of conformance to requirements templates using natural language processing. *IEEE transactions on Software Engineering*, *41*(10), 944–968.

Berendt, B., Günther, O., & Spiekermann, S. (2005). Privacy in e-commerce: stated preferences vs. actual behavior. *Communications of the ACM*, *48*(4), 101–106.

Berry, D. M., Kamsties, E., & Krieger, M. M. (2003). From contract drafting to software specification: Linguistic sources of ambiguity.

Bird, S. (2006). Nltk: the natural language toolkit. In *Proceedings of the coling/acl on interactive presentation sessions* (pp. 69–72).

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Breaux, T. D., & Antòn, A. I. (2005). Analyzing goal semantics for rights, permissions, and obligations. In *Requirements engineering, 2005. proceedings. 13th ieee international conference on* (pp. 177–186).

Breaux, T. D., Hibshi, H., & Rao, A. (2014). Eddy, a formal language for specifying and analyzing data flow specifications for conflicting privacy requirements. *Requirements Engineering*, *19*(3), 281–307.

Breaux, T. D., Vail, M. W., & Anton, A. I. (2006). Towards regulatory compliance: Extracting rights and obligations to align requirements with regulations. In *Requirements engineering, 14th ieee international conference* (pp. 49–58).

Carnegie Mellon University. (2009). *Information security office.* Retrieved from `http://www.cmu.edu/iso/governance/roles/data-custodian.html`

Cate, F. H. (2006). The failure of fair information practice principles. *Consumer protection in the age of the information economy*.

Cheong, L. K., & Chang, V. (2007, Sep). The need for data governance: a case study. *ACIS 2007 Proceedings*, 100.

Choi, J. D., Tetreault, J. R., & Stent, A. (2015). It depends: Dependency parser comparison using a web-based evaluation tool. In *Acl (1)* (pp. 387–396).

Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, *37*(1), 51–89.

Chung, L., & do Prado Leite, J. (2009). On non-functional requirements in software engineering. *Conceptual modeling: Foundations and applications*, 363–379.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, *70*(4), 213.

da Silva, A. R., Caramujo, J., Monfared, S., Calado, P., & Breaux, T. (2016). Improving the specification and analysis of privacy policies. *ICEIS 2016*, 336.

Deloitte Privacy Services. (2017, June-July). validation interview.

De Marneffe, M.-C., & Manning, C. D. (2008). *Stanford typed dependencies manual* (Tech. Rep.). Technical report, Stanford University.

Demsetz, H. (1974). Toward a theory of property rights. In *Classic papers in natural resource economics* (pp. 163–177). Springer.

European Parliament and the Council of European Union. (2016). *Council regulation (EU) no 2016/269.* (`http://eur-lex.europa.eu/legal-content/NL/TXT/&uri=CELEX:3A32016R0679`)

Evans, M. C., Bhatia, J., Wadkar, S., & Breaux, T. D. (2017). An evaluation of constituency-based hyponymy extraction from privacy policies. In *25th ieee international requirements engineering conference.*

Facebook. (2016, Sep). *Data policy.* Retrieved from `https://www.facebook.com/policy.php`

Falessi, D., Cantone, G., & Canfora, G. (2013). Empirical principles and an industrial case study in retrieving equivalent requirements via natural language processing techniques. *IEEE Transactions on Software Engineering*, *39*(1), 18–44.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data.* Cambridge university press.

Fellbaum, C. (1998). *Wordnet.* Wiley Online Library.

Gharib, M., Giorgini, P., & Mylopoulos, J. (2016). Ontologies for privacy requirements engineering: A systematic literature review. *arXiv preprint arXiv:1611.10097.*

Giorgini, P., Massacci, F., Mylopoulos, J., & Zannone, N. (2005). Modeling security requirements through ownership, permission and delegation. In *Requirements engineering, 2005. proceedings. 13th ieee international conference on* (pp. 167–176).

Hartley, J. (2016). Is time up for the flesch measure of reading ease? *Scientometrics*, *107*(3), 1523–1526.

He, Q., Antòn, A. I., et al. (2003). A framework for modeling privacy requirements in role engineering. In *Proc. of refsq* (Vol. 3, pp. 137–146).

Hosseini, M. B., Wadkar, S., Breaux, T. D., & Niu, J. (2016). Lexical similarity of information type hypernyms, meronyms and synonyms in privacy policies. In *2016 aaai fall symposium series.*

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, *9*(3), 90–95.

IEC/ISO 25010. (2011). 25010 (2011) systems and software engineering-systems and software quality requirements and evaluation (square)-system and software quality models. *International Organization for Standardization, Geneva, Switzerland.*

IEEE. (1990). Ieee standard glossary of software engineering terminology (ieee std 610.12-1990). los alamitos. *CA: IEEE Computer Society.*

Information technology — Security techniques — Privacy framework (Vol. 2011; Standard). (2011, December). Geneva, CH: International Organization for Standardization.

Karjoth, G., & Schunter, M. (2002). A privacy policy model for enterprises. In *Computer security foundations workshop, 2002. proceedings. 15th ieee* (pp. 271–281).

Laplante, P. A. (2013). *Requirements engineering for software and systems.* CRC Press.

Loper, E., & Bird, S. (2002). Nltk: The natural language toolkit. In *Proceedings of the acl-02 workshop on effective tools and methodologies for teaching natural language processing and computational linguistics-volume 1* (pp. 63–70).

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing* (Vol. 999). MIT Press.

Massey, A. K., Eisenstein, J., Antòn, A. I., & Swire, P. P. (2013). Automated text mining for requirements analysis of policy documents. In *Requirements engineering conference (re), 2013 21st ieee international* (pp. 4–13).

Matteucci, I., Mori, P., & Petrocchi, M. (2013). Prioritized execution of privacy policies. In *Data privacy management and autonomous spontaneous security* (pp. 133–145). Springer.

Mylopoulos, J. (1992). Conceptual modelling and telos. *Conceptual Modelling, Databases, and CASE: an Integrated View of Information System Development, New York: John Wiley & Sons*, 49–68.

Nissenbaum, H. (2011). A contextual approach to privacy online. *Daedalus*, *140*(4), 32–48.

NLTK. (2017). Nltk 3.2.4 documentation: nltk.tokenize package [Computer software manual].

Nuseibeh, B., & Easterbrook, S. (2000). Requirements engineering: a roadmap. In *Proceedings of the conference on the future of software engineering* (pp. 35–46).

och Dag, J. N., Regnell, B., Gervasi, V., & Brinkkemper, S. (2005). A linguistic-engineering approach to large-scale requirements management. *IEEE software*, *22*(1), 32–39.

Otto, P. N., & Antón, A. I. (2007). Addressing legal requirements in requirements engineering. In *Requirements engineering conference, 2007. re'07. 15th ieee international* (pp. 5–14).

Pandey, D., Suman, U., & Ramani, A. (2010). An effective requirement engineering process model for software development and requirements management. In *Advances in recent technologies in communication and computing (artcom), 2010 international conference on* (pp. 287–291).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Pohl, K. (1994). The three dimensions of requirements engineering: a framework and its applications. *Information systems*, *19*(3), 243–258.

Reidenberg, J. R. (1994). Setting standards for fair information practice in the us private sector. *Iowa L. Rev.*, *80*, 497.

Reidenberg, J. R., Bhatia, J., Breaux, T. D., & Norton, T. B. (2016). Ambiguity in privacy policies and the impact of regulation. *The Journal of Legal Studies*, *45*(S2), S163–S190.

Richards, N. M., & King, J. H. (2014). Big data ethics.

Robertson, S., & Robertson, J. (2012). *Mastering the requirements process: Getting requirements right.* Addison-wesley.

Sargent, R. G. (2005). Verification and validation of simulation models. In *Proceedings of the 37th conference on winter simulation* (pp. 130–143).

Schaad, A., & Moffett, J. D. (2002). Delegation of obligations. In *Policies for distributed systems and networks, 2002. proceedings. third international workshop on* (pp. 25–35).

Shell. (2017, April). *Privacy policy.* Retrieved from `http://www.shell.com/privacy.html`

Siena, A., Perini, A., Susi, A., & Mylopoulos, J. (2009). A meta-model for modelling law-compliant requirements. In *Requirements engineering and law (relaw), 2009 second international workshop on* (pp. 45–51).

Singh, M. P. (2013). Norms as a basis for governing sociotechnical systems. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *5*(1), 21.

Sokolowski, J. A., & Banks, C. M. (2010). *Modeling and simulation fundamentals: theoretical underpinnings and practical domains.* John Wiley & Sons.

Sommerville, I. (2011). *Software engineering* (9th ed.).

spaCy. (2017). *Facts & figures: Benchmarks.* Retrieved from `https://spacy.io/docs/api/#benchmarks`

Stemler, S. (2001). An overview of content analysis. *Practical assessment, research & evaluation*, *7*(17), 137–146.

Talburt, J. (1986). The flesch index: An easily programmable readability analysis algorithm. In *Proceedings of the 4th annual international conference on systems documentation* (pp. 114–122).

Tan, A.-H., et al. (1999). Text mining: The state of the art and the challenges. In *Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases* (Vol. 8, pp. 65–70).

Tankard, C. (2016). What the gdpr means for businesses. *Network Security*, *2016*(6), 5–8.

Van Alstyne, M., Brynjolfsson, E., & Madnick, S. (1995). Why not one big database? principles for data ownership. *Decision Support Systems*, *15*(4), 267–284.

Van Lamsweerde, A. (2001). Goal-oriented requirements engineering: A guided tour. In *Requirements engineering, 2001. proceedings. fifth ieee international symposium on* (pp. 249–262).

Weerd, I. v., Souer, J., Versendaal, J., & Brinkkemper, S. (2005). Situational requirements engineering of web content management implementations. In *Proceedings of the international workshop on situational requirements engineering processes (srep'05)* (Vol. 1, pp. 13–30).

Wikipedia. (2017). *Cross Industry Standard Process for Data Mining — Wikipedia, the free encyclopedia.* `http://en.wikipedia.org/w/index.php?title=Cross%20Industry%20Standard%20Process%20for%20Data%20Mining&oldid=798055135`. ([Online; accessed 24-September-2017])

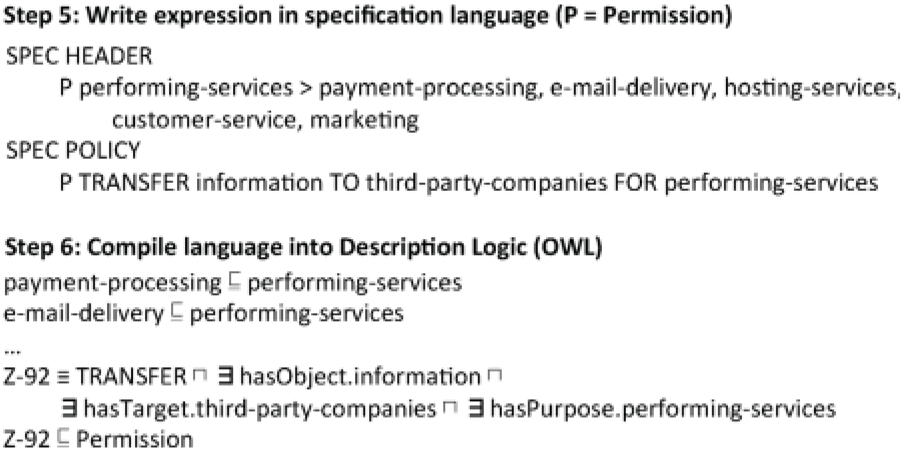
Wirth, R., & Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (pp. 29–39).

Appendix A

Used Literature/Data

Table A.1: Common phrases used in policies regarding the indication of rights and obligations. Extracted from a table from Breaux et al. (2006).

Phrase	**Modality**
does not have a right to	Anti-Right
has a right to	Right
is not required to	Anti-Obligation
may	Right
may deny	Right
may not	Obligation
may not require	Obligation
may require	Right
must	Obligation
must deny	Obligation
must permit	Obligation
must request	Obligation
retains the right to	Right



Step 5: Write expression in specification language (P = Permission)

```
SPEC HEADER
    P performing-services > payment-processing, e-mail-delivery, hosting-services,
        customer-service, marketing
SPEC POLICY
    P TRANSFER information TO third-party-companies FOR performing-services
```

Step 6: Compile language into Description Logic (OWL)

```
payment-processing ⊑ performing-services
e-mail-delivery ⊑ performing-services
...
Z-92 ≡ TRANSFER ⊓ ∃ hasObject.information ⊓
    ∃ hasTarget.third-party-companies ⊓ ∃ hasPurpose.performing-services
Z-92 ⊑ Permission
```

Figure A.1: Examples of encoding the requirement in step 5 and 6 of the requirement extracting process. Taken from Breaux et al. (2014).

Appendix B

Framework Validation

B.1 Interview Protocol

Introduction
I'm going to analyze existing privacy statements with text mining. The goal of the project is to find the best combination text mining techniques that can automatically label elements (building blocks) from requirements in privacy statements. Based on the literature I created a framework that models the elements in relation to each other and the privacy statement. This interview is part of the validation of the purpose of the framework.

General questions
Some questions are asked regarding the work the expert does around privacy statements.

1. Do you mind if I record the interview?
2. What is your role (job description) at Deloitte?

Questions to define the problem space

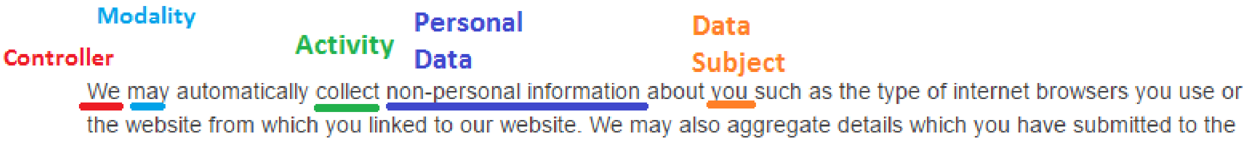
3. When, during your work activities, are you dealing with Privacy Statements?
4. What is your contribution to the creation or evaluation of Privacy Statements?
5. What do you communicate to clients about Privacy Statements?
6. How do you communicate to clients about Privacy Statements?
7. What are your main problems during this communication?
8. How do you solve these problems?

After the first questions the conceptual model is introduced and explained briefly (see Figure 3.1).This model is based on the literature and models the content of privacy statements. Each element is considered to be a fundamental or building block of a privacy statement. Additionally, Example B.1 is shown to explain the element labeling.

Consider that these elements can be automatically subtracted from a Privacy Statement.

9. How can the tool (or model) help reducing the main problem(s) you just mentioned?
10. How can this tool assist your communication regarding Privacy Statements towards clients?
11. Are there others, with similar problems, who could benefit from this tool as well?

Asking for additional comments and thanking for their time.



We may automatically collect non-personal information about you such as the type of internet browsers you use or
the website from which you linked to our website. We may also aggregate details which you have submitted to the

Figure B.1: Example of labeling elements of a privacy requirement of the Shell (2017) privacy statement.

B.2 Interviews

nr	Questions	Respondent 1	Respondent 2
2	What is your role (job description) at Deloitte?	Junior Manager at Privacy Services	Manager at Privacy Services
3	When, during your work activities, are you dealing with Privacy Statements?	We create them. It's mostly part of a project for big project. We never just write one.	When doing gap assessments. Mostly to check if the privacy statements are there and if they are good enough.
4	What is your contribution to the creation or evaluation of Privacy Statements?	Statements are written followed by an extra check with the company. It's an iterative process of writing and getting feedback.	I evaluate them. I did not write them myself from scratch so far. It is mostly reviewing.
5	What do you communicate to clients about Privacy Statements?	It's depending on the kind of organization. With your privacy statement you can be compliant with the law very well, but you can also choose to be a little vague in your description about what you do with data. In the case of keeping it a little vague there is more space to change the approach of data use. As an organization you are more flexible. Also, with regard to the clients own customers, I advise how they should approach their customers. If the statement is external, mostly individuals are addressed you cannot use jargon that is typical for the field, but you should focus on what they understand, take their hand and lead them through your privacy statement.	I indicate the subjects that need to be covered by the statements. Sometimes I recommend 'layered' privacy statements. That are statements that have different levels of details. So for example the highest level (less detailed) statement is a simple, easy-to-read statement that summarizes the whole (complex) statement. But layered statements are not in every country allowed so we have to advised our clients on that. Writing different statements for different countries of one organization takes a lot of time.

6	How do you communicate to clients about Privacy Statements?	It's depending on the relationship you have with your client. It's mostly advising how to approach they customers and the content of the statement rather than writing it myself.	It is depending on what your client already knows about writing statements. Sometimes giving some tips or tricks will be sufficient. However, when the organization is less experienced we will need to explain the content step by step. Sometime we have a framework or template to help them. We also tell clients that a statements should be understandable for everyone. So instead of a legal text, the language should be clear and understandable.
7	What are your main problems during this communication?	While creating a statement it's finding the balance between something that is useful (informative) and what it's allowed by regulations. Also, there are a lot of stakeholders, and therefore interests, involved in the process of creating a statement. That takes time.	Explaining what a statement should contain, like the subjects and structure, and how detailed it must or must not be.
8	How do you solve these problems?	–	Clearly pointing out the subjects, gathering examples, and best practice methods. The structure is often similar in existing statements so that can be copied.
	Showing the conceptual model		
9	How can the an application of the framework help reduce the main problem(s) you just mentioned?	There are a lot of small organizations that are not able to spend that much money on getting assistance with writing a privacy statement. A tool that can create a basic template or that can give some guidance can be useful to them.	It is really useful to know what you can find where in the statement, especially knowing the processing activity in combination with the purpose is really important, especially when using data for profiling. I think that this framework is also really useful as an assistance for writing a policy. Another application could be a tool that is able to split large amounts of text into smaller once.

10	How can this tool assist your communication regarding Privacy Statements towards clients?	First, feedback on a privacy statement can be given based on an analysis of a model that can show what kind of data is used and how it is used. As a result that people who are interested in a particular piece of information are able to see everything activity involving that piece of information. Second, organizations who want to be more transparent can create an easy-to-read summary. However, organizations are not likely to offer a tool that make their statement more readable.	I would see it as a framework, maybe a guideline or instruction document. It could serve as a template for writing a statement or even as a tool that can automatically create them based on input.
11	Are there others, with similar problems, who could benefit from this tool as well?	First, for customers who are only interested in particular information (data object) who can only ask for information about that piece of data. Also, mid-/small-organizations who still need to make a statement that can use the tool to create a statement.	People who do the actual writing could really benefit of a tool that can create statements (or assist with creating them).

Table B.1: Purpose validation interviews with answers of Respondent 1 and Respondent 2.

nr	**Questions**	**Respondent 3**	**Respondent 4**
2	What is your role (job description) at Deloitte?	Senior Manager at Privacy Services	Senior Consultant at Privacy Services
3	When, during your work activities, are you dealing with Privacy Statements?	Not that much. However, a privacy statement is involved in almost every project. We write, check on completeness or sometimes implement statements. Implementation can be one statement for multiple products/services. A privacy statement is mostly used as a communication tool (?). It is not mandatory to have one. However, the GDPR obliges each organization to have a privacy statement.	By exception we write them, but also updating existing privacy statements when it turns out that the statement is not compliant with (new) processes that the organization adapted. Evaluating privacy statements is often part of a PIA or gap assessment. We will evaluate the content. With the GDPR coming up, more mandatory information is required, almost every privacy statement needs to be revised.

4	What is your contribution to the creation or evaluation of Privacy Statements?	I manage the writing of a statement.	We check if all the mandatory topics from the regulation are included in the statement. But also, if what is in there is clear and understandable. Some organizations also do a legal check, but make sure everything is covered according to the law. It is not just tick-the-box.
5	What do you communicate to clients about Privacy Statements?	I tell them that they have to communicate about what they are doing regarding privacy. They could use a privacy statement for that. The WP29 can be used as a guideline of the content of a privacy statement. Each chapter covers a topic that should be in the statement. A lot of organizations see a privacy statement as something they should have and having it means checking a box. The drawback of this approach is that most statements have a lot of legal terms and therefore hard to understand. A privacy statement can also be in the form of a informative text instead of a set requirements or articles.	It depends on the project, but we provide our clients, depending on their expertise, with tools(?) that can help them. When a preforming a PIA (Privacy Impact Assessment), there are important steps. 1) Is there a privacy statement at all. 2) When was the last revision. 3) Is the statement still aligned with the processes found during conduction of the PIA. And based on the first 3 steps, 4) specifying if any updates are needed.
6	How do you communicate to clients about Privacy Statements?	Checklists with topics, but I also try to let them think what is important for their organization. A plain text document is hard to read and often too legally written. I am a fan of an informative text that is focused on its reader. Those approaches are rare, only big organizations like Google and Microsoft have a big focus on how they present their information about processing personal data.	This is also depending on the project and the client. It could be an Excel sheet with requirements. Those requirements are based on the regulation, but summarized. If the client is experience in privacy, it might also happen that I explain the content of a statement during a meeting. You have to adjust your communication based on the experience of client.

7	What are your main problems during this communication?	The approach of "check-the-box, we are having a statement" while it is a hard to understand, legal document. That does not work. What is the alternative? Telling an informative story. Another problem is that the purpose of the processing activity is rarely given. For example if they say that your data will be used for 'marketing purposes'. What does that mean? The problem is that they do not what to be more specific, because then they have to spend a lot of time updating the statements if something changes. Additionally, there is almost never in a statement what they do not do with your data.	The new regulation (GDPR) is quite clear regarding the content of a privacy statement, the problem is the lack of a specified format. Also, organizations are often not giving enough details on data processing, they do not want to be too specific because then they have to rewrite their statement more often in the case they change processes. Besides that, they are sometimes not even able to provide an overview of their activity processes, because they simply do not know all the activity processes around personal data. Actually, the new regulation (GDPR) is quite solid, you need to have a data inventory and a statement, but when you have one, you can have the other. They are both linked.
8	How do you solve these problems?	I had a solution, it would have been solved. But having the knowledge stored in a central database could be really important. It would be really hard to have each processing activity stored with every detail, but gathering that knowledge would be useful. It will improve the flexibility of the organization and statements could be updated much faster. On the other hand, having to much detail in a statement would reduce the number people reading it.	An organization needs a strategy first, without looking at privacy. Sometimes I advise my clients to be more critically about what they need (data minimization). Sometimes, to be more detailed, but also to define the purpose of a processing activity better. Who's your client? What do want to offer? Then you can start asking: "What do you need to offer your services?" When they want to store more data then they will use, it is about the risk they want to take. Do you want the risk of getting a fine or maybe higher the chance of having a data-leak.
	Showing the conceptual model		

9	How can the an application of the framework help reduce the main problem(s) you just mentioned?	I will only read a statement if I am curious about some specific piece of information they have from me. So being able to search on one specific piece of information would be nice. For example a chat bot that can answer your question about what an organization is doing with your data.	The problem I described was centered towards the purpose and personal data. This conceptual model can provide me with an overview of all the data that is processed according to the privacy statement.
10	How can this tool assist your communication regarding Privacy Statements towards clients?	It could be a tool that could help our clients comparing their own statements with each other.	The application can decompose a privacy statement at once and create a list with all available personal data (and its processes) and check how it is aligned with the data inventory. That list could also be used to test the purpose of the activity process regarding the personal data. The benefit of an application like that is that you do not have to be a privacy specialist nor a lawyer.
11	Are there others, with similar problems, who could benefit from this tool as well?	-	It could be used by our clients who have multiple statements that might need a update as a result of new or changed process activities. They can use an application to subtract the involved data and therewith update the statements according to the output. A second application could be a tool that creates a privacy statement based on input, so using the model the other way around.

Table B.2: Purpose validation interviews with answers of Respondent 3 and Respondent 4.

Appendix C

Method

C.1 Python Anaconda Environment

```
# This file may be used to create an environment using:
# $ conda create --name <env> --file <this file>
# platform: win-64
@EXPLICIT
https://conda.anaconda.org/conda-forge/win-64/asn1crypto-0.22.0-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/backports
 .shutil_get_terminal_size-1.0.0-py27_1.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/backports_abc-0.5-py27_0.tar.bz2
https://repo.continuum.io/pkgs/free/win-64/boto-2.48.0-py27_0.tar.bz2
https://repo.continuum.io/pkgs/free/win-64/bz2file-0.98-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/ca-certificates-2017.7.27.1-0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/certifi-2017.7.27.1-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/cffi-1.10.0-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/chardet-3.0.2-py27_1.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/colorama-0.3.9-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/configparser-3.5.0-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/cryptography-1.9-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/cycler-0.10.0-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/cymem-1.31.2-py27_vc9_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/cytoolz-0.8.2-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/decorator-4.1.2-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/dill-0.2.6-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/empath-0.89-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/entrypoints-0.2.3-py27_1.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/enum34-1.1.6-py27_1.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/freetype-2.7-vc9_1.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/ftfy-4.4.2-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/functools32-3.2.3.2-py27_1.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/gensim-2.3.0-py27_vc9_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/html5lib-0.999999999-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/icu-58.1-vc9_1.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/idna-2.5-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/ipaddress-1.0.18-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/ipykernel-4.6.1-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/ipython-5.4.1-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/ipython_genutils-0.2.0-py27_0.tar.bz2
```

```
https://conda.anaconda.org/conda-forge/win-64/ipywidgets-6.0.0-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/jieba-0.39-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/jinja2-2.9.5-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/jpeg-9b-vc9_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/jsonschema-2.5.1-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/jupyter-1.0.0-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/jupyter_client-5.1.0-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/jupyter_console-5.1.0-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/jupyter_core-4.3.0-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/libpng-1.6.28-vc9_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/markupsafe-1.0-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/matplotlib-2.0.2-py27_1.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/mistune-0.7.4-py27_0.tar.bz2
https://repo.continuum.io/pkgs/free/win-64/mkl-2017.0.3-0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/murmurhash-0.26.4-py27_vc9_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/nbconvert-4.2.0-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/nbformat-4.3.0-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/nltk-3.2.4-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/notebook-5.0.0-py27_0.tar.bz2
https://repo.continuum.io/pkgs/free/win-64/numpy-1.12.1-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/openssl-1.0.2l-vc9_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/pandas-0.20.3-py27_1.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/pathlib-1.0.1-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/pathlib2-2.3.0-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/patsy-0.4.1-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/pickleshare-0.7.3-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/pip-9.0.1-py27_0.tar.bz2
https://repo.continuum.io/pkgs/free/win-64/plac-0.9.6-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/preshed-1.0.0-py27_vc9_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/prompt_toolkit-1.0.15-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/pycparser-2.18-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/pygments-2.2.0-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/pyopenssl-16.2.0-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/pyparsing-2.2.0-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/pyqt-5.6.0-py27_4.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/pyreadline-2.1-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/pysocks-1.6.7-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/python-2.7.13-1.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/python-dateutil-2.6.1-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/pytz-2017.2-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/pyzmq-16.0.2-py27_2.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/qt-5.6.2-vc9_1.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/qtconsole-4.3.0-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/regex-2017.07.28-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/requests-2.18.3-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/scandir-1.5-py27_1.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/scattertext-0.0.2.9.9-py27_0.tar.bz2
https://repo.continuum.io/pkgs/free/win-64/scikit-learn-0.18.2-np112py27_0.tar.bz2
https://repo.continuum.io/pkgs/free/win-64/scipy-0.19.1-np112py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/seaborn-0.8.0-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/setuptools-36.2.2-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/simplegeneric-0.8.1-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/singledispatch-3.4.0.3-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/sip-4.18-py27_1.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/six-1.10.0-py27_1.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/smart_open-1.5.3-py27_0.tar.bz2
```

```
https://conda.anaconda.org/conda-forge/win-64/spacy-1.9.0-np112py27_vc9_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/ssl_match_hostname-3.5.0.1-py27_1.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/statsmodels-0.8.0-np112py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/termcolor-1.1.0-py27_1.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/thinc-6.5.2-np112py27_vc9_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/toolz-0.8.2-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/tornado-4.5.1-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/tqdm-4.15.0-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/traitlets-4.3.2-py27_0.tar.bz2
https://repo.continuum.io/pkgs/free/win-64/ujson-1.35-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/urllib3-1.21.1-py27_1.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/vc-9-0.tar.bz2
https://repo.continuum.io/pkgs/free/win-64/vs2008_runtime-9.00.30729.5054-0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/wcwidth-0.1.7-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/webencodings-0.5-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/wheel-0.29.0-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/widgetsnbextension-2.0.0-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/win_inet_pton-1.0.1-py27_1.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/win_unicode_console-0.5-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/wincertstore-0.2-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/wrapt-1.10.8-py27_0.tar.bz2
https://conda.anaconda.org/conda-forge/win-64/zlib-1.2.11-vc9_0.tar.bz2
```

C.2 Privacy Statements

C.2.1 Original Statements

[not included in this printed version]

Depending on which Services you use, we collect different kinds of information from or about you. Things you do and information you provide. We collect the content and other information you provide when you use our Services, including when you sign up for an account, create or share, and message or communicate with others. This can include information in or about the content you provide, such as the location of a photo or the date a file was created. We also collect information about how you use our Services, such as the types of content you view or engage with or the frequency and duration of your activities. Things others do and information they provide. We also collect content and information that other people provide when they use our Services, including information about you, such as when they share a photo of you, send a message to you, or upload, sync or import your contact information. Your networks and connections. We collect information about the people and groups you are connected to and how you interact with them, such as the people you communicate with the most or the groups you like to share with. We also collect contact information you provide if you upload, sync or import this information (such as an address book) from a device. Information about payments. If you use our Services for purchases or financial transactions (like when you buy something on Facebook, make a purchase in a game, or make a donation), we collect information about the purchase or transaction. This includes your payment information, such as your credit or debit card number and other card information, and other account and authentication information, as well as billing, shipping and contact details. Device information. We collect information from or about the computers, phones, or other devices where you install or access our Services, depending on the permissions you've granted. We may associate the information we collect from your different devices, which helps us provide consistent Services across your devices. Here are some examples of the device information we collect:

Attributes such as the operating system, hardware version, device settings, file and software names and types, battery and signal strength, and device identifiers. Device locations, including specific geographic locations, such as through GPS, Bluetooth, or WiFi signals. Connection information such as the name of your mobile operator or ISP, browser type, language and time zone, mobile phone number and IP address. Information from websites and apps that use our Services. We collect information when you visit or use third-party websites and apps that use our Services (like when they offer our Like button or Facebook Log In or use our measurement and advertising services). This includes information about the websites and apps you visit, your use of our Services on those websites and apps, as well as information the developer or publisher of the app or website provides to you or us. Information from third-party partners. We receive information about you and your activities on and off Facebook from third-party partners, such as information from a partner when we jointly offer services or from an advertiser about your experiences or interactions with them. Facebook companies. We receive information about you from companies that are owned or operated by Facebook, in accordance with their terms and policies. Learn more about these companies and their privacy policies.

How do we use this information?

We are passionate about creating engaging and customized experiences for people. We use all of the information we have to help us provide and support our Services. Here's how: Provide, improve and develop Services. We are able to deliver our Services, personalize content, and make suggestions for you by using this information to understand how you use and interact with our Services and the people or things you're connected to and interested in on and off our Services.

We also use information we have to provide shortcuts and suggestions to you. For example, we are able to suggest that your friend tag you in a picture by comparing your friend's pictures to information we've put together from your profile pictures and the other photos in which you've been tagged. If this feature is enabled for you, you can control whether we suggest that another user tag you in a photo using the "Timeline and Tagging" settings.

When we have location information, we use it to tailor our Services for you and others, like helping you to check-in and find local events or offers in your area or tell your friends that you are nearby.

We conduct surveys and research, test features in development, and analyze the information we have to evaluate and improve products and services, develop new products or features, and conduct audits and troubleshooting activities. Communicate with you. We use your information to send you marketing communications, communicate with you about our Services and let you know about our policies and terms. We also use your information to respond to you when you contact us. Show and measure ads and services. We use the information we have to improve our advertising and measurement systems so we can show you relevant ads on and off our Services and measure the effectiveness and reach of ads and services. Learn more about advertising on our Services and how you can control how information about you is used to personalize the ads you see. Promote safety and security. We use the information we have to help verify accounts and activity, and to promote safety and security on and off of our Services, such as by investigating suspicious activity or violations of our terms or policies. We work hard to protect your account using teams of engineers, automated systems, and advanced technology such as encryption and machine learning. We also offer easy-to-use security tools that add an extra layer of security to your account. For more information about promoting safety on Facebook, visit the Facebook Security Help Center. We use cookies and similar technologies to provide and support our Services and each of the uses outlined and described in this section of our policy. Read our Cookie Policy to learn more.

How is this information shared?

Sharing On Our Services People use our Services to connect and share with others. We make this possible by sharing your information in the following ways: People you share and communicate with. When you share and communicate using our Services, you choose the audience who can see what you share. For example, when you post on Facebook, you select the audience for the post, such as a customized group of individuals, all of your Friends, or members of a Group. Likewise, when you use Messenger, you also choose the people you send photos to or message.

Public information is any information you share with a public audience, as well as information in your Public Profile, or content you share on a Facebook Page or another public forum. Public information is available to anyone on or off our Services and can be seen or accessed through online search engines, APIs, and offline media, such as on TV.

In some cases, people you share and communicate with may download or re-share this content with others on and off our Services. When you comment on another person's post or like their content on Facebook, that person decides the audience who can see your comment or like. If their audience is public, your comment will also be public. People that see content others share about you. Other people may use our Services to share content about you with the audience they choose. For example, people may share a photo of you, mention or tag you at a location in a post, or share information about you that you shared with them. If you have concerns with someone's post, social reporting is a way for people to quickly and easily ask for help from someone they trust. Learn More. Apps, websites and third-party integrations on or using our Services. When you use third-party apps, websites or other services that use, or are integrated with, our Services, they may receive information about what you post or share. For example, when you play a game with your Facebook friends or use the Facebook Comment or Share button on a website, the game developer or website may get information about your activities in the game or receive a comment or link that you share from their website on Facebook. In addition, when you download or use such third-party services, they can access your Public Profile, which includes your username or user ID, your age range and country/language, your list of friends, as well as any information that you share with them. Information collected by these apps, websites or integrated services is subject to their own terms and policies.

Learn more about how you can control the information about you that you or others share with these apps and websites. Sharing within Facebook companies. We share information we have about you within the family of companies that are part of Facebook. Learn more about our companies. New owner. If the ownership or control of all or part of our Services or their assets changes, we may transfer your information to the new owner.

Sharing With Third-Party Partners and Customers We work with third party companies who help us provide and improve our Services or who use advertising or related products, which makes it possible to operate our companies and provide free services to people around the world.

Here are the types of third parties we can share information with about you: Advertising, Measurement and Analytics Services (Non-Personally Identifiable Information Only). We want our advertising to be as relevant and interesting as the other information you find on our Services. With this in mind, we use all of the information we have about you to show you relevant ads. We do not share information that personally identifies you (personally identifiable information is information like name or email address that can by itself be used to contact you or identifies who you are) with advertising, measurement or analytics partners unless you give us permission. We may provide these partners with information about the reach and effectiveness of their advertising without providing information that personally identifies you, or if we have aggregated the information so that it does not personally identify you. For example, we may tell an advertiser how its ads performed, or how many people viewed their ads or installed an app after seeing an ad, or provide non-personally identifying demographic information (such as 25 year old female, in Madrid, who likes software engineering) to these partners to help them understand their audience or customers, but only after the advertiser has agreed to abide by our advertiser guidelines.

Please review your advertising preferences to understand why you're seeing a particular ad on Facebook. You can adjust your ad preferences if you want to control and manage your ad experience on Facebook. Vendors, service providers and other partners. We transfer information to vendors, service providers, and other partners who globally support our business, such as providing technical infrastructure services, analyzing how our Services are used, measuring the effectiveness of ads and services, providing customer service, facilitating payments, or conducting academic research and surveys. These partners must adhere to strict confidentiality obligations in a way that is consistent with this Data Policy and the agreements we enter into with them.

How can I manage or delete information about me?

You can manage the content and information you share when you use Facebook through the Activity Log tool. You can also download information associated with your Facebook account through our Download Your Information tool.

We store data for as long as it is necessary to provide products and services to you and others, including those described above. Information associated with your account will be kept until your account is deleted, unless we no longer need the data to provide products and services.

You can delete your account any time. When you delete your account, we delete things you have posted, such as your photos and status updates. If you do not want to delete your account, but want to temporarily stop using Facebook, you may deactivate your account instead. To learn more about deactivating or deleting your account, click here. Keep in mind that information that others have shared about you is not part of your account and will not be deleted when you delete your account.

How do we respond to legal requests or prevent harm?

We may access, preserve and share your information in response to a legal request (like a search warrant, court order or subpoena) if we have a good faith belief that the law requires us to do so. This may include responding to legal requests from jurisdictions outside of the United States where we have a good faith belief that the response is required by law in that jurisdiction, affects users in that jurisdiction, and is consistent with internationally recognized standards. We may also access, preserve and share information when we have a good faith belief it is necessary to: detect, prevent and address fraud and other illegal activity; to protect ourselves, you and others, including as part of investigations; or to prevent death or imminent bodily harm. For example, we may provide information to third-party partners about the reliability of your account to prevent fraud and abuse on and off of our Services. Information we receive about you, including financial transaction data related to purchases made with Facebook, may be accessed, processed and retained for an extended period of time when it is the subject of a legal request or obligation, governmental investigation, or investigations concerning possible violations of our terms or policies, or otherwise to prevent harm. We also may retain information from accounts disabled for violations of our terms for at least a year to prevent repeat abuse or other violations of our terms.

How our global services operate

Facebook may share information internally within our family of companies or with third parties for purposes described in this policy. Information collected within the European Economic Area ("EEA") may, for example, be transferred to countries outside of the EEA for the purposes as described in this policy. We utilize standard contract clauses approved by the European Commission, adopt other means under European Union law, and obtain your consent to legitimize data transfers from the EEA to the United States and other countries.

You can contact us using the information provided below with questions or concerns. We also may resolve disputes you have with us in connection with our privacy policies and practices through TRUSTe. You can contact TRUSTe through their website.

How will we notify you of changes to this policy?

We'll notify you before we make changes to this policy and give you the opportunity to review and comment on the revised policy before continuing to use our Services.

Google

Information we collect

We collect information to provide better services to all of our users – from figuring out basic stuff like which language you speak, to more complex things like which ads you'll find most useful, the people who matter most to you online, or which YouTube videos you might like.

We collect information in the following ways:

Information you give us. For example, many of our services require you to sign up for a Google Account. When you do, we'll ask for personal information, like your name, email address, telephone number or credit card to store with your account. If you want to take full advantage of the sharing features we offer, we might also ask you to create a publicly visible Google Profile, which may include your name and photo.

Information we get from your use of our services. We collect information about the services that you use and how you use them, like when you watch a video on YouTube, visit a website that uses our advertising services, or view and interact with our ads and content. This information includes:

Device information

We collect device-specific information (such as your hardware model, operating system version, unique device identifiers, and mobile network information including phone number). Google may associate your device identifiers or phone number with your Google Account.

Log information

When you use our services or view content provided by Google, we automatically collect and store certain information in server logs. This includes:

details of how you used our service, such as your search queries. telephony log information like your phone number, calling-party number, forwarding numbers, time and date of calls, duration of calls, SMS routing information and types of calls. Internet protocol address. device event information such as crashes, system activity, hardware settings, browser type, browser language, the date and time of your request and referral URL. cookies that may uniquely identify your browser or your Google Account. Location information

When you use Google services, we may collect and process information about your actual location. We use various technologies to determine location, including IP address, GPS, and other sensors that may, for example, provide Google with information on nearby devices, Wi-Fi access points and cell towers.

Unique application numbers

Certain services include a unique application number. This number and information about your installation (for example, the operating system type and application version number) may be sent to Google when you install or uninstall that service or when that service periodically contacts our servers, such as for automatic updates.

Local storage

We may collect and store information (including personal information) locally on your device using mechanisms such as browser web storage (including HTML 5) and application data caches.

Cookies and similar technologies

We and our partners use various technologies to collect and store information when you visit a Google service, and this may include using cookies or similar technologies to identify your browser or device. We also use these technologies to collect and store information when you interact with services we offer to our partners, such as advertising services or Google features that may appear on other sites. Our Google Analytics product helps businesses and site owners analyze the traffic to their websites and apps. When used in conjunction with our advertising services, such as those using the DoubleClick cookie, Google Analytics information is linked, by the Google Analytics customer or by Google, using Google technology, with information about visits to multiple sites.

Information we collect when you are signed in to Google, in addition to information we obtain about you from partners, may be associated with your Google Account. When information is associated with your Google Account, we treat it as personal information. For more information about how you can access, manage or delete information that is associated with your Google Account, visit the Transparency and choice section of this policy.

Back to top How we use information we collect

We use the information we collect from all of our services to provide, maintain, protect and improve them, to develop new ones, and to protect Google and our users. We also use this information to offer you tailored content – like giving you more relevant search results and ads.

We may use the name you provide for your Google Profile across all of the services we offer that require a Google Account. In addition, we may replace past names associated with your Google Account so that you are represented consistently across all our services. If other users already have your email, or other information that identifies you, we may show them your publicly visible Google Profile information, such as your name and photo.

If you have a Google Account, we may display your Profile name, Profile photo, and actions you take on Google or on third-party applications connected to your Google Account (such as +1's, reviews you write and comments you post) in our services, including displaying in ads and other commercial contexts. We will respect the choices you make to limit sharing or visibility settings in your Google Account.

When you contact Google, we keep a record of your communication to help solve any issues you might be facing. We may use your email address to inform you about our services, such as letting you know about upcoming changes or improvements.

We use information collected from cookies and other technologies, like pixel tags, to improve your user experience and the overall quality of our services. One of the products we use to do this on our own services is Google Analytics. For example, by saving your language preferences, we'll be able to have our services appear in the language you prefer. When showing you tailored ads, we will not associate an identifier from cookies or similar technologies with sensitive categories, such as those based on race, religion, sexual orientation or health.

Our automated systems analyze your content (including emails) to provide you personally relevant product features, such as customized search results, tailored advertising, and spam and malware detection.

We may combine personal information from one service with information, including personal information, from other Google services – for example to make it easier to share things with people you know. Depending on your account settings, your activity on other sites and apps may be associated with your personal information in order to improve Google's services and the ads delivered by Google.

We will ask for your consent before using information for a purpose other than those that are set out in this Privacy Policy.

Google processes personal information on our servers in many countries around the world. We may process your personal information on a server located outside the country where you live.

Back to top Transparency and choice

People have different privacy concerns. Our goal is to be clear about what information we collect, so that you can make meaningful choices about how it is used. For example, you can:

Review and update your Google activity controls to decide what types of data, such as videos you've watched on YouTube or past searches, you would like saved with your account when you use Google services. You can also visit these controls to manage whether certain activity is stored in a cookie or similar technology on your device when you use our services while signed-out of your account. Review and control certain types of information tied to your Google Account by using Google Dashboard. View and edit your preferences about the Google ads shown to you on Google and across the web, such as which categories might interest you, using Ads Settings. You can also visit that page to opt out of certain Google advertising services. Adjust how the Profile associated with your Google Account appears to others. Control who you share information with through your Google Account. Take information associated with your Google Account out of many of our services. Choose whether your Profile name and Profile photo appear in shared endorsements that appear in ads. You may also set your browser to block all cookies, including cookies associated with our services, or to indicate when a cookie is being set by us. However, it's important to remember that many of our services may not function properly if your cookies are disabled. For example, we may not remember your language preferences.

Back to top Information you share

Many of our services let you share information with others. Remember that when you share information publicly, it may be indexable by search engines, including Google. Our services provide you with different options on sharing and removing your content.

Back to top Accessing and updating your personal information

Whenever you use our services, we aim to provide you with access to your personal information. If that information is wrong, we strive to give you ways to update it quickly or to delete it – unless we have to keep that information for legitimate business or legal purposes. When updating your personal information, we may ask you to verify your identity before we can act on your request.

We may reject requests that are unreasonably repetitive, require disproportionate technical effort (for example, developing a new system or fundamentally changing an existing practice), risk the privacy of others, or would be extremely impractical (for instance, requests concerning information residing on backup systems).

Where we can provide information access and correction, we will do so for free, except where it would require a disproportionate effort. We aim to maintain our services in a manner that protects information from accidental or malicious destruction. Because of this, after you delete information from our services, we may not immediately delete residual copies from our active servers and may not remove information from our backup systems.

Back to top Information we share

We do not share personal information with companies, organizations and individuals outside of Google unless one of the following circumstances applies:

With your consent

We will share personal information with companies, organizations or individuals outside of Google when we have your consent to do so. We require opt-in consent for the sharing of any sensitive personal information.

With domain administrators

If your Google Account is managed for you by a domain administrator (for example, for G Suite users) then your domain administrator and resellers who provide user support to your organization will have access to your Google Account information (including your email and other data). Your domain administrator may be able to:

view statistics regarding your account, like statistics regarding applications you install. change your account password. suspend or terminate your account access. access or retain information stored as part of your account. receive your account information in order to satisfy applicable law, regulation, legal process or enforceable governmental request. restrict your ability to delete or edit information or privacy settings. Please refer to your domain administrator's privacy policy for more information.

For external processing

We provide personal information to our affiliates or other trusted businesses or persons to process it for us, based on our instructions and in compliance with our Privacy Policy and any other appropriate confidentiality and security measures.

For legal reasons

We will share personal information with companies, organizations or individuals outside of Google if we have a good-faith belief that access, use, preservation or disclosure of the information is reasonably necessary to:

meet any applicable law, regulation, legal process or enforceable governmental request. enforce applicable Terms of Service, including investigation of potential violations. detect, prevent, or otherwise address fraud, security or technical issues. protect against harm to the rights, property or safety of Google, our users or the public as required or permitted by law. We may share non-personally identifiable information publicly and with our partners – like publishers, advertisers or connected sites. For example, we may share information publicly to show trends about the general use of our services.

If Google is involved in a merger, acquisition or asset sale, we will continue to ensure the confidentiality of any personal information and give affected users notice before personal information is transferred or becomes subject to a different privacy policy.

Back to top Information security

We work hard to protect Google and our users from unauthorized access to or unauthorized alteration, disclosure or destruction of information we hold. In particular:

We encrypt many of our services using SSL. We offer you two step verification when you access your Google Account, and a Safe Browsing feature in Google Chrome. We review our information collection, storage and processing practices, including physical security measures, to guard against unauthorized access to systems. We restrict access to personal information to Google employees, contractors and agents who need to know that information in order to process it for us, and who are subject to strict contractual confidentiality obligations and may be disciplined or terminated if they fail to meet these obligations. Back to top When this Privacy Policy applies

Our Privacy Policy applies to all of the services offered by Google Inc. and its affiliates, including YouTube, services Google provides on Android devices, and services offered on other sites (such as our advertising services), but excludes services that have separate privacy policies that do not incorporate this Privacy Policy.

Our Privacy Policy does not apply to services offered by other companies or individuals, including products or sites that may be displayed to you in search results, sites that may include Google services, or other sites linked from our services. Our Privacy Policy does not cover the information practices of other companies and organizations who advertise our services, and who may use cookies, pixel tags and other technologies to serve and offer relevant ads.

Back to top Compliance and cooperation with regulatory authorities

We regularly review our compliance with our Privacy Policy. We also adhere to several self regulatory frameworks, including the EU-US and Swiss-US Privacy Shield Frameworks. When we receive formal written complaints, we will contact the person who made the complaint to follow up. We work with the appropriate regulatory authorities, including local data protection authorities, to resolve any complaints regarding the transfer of personal data that we cannot resolve with our users directly.

Back to top Changes

Our Privacy Policy may change from time to time. We will not reduce your rights under this Privacy Policy without your explicit consent. We will post any privacy policy changes on this page and, if the changes are significant, we will provide a more prominent notice (including, for certain services, email notification of privacy policy changes). We will also keep prior versions of this Privacy Policy in an archive for your review.

Back to top Specific product practices

The following notices explain specific privacy practices with respect to certain Google products and services that you may use:

Chrome and Chrome OS Play Books Payments Fiber Project Fi G Suite for Education For more information about some of our most popular services, you can visit the Google Product Privacy Guide.

Back to top Other useful privacy and security related materials

Further useful privacy and security related materials can be found through Google's policies and principles pages, including:

Information about our technologies and principles, which includes, among other things, more information on how Google uses cookies. technologies we use for advertising. how we recognize patterns like faces. A page that explains what data is shared with Google when you visit websites that use our advertising, analytics and social products. The Privacy Checkup tool, which makes it easy to review your key privacy settings. Google's safety center, which provides information on how to stay safe and secure online.

KLM

We may collect and process the following categories of personal information:

For example, we may record your name, title, gender and date of birth, your nationality country of residence and passport number

Your contact details may include your address, telephone number and email address. When you create a personal account or register for a service, we may also record your sign in details and other information you fill out on your personal account or registration form. For business travellers, we also collect information relating to your company such as company name and business location.

When you make a reservation or book a flight with us, we process your reservation and booking information. This information may include details about your flight, prices and the date of your reservation or booking. In addition hereto we process information in relation to ancillary services (such as upgrades and extra luggage) and products you purchase (for example via the KLM webshop).

When you travel with us, we process information in relation to your journey. Such as your travel itinerary, (online) check-in, your (mobile) boarding pass and your travel companions. We may also record any specified medical needs or dietary requests you have and any additional assistance you require.

When you become a member of our loyalty programs, we process your membership number, miles or credits balance, awards and benefits, type and level of membership and other information in relation to your membership.

When you send us an email or chat with us online or via social media, we register your communication with us. When you call us, our customer support will register your questions or complaints in our database. We may also record telephone calls for training purposes or to prevent or combat fraud.

When you visit our websites or use one of our apps, we may register your IP address, browser type, operating system, referring website, web-browsing behavior and app use. We may receive an automatic notification when you open a newsletter or click on a link in such newsletter. With your permission we may also receive your location data. You can also agree to provide us with access to certain data stored on your mobile phone (such as photos and contacts). For more information, please see below (cookies).

Depending on your social network settings we may receive information from your social network provider. For example when you sign in for our services using a social network account, we may receive your social network profile including your contact details, interests and contacts. For more information on the personal data that we receive from your social network provider and how to change your settings, please check the website and privacy policy of your social network provider.

You may choose to share information with us, for example when you leave a comment for us on Facebook, fill out a customer survey or submit an entry for a contest.

Sensitive data. Certain categories of personal information we collect and use, such as data about race, religion or health, can be considered "sensitive personal data" under Dutch data protection law.

General. We may be required to collect, use and share such data with third parties for the purposes as described in this privacy policy. For example, we keep records of certain passengers who have been found to be carrying illegal drugs or who have interfered with the safety.

Data you share. You may also choose to share such information with us, for example because you have requested for specific medical assistance, you have sought clearance to fly with a medical condition or you request a particular type of special meal which may imply that you hold particular religious beliefs. By providing any personal information that is, or could be considered to be, sensitive personal information, you agree that we may collect, use and share this information with third parties as described in this privacy policy. If you withdraw your consent, it may mean we will not be able to provide all or parts of the services you have requested from us. Please be aware that in these circumstances you will not be able to cancel or obtain a refund of any fees you have paid.

Cookies and similar technologies. When you use our website or mobile apps we collect information via cookies and similar technologies. For more information please read our cookie policy on the website or mobile app you use.

Specific services, apps or events. For specific services, apps or events, we may collect other types of data and use such data for different purposes than described in this privacy policy. We will inform you about this when you register for a service or an event or download an app.

KLM may obtain your personal data in a number of ways, for example when you book a flight, you register for one of our loyalty programs, use one of our apps, communicate with us via social media or when you subscribe to our newsletter.

We may also receive personal information from our group companies, partners or from service providers.

For example, when you use a third party platform to search and book a flight, we may receive your contact details from that provider. When you purchase a service from our loyalty partners, they share your earned miles with us. Depending on your social network settings, we may also receive information from your social network provider (please see above).

Some information we receive from public authorities.

To handle your reservations and bookings and to fulfill your travel arrangements and purchases, we need to process most of the information described above. For example, we need your name, passport number and other identifying information to issue your ticket. To inform you about changes in departure times, we need your contact details. And, to ensure that you receive the required care, we require your specified medical needs.

To let you or your company benefit from the discounts and rewards under our loyalty programs, we use your membership information and your booking information.

For example, we use your name and flight details when you check-in for your flight with our app. Some of our online services and apps use your location, for example to show you the nearest location of your interest. To ease your use of our online services or apps, we may analyze the data we collect when you use our digital media and combine it with information collected via cookies and similar technologies (please see above). For example, to understand which digital channel (email, social media) or device (desktop, table or mobile) you prefer, so we can restrict our communication to that channel or device.

We use your contact details to communicate with you, answer your questions or handle your complaints.

General. We use automatic tools to perform statistical research into general trends regarding the use of our services, loyalty programs, websites, apps and social media and the behavior and preferences of our customers and users.

Data. To perform our research, we may merge and analyze the different types of data as described above. We will only use aggregated data and do not use name, email address or other directly identifying information. We may also combine such aggregated data with information we receive from our group companies or from public sources (such as Statistics Netherlands). Without your consent we will not use sensitive data for this statistical research.

Benefits. Statistical research helps us to develop better services and offerings, to improve our loyalty programs, to provide more responsive customer support and to improve the design and content of our websites and mobile apps.

General. We may use your personal information to send you newsletters, magazines, promotions or other marketing communications.

Categories of data. To understand what is relevant to you, we may use automatic tools to analyze your personal information. For this purpose we may use and combine the information described above. We may also combine such aggregated data with information we receive from our group companies or from public sources (such as Statistics Netherlands). Without your consent we will not use sensitive data for this statistical research.

Benefits. We use the results of our analysis to tailor our marketing communications to your specific interests and preferences. For example, if our analysis shows that you may be interested in certain types of travel or a specific destination, we may customize our newsletter and websites with offers and content that we believe will be relevant to you. Channels. We may use different channels for our marketing communications, such as e-mail, apps, social media and your personal online account.

Custom audience targeting. We may participate in Facebook's Custom Audience program, which enables us to display personalized ads to you when you visit Facebook. We may share your email address or another identifier to Facebook to enable Facebook to determine if you have a Facebook account. We may use similar programs from other social networks. You may opt-out of participation in our Facebook Custom Audience (or similar programs) by sending an email, from the email address you are opting out of, to KLMPrivacyOffice@klm.com. For more information on custom audience targeting, please see the website of your social media provider.

Object or revoke. You may object or revoke your consent for receiving marketing communications at any time by following the instructions in the relevant marketing communication or by contacting us at KLMPrivacyOffice@klm.com.

Illegal drugs. The names of passengers who have disembarked at Amsterdam Airport Schiphol and who have been found by the Royal Netherlands Marechaussee to be carrying illegal drugs will be recorded by the State of the Netherlands. KLM receives the names of these persons from the State of the Netherlands. KLM may refuse to enter into any transport contract with these persons for a period of three years for direct flights from Schiphol to Suriname, Aruba, Bonaire, St Martin or Curacao and for direct flights from these countries to Schiphol. You may request permission to examine this data or to make changes to its accuracy by submitting this request in writing to the Royal Netherlands Marechaussee at P.O. Box 90615, 2509 LP The Hague, The Netherlands. If you reside on Aruba, the Netherlands Antilles, in Suriname or in Venezuela, you should accompany this written request with a copy of your passport.

Instagram

1. INFORMATION WE COLLECT

We collect the following types of information.

Information you provide us directly: Your username, password and e-mail address when you register for an Instagram account. Profile information that you provide for your user profile (e.g., first and last name, picture, phone number). This information allows us to help you or others be "found" on Instagram. User Content (e.g., photos, comments, and other materials) that you post to the Service. Communications between you and Instagram. For example, we may send you Service-related emails (e.g., account verification, changes/updates to features of the Service, technical and security notices). Note that you may not opt out of Service-related e-mails. Finding your friends on Instagram: If you choose, you can use our "Find friends" feature to locate other people with Instagram accounts either through (i) your contacts list, (ii) third-party social media sites or (iii) through a search of names and usernames on Instagram. If you choose to find your friends through (i) your device's contacts list, then Instagram will access your contacts list to determine whether or not someone associated with your contact is using Instagram. If you choose to find your friends through a (ii) third-party social media site, then you will be prompted to set up a link to the third-party service and you understand that any information that such service may provide to us will be governed by this Privacy Policy. If you choose to find your friends (iii) through a search of names or usernames on Instagram then simply type a name to search and we will perform a search on our Service. Note about "Invite Friends" feature: If you choose to invite someone to the Service through our "Invite friends" feature, you may select a person directly from the contacts list on your device and send a text or email from your personal account. You understand and agree that you are responsible for any charges that apply to communications sent from your device, and because this invitation is coming directly from your personal account, Instagram does not have access to or control this communication. Analytics information: We use third-party analytics tools to help us measure traffic and usage trends for the Service. These tools collect information sent by your device or our Service, including the web pages you visit, add-ons, and other information that assists us in improving the Service. We collect and use this analytics information with analytics information from other Users so that it cannot reasonably be used to identify any particular individual User. Cookies and similar technologies: When you visit the Service, we may use cookies and similar technologies like pixels, web beacons, and local storage to collect information about how you use Instagram and provide features to you. We may ask advertisers or other partners to serve ads or services to your devices, which may use cookies or similar technologies placed by us or the third party. More information is available in our About Cookies section Log file information: Log file information is automatically reported by your browser each time you make a request to access (i.e., visit) a web page or app. It can also be provided when the content of the webpage or app is downloaded to your browser or device. When you use our Service, our servers automatically record certain log file information, including your web request, Internet Protocol ("IP") address, browser type, referring / exit pages and URLs, number of clicks and how you interact with links on the Service, domain names, landing pages, pages viewed, and other such information. We may also collect similar information from emails sent to our Users which then help us track which emails are opened and which links are clicked by recipients. The information allows for more accurate reporting and improvement of the Service. Device identifiers: When you use a mobile device like a tablet or phone to access our Service, we may access, collect, monitor, store on your device, and/or remotely store one or more "device identifiers." Device identifiers are small data files or similar data structures stored on or associated with your mobile device, which uniquely identify your mobile device. A device identifier may be data stored in connection with the device hardware, data stored in connection with the device's operating system or other software, or data sent to the device by Instagram. A device identifier may deliver information to us or to a third party partner about how you browse and use the Service and may help us or others provide reports or personalized content and ads. Some features of the Service may not function properly if use or availability of device identifiers is impaired or disabled. Metadata: Metadata is usually technical data that is associated with User Content. For example, Metadata can describe how, when and by whom a piece of User Content was collected and how that content is formatted. Users can add or may have Metadata added to their User Content including a hashtag (e.g., to mark keywords when you post a photo), geotag (e.g., to mark your location to a photo), comments or other data. This makes your User Content more searchable by others and more interactive. If you geotag your photo or tag your photo using other's APIs then, your latitude and longitude will be stored with the photo and searchable (e.g., through a location or map feature) if your photo is made public by you in accordance with your privacy settings. 2. HOW WE USE YOUR INFORMATION

In addition to some of the specific uses of information we describe in this Privacy Policy, we may use information that we receive to: help you efficiently access your information after you sign in remember information so you will not have to re-enter it during your visit or the next time you visit the Service; provide personalized content and information to you and others, which could include online ads or other forms of marketing provide, improve, test, and monitor the effectiveness of our Service develop and test new products and features monitor metrics such as total number of visitors, traffic, and demographic patterns diagnose or fix technology problems automatically update the Instagram application on your device Instagram or other Users may run contests, special offers or other events or activities ("Events") on the Service. If you do not want to participate in an Event, do not use the particular Metadata (i.e. hashtag or geotag) associated with that Event. 3. SHARING OF YOUR INFORMATION

We will not rent or sell your information to third parties outside Instagram (or the group of companies of which Instagram is a part) without your consent, except as noted in this Policy.

Parties with whom we may share your information: We may share User Content and your information (including but not limited to, information from cookies, log files, device identifiers, location data, and usage data) with businesses that are legally part of the same group of companies that Instagram is part of, or that become part of that group ("Affiliates"). Affiliates may use this information to help provide, understand, and improve the Service (including by providing analytics) and Affiliates' own services (including by providing you with better and more relevant experiences). But these Affiliates will honor the choices you make about who can see your photos. We also may share your information as well as information from tools like cookies, log files, and device identifiers and location data, with third-party organizations that help us provide the Service to you ("Service Providers"). Our Service Providers will be given access to your information as is reasonably necessary to provide the Service under reasonable confidentiality terms. We may also share certain information such as cookie data with third-party advertising partners. This information would allow third-party ad networks to, among other things, deliver targeted advertisements that they believe will be of most interest to you. We may remove parts of data that can identify you and share anonymized data with other parties. We may also combine your information with other information in a way that it is no longer associated with you and share that aggregated information. Parties with whom you may choose to share your User Content: Any information or content that you voluntarily disclose for posting to the Service, such as User Content, becomes available to the public, as controlled by any applicable privacy settings that you set. To change your privacy settings on the Service, please change your profile setting. Once you have shared User Content or made it public, that User Content may be re-shared by others. Subject to your profile and privacy settings, any User Content that you make public is searchable by other Users and subject to use under our Instagram API. The use of the Instagram API is subject to the API Terms of Use which incorporates the terms of this Privacy Policy. If you remove information that you posted to the Service, copies may remain viewable in cached and archived pages of the Service, or if other Users or third parties using the Instagram API have copied or saved that information. What happens in the event of a change of control: If we sell or otherwise transfer part or the whole of Instagram or our assets to another organization (e.g., in the course of a transaction like a merger, acquisition, bankruptcy, dissolution, liquidation), your information such as name and email address, User Content and any other information collected through the Service may be among the items sold or transferred. You will continue to own your User Content. The buyer or transferee will have to honor the commitments we have made in this Privacy Policy. Responding to legal requests and preventing harm: We may access, preserve and share your information in response to a legal request (like a search warrant, court order or subpoena) if we have a good faith belief that the law requires us to do so. This may include responding to legal requests from jurisdictions outside of the United States where we have a good faith belief that the response is required by law in that jurisdiction, affects users in that jurisdiction, and is consistent with internationally recognized standards. We may also access, preserve and share information when we have a good faith belief it is necessary to: detect, prevent and address fraud and other illegal activity; to protect ourselves, you and others, including as part of investigations; and to prevent death or imminent bodily harm. Information we receive about you may be accessed, processed and retained for an extended period of time when it is the subject of a legal request or obligation, governmental investigation, or investigations concerning possible violations of our terms or policies, or otherwise to prevent harm. 4. HOW WE STORE YOUR INFORMATION

Storage and Processing: Your information collected through the Service may be stored and processed in the United States or any other country in which Instagram, its Affiliates or Service Providers maintain facilities. Instagram, its Affiliates, or Service Providers may transfer information that we collect about you, including personal information across borders and from your country or jurisdiction to other countries or jurisdictions around the world. If you are located in the European Union or other regions with laws governing data collection and use that may differ from U.S. law, please note that we may transfer information, including personal information, to a country and jurisdiction that does not have the same data protection laws as your jurisdiction. By registering for and using the Service you consent to the transfer of information to the U.S. or to any other country in which Instagram, its Affiliates or Service Providers maintain facilities and the use and disclosure of information about you as described in this Privacy Policy. We use commercially reasonable safeguards to help keep the information collected through the Service secure and take reasonable steps (such as requesting a unique password) to verify your identity before granting you access to your account. However, Instagram cannot ensure the security of any information you transmit to Instagram or guarantee that information on the Service may not be accessed, disclosed, altered, or destroyed. Please do your part to help us. You are responsible for maintaining the secrecy of your unique password and account information, and for controlling access to emails between you and Instagram, at all times. Your privacy settings may also be affected by changes the social media services you connect to Instagram make to their services. We are not responsible for the functionality, privacy, or security measures of any other organization. 5. YOUR CHOICES ABOUT YOUR INFORMATION

Your account information and profile/privacy settings: Update your account at any time by logging in and changing your profile settings. Unsubscribe from email communications from us by clicking on the "unsubscribe link" provided in such communications. As noted above, you may not opt out of Service-related communications (e.g., account verification, purchase and billing confirmations and reminders, changes/updates to features of the Service, technical and security notices). Learn more about reviewing or modifying your account information. How long we keep your User Content: Following termination or deactivation of your account, Instagram, its Affiliates, or its Service Providers may retain information (including your profile information) and User Content for a commercially reasonable time for backup, archival, and/or audit purposes. Learn more about deleting your account. 6. CHILDREN'S PRIVACY

Instagram does not knowingly collect or solicit any information from anyone under the age of 13 or knowingly allow such persons to register for the Service. The Service and its content are not directed at children under the age of 13. In the event that we learn that we have collected personal information from a child under age 13 without parental consent, we will delete that information as quickly as possible. If you believe that we might have any information from or about a child under 13, please contact us.

7. OTHER WEB SITES AND SERVICES

We are not responsible for the practices employed by any websites or services linked to or from our Service, including the information or content contained within them. Please remember that when you use a link to go from our Service to another website or service, our Privacy Policy does not apply to those third-party websites or services. Your browsing and interaction on any third-party website or service, including those that have a link on our website, are subject to that third party's own rules and policies. In addition, you agree that we are not responsible and do not have control over any third-parties that you authorize to access your User Content. If you are using a third-party website or service and you allow them to access your User Content you do so at your own risk.

8. HOW TO CONTACT US ABOUT A DECEASED USER

In the event of the death of an Instagram User, please contact us. We will usually conduct our communication via email; should we require any other information, we will contact you at the email address you have provided in your request.

9. HOW TO CONTACT US

If you have any questions about this Privacy Policy or the Service, please find the appropriate support channel in the Help Center at which to contact us.

10. CHANGES TO OUR PRIVACY POLICY

Instagram may modify or update this Privacy Policy from time to time, so please review it periodically. We may provide you additional forms of notice of modifications or updates as appropriate under the circumstances. Your continued use of Instagram or the Service after any modification to this Privacy Policy will constitute your acceptance of such modification.

Lufthansa

Any personal information collected during visits to our websites is processed in accordance with the provisions of German law. Our data protection policy also complies with the internal data protection guidelines of the Lufthansa Group.

Our websites may contain links to other providers' websites that are not covered by this privacy statement.

Lufthansa respects your personal privacy. Collection and processing of personal data

Lufthansa stores your personal information if you provide it to us yourself, for example, during registration, as part of a survey or competition, or during contractual transactions (e.g. booking a flight).

Furthermore Lufthansa analyses traffic on its website in order to understand our customers' requirements and, based on these, to continually improve our site. For this reason we store the IP address of a visitor's Internet Service Provider as standard. The IP address is not linked to a specific person. Only anonymous, aggregate data is evaluated for statistical purposes during web analysis.

Data from individual uses of our website is stored for error analysis. This information is used solely for correcting errors and is deleted after a two week storage deadline.

Lufthansa uses cookies to track visitor preferences and to improve the design of its websites accordingly.

When a rating is submitted for a website , anonymised and non-personal data will be forwarded to a third party supplier. Use of the web analysis tool Webtrends

Lufthansa carries out access measurements on Lufthansa websites using the web analysis tool Webtrends. Cookies are used for measurement and the access data is collected in an anonymised form so that no connection can be made to a user. In particular this is done by anonymising the IP address.

The information produced about the use of the websites is transferred in anonymised form to the statistics server (statse.webtrendslive.com) that is operated by Webtrends Inc, 555 SW Oak Street, Suite 300, Portland, OR 97204, USA ("Webtrends") in the United States.

Only authorised persons have access to this anonymised data. Limitations on use and disclosure of personal data

Lufthansa will process and use your personal data only in connection with services related to the website LH.com and the Lufthansa newsletter. This enables us to offer you a customized service and/or can also save you from having to enter the same information twice. We only utilize as much information as is necessary.

We will not share your personal data with third parties. Any partner services offered on LH.com are booked directly with these partners. You will be advised of this accordingly.

Some of our services give you the opportunity to store personal information (e.g. FAQ). Such information will only be used to respond to your current inquiry. After your inquiry is closed, this information will not be retained without requesting your permission to do so.

If Government agencies or authorities ask us to collect or share personal data, we will do so only within the appropriate legal frameworks. We require our employees, suppliers and partners to maintain confidentiality and data secrecy in accordance with Article 5 of the German Federal Data Protection Act.

All airlines are required by law to provide the US Customs and Border Protection with the flight and reservation details of every passenger entering the USA. This information is to be used exclusively for security purposes.

Further information is available here.

Opt-in/opt-out

We would like to use your data to tell you about our products and services and those of our partners, and also to conduct occasional surveys. You decide whether or not you would like to be contacted by us. If you change your mind, you can alter your preferences at any time in the 'Newsletter and SMS services' section under 'My Account' > 'My Profile'. Accessing personal data

If requested, Lufthansa will confirm whether we store personal data about you and what this data is. Despite our best efforts to ensure that this stored data is accurate and up-to-date, it may be incorrect. If it is, we will correct it on request. You can update most of your personal details yourself at any time under 'My Account' > 'My Profile'. If changes occur to your personal details, please update your profile as soon as possible. Data security

Lufthansa uses technical and organizational security measures to protect the data we hold about you against accidental or deliberate manipulation, loss, deletion or unauthorized access. Our security measures are being improved continuously as new technology develops. The processing and transmission of data is encrypted by the SSL (Secure Sockets Layer) protocol.

C.2.2 Preprocessed Statements

ID	source	requirement	cleaned_requirement
0	klm	We may collect and process the following categories of personal information	We may collect and process the following categories
3	klm	When you create a personal account or register for a service, we may also record your sign in details and other information you fill out on your personal account or registration form	, we may also record your sign you fill out
4	klm	For business travellers, we also collect information relating to your company such as company name and business location	, we also collect information relating
5	klm	When you make a reservation or book a flight with us, we process your reservation and booking information	, we process your reservation and booking information

7	klm	In addition hereto we process information in relation to ancillary services (such as upgrades and extra luggage) and products you purchase (for example via the KLM webshop)	we process information in relation to ancillary services (such as upgrades and extra luggage) and products you purchase ( )
8	klm	When you travel with us, we process information in relation to your journey	, we process information
10	klm	We may also record any specified medical needs or dietary requests you have and any additional assistance you require	We may also record any specified medical needs or dietary requests you have and any additional assistance you require
11	klm	When you become a member of our loyalty programs, we process your membership number, miles or credits balance, awards and benefits, type and level of membership and other information in relation to your membership	, we process your membership number, miles or credits balance, awards and benefits, type and level
12	klm	When you send us an email or chat with us online or via social media, we register your communication with us	, we register your communication
13	klm	When you call us, our customer support will register your questions or complaints in our database	, our customer support will register your questions or complaints
14	klm	We may also record telephone calls for training purposes or to prevent or combat fraud	We may also record telephone calls
15	klm	When you visit our websites or use one of our apps, we may register your IP address, browser type, operating system, referring website, web-browsing behavior and app use	use one , we may register your IP address, browser type, operating system, referring website, web-browsing behavior and app use
16	klm	We may receive an automatic notification when you open a newsletter or click on a link in such newsletter	We may receive an automatic notification when you open a newsletter or click
17	klm	With your permission we may also receive your location data	With your permission we may also receive your location data
20	klm	Depending on your social network settings we may receive information from your social network provider	we may receive information
21	klm	For example when you sign in for our services using a social network account, we may receive your social network profile including your contact details, interests and contacts	, we may receive your social network profile including your contact details, interests and contacts

23	klm	You may choose to share information with us, for example when you leave a comment for us on Facebook, fill out a customer survey or submit an entry for a contest	You may choose to share information , , fill out a customer survey or submit an entry
25	klm	We may be required to collect, use and share such data with third parties for the purposes as described in this privacy policy	We may be required to collect,
26	klm	For example, we keep records of certain passengers who have been found to be carrying illegal drugs or who have interfered with the safety	, we keep records
28	klm	By providing any personal information that is, or could be considered to be, sensitive personal information, you agree that we may collect, use and share this information with third parties as described in this privacy policy	By providing any personal information that is, or could be considered to be, sensitive personal information, you agree that we may collect, use and share this information
29	klm	If you withdraw your consent, it may mean we will not be able to provide all or parts of the services you have requested from us	, it may mean we will not be able to provide all or parts you have requested
30	klm	Please be aware that in these circumstances you will not be able to cancel or obtain a refund of any fees you have paid	Please be aware that you will not be able to cancel or obtain a refund you have paid
32	klm	When you use our website or mobile apps we collect information via cookies and similar technologies	we collect information
35	klm	For specific services, apps or events, we may collect other types of data and use such data for different purposes than described in this privacy policy	For specific services, apps or events, we may collect other types
36	klm	We will inform you about this when you register for a service or an event or download an app	We will inform you
37	klm	KLM may obtain your personal data in a number of ways, for example when you book a flight, you register for one of our loyalty programs, use one of our apps, communicate with us via social media or when you subscribe to our newsletter	KLM may obtain your personal data in a number , , you register for one , use one , communicate

38	klm	We may also receive personal information from our group companies, partners or from service providers	We may also receive personal information from our group companies, partners or
39	klm	For example, when you use a third party platform to search and book a flight, we may receive your contact details from that provider	, book a flight, we may receive your contact details
40	klm	When you purchase a service from our loyalty partners, they share your earned miles with us	, they share your earned miles
41	klm	Depending on your social network settings, we may also receive information from your social network provider (please see above)	, we may also receive information (please see above)
43	klm	To handle your reservations and bookings and to fulfill your travel arrangements and purchases, we need to process most of the information described above	To handle your reservations and bookings and to fulfill your travel arrangements and purchases, we need to process most
44	klm	For example, we need your name, passport number and other identifying information to issue your ticket	, we need your name, passport number and other identifying information
45	klm	To inform you about changes in departure times, we need your contact details	, we need your contact details
46	klm	And, to ensure that you receive the required care, we require your specified medical needs	And, to ensure that you receive the required care, we require your specified medical needs
47	klm	To let you or your company benefit from the discounts and rewards under our loyalty programs, we use your membership information and your booking information	To let , we use your membership information and your booking information
48	klm	For example, we use your name and flight details when you check-in for your flight with our app	, we use your name and flight details when you check-in
49	klm	Some of our online services and apps use your location, for example to show you the nearest location of your interest	Some use your location, to show you the nearest location
50	klm	To ease your use of our online services or apps, we may analyze the data we collect when you use our digital media and combine it with information collected via cookies and similar technologies (please see above)	, we may analyze the data we collect (please see above)

52	klm	We use your contact details to communicate with you, answer your questions or handle your complaints	We use your contact details to communicate , answer your questions or handle your complaints
54	klm	To perform our research, we may merge and analyze the different types of data as described above	, we may merge and analyze the different types
55	klm	We will only use aggregated data and do not use name, email address or other directly identifying information	We will only use aggregated data and do not use name, email address or other directly identifying information
56	klm	We may also combine such aggregated data with information we receive from our group companies or from public sources (such as Statistics Netherlands)	We may also combine such aggregated data we receive from our group companies or from public sources ()
57	klm	Without your consent we will not use sensitive data for this statistical research	we will not use sensitive data
59	klm	We may use your personal information to send you newsletters, magazines, promotions or other marketing communications	We may use your personal information to send you newsletters, magazines, promotions or other marketing communications
60	klm	To understand what is relevant to you, we may use automatic tools to analyze your personal information	, we may use automatic tools to analyze your personal information
61	klm	For this purpose we may use and combine the information described above	we may use and combine the information described
62	klm	We may also combine such aggregated data with information we receive from our group companies or from public sources (such as Statistics Netherlands)	We may also combine such aggregated data we receive from our group companies or from public sources ()
63	klm	Without your consent we will not use sensitive data for this statistical research	we will not use sensitive data
64	klm	We use the results of our analysis to tailor our marketing communications to your specific interests and preferences	We use the results to tailor our marketing communications
66	klm	We may use different channels for our marketing communications, such as e-mail, apps, social media and your personal online account	We may use different channels for our marketing communications, such as e-mail, apps, social media and your personal online account
69	klm	We may share your email address or another identifier to Facebook to enable Facebook to determine if you have a Facebook account	We may share your email address or another identifier to Facebook to enable Facebook to determine if you have a Facebook account

77	klm	You may request permission to examine this data or to make changes to its accuracy by submitting this request in writing to the Royal Netherlands Marechaussee at P	You may request permission to examine this data or
79	klm	If you reside on Aruba, the Netherlands Antilles, in Suriname or in Venezuela, you should accompany this written request with a copy of your passport	If you reside on Aruba, the Netherlands Antilles, , you should accompany this
81	facebook	Depending on which Services you use, we collect different kinds of information from or about you	, we collect different kinds
83	facebook	We collect the content and other information you provide when you use our Services; including when you sign up for an account; create or share; and message or communicate with others	We collect the content and other information you provide ; including ; create or share; and message or communicate
85	facebook	We also collect information about how you use our Services; such as the types of content you view or engage with or the frequency and duration of your activities	We also collect information ;
87	facebook	We also collect content and information that other people provide when they use our Services; including information about you; such as when they share a photo of you; send a message to you; or upload; sync or import your contact information	We also collect content and information that other people provide ; ; such as when they share a photo ; send a message to you; or upload; sync or import your contact information
89	facebook	We collect information about the people and groups you are connected to and how you interact with them; such as the people you communicate with the most or the groups you like to share with	We collect information how you interact ;
90	facebook	We also collect contact information you provide if you upload; sync or import this information (such as an address book) from a device	We also collect contact information you provide if you upload; sync or import this information (such as an address book)
92	facebook	If you use our Services for purchases or financial transactions (like when you buy something on Facebook; make a purchase in a game; or make a donation); we collect information about the purchase or transaction	If you use our Services for purchases or financial transactions (like ; make a purchase ; or make a donation); we collect information

94	facebook	We collect information from or about the computers; phones; or other devices where you install or access our Services; depending on the permissions you've granted	We collect information from or about the computers; phones; or other devices where you install or access our Services; depending on the permissions you've granted
95	facebook	We may associate the information we collect from your different devices; which helps us provide consistent Services across your devices	We may associate the information we collect from your different devices; which helps us provide consistent Services
101	facebook	We collect information when you visit or use third-party websites and apps that use our Services (like when they offer our Like button or Facebook Log In or use our measurement and advertising services)	We collect information when you visit or use third-party websites and apps that use our Services ( )
104	facebook	We receive information about you and your activities on and off Facebook from third-party partners; such as information from a partner when we jointly offer services or from an advertiser about your experiences or interactions with them	We receive information Facebook from third-party partners;
105	facebook	We receive information about you from companies that are owned or operated by Facebook; in accordance with their terms and policies	We receive information ;
112	facebook	We also use information we have to provide shortcuts and suggestions to you	We also use information we have to provide shortcuts and suggestions
115	facebook	When we have location information; we use it to tailor our Services for you and others; like helping you to check-in and find local events or offers in your area or tell your friends that you are nearby	; we use it to tailor our Services ; like helpg you to check- and fd local events or offers your area or tell your friends that you are nearby
117	facebook	We use your information to send you marketing communications; communicate with you about our Services and let you know about our policies and terms	We use your information to send you marketing communications; communicate and let you know
118	facebook	We also use your information to respond to you when you contact us	We also use your information to respond
120	facebook	We use the information we have to improve our advertising and measurement systems so we can show you relevant ads on and off our Services and measure the effectiveness and reach of ads and services	We use the information we have to improve our advertising and measurement systems

123	facebook	We use the information we have to help verify accounts and activity; and to promote safety and security on and off of our Services; such as by investigating suspicious activity or violations of our terms or policies	We use the information we have to help verify accounts and activity; and to promote safety and security ;
140	facebook	When you comment on another person's post or like their content on Facebook; that person decides the audience who can see your comment or like	When you comment on another person's post or like their content ; that person decides the audience who can see your comment or like
143	facebook	Other people may use our Services to share content about you with the audience they choose	Other people may use our Services to share content they choose
147	facebook	When you use third-party apps; websites or other services that use; or are integrated with; our Services; they may receive information about what you post or share	When you use third-party apps; websites or other services that use; or are integrated ; our Services; they may receive information
149	facebook	In addition; when you download or use such third-party services; they can access your Public Profile; which includes your username or user ID; your age range and country/language; your list of friends; as well as any information that you share with them	; when you download or use such third-party services; they can access your Public Profile; which includes your username or user ID; your age range and country/language; your list of friends; as well as any information that you share
153	facebook	We share information we have about you within the family of companies that are part of Facebook	We share information we have
155	facebook	If the ownership or control of all or part of our Services or their assets changes; we may transfer your information to the new owner	; we may transfer your information
161	facebook	With this in mind; we use all of the information we have about you to show you relevant ads	; we use all
162	facebook	We do not share information that personally identifies you (personally identifiable information is information like name or email address that can by itself be used to contact you or identifies who you are) with advertising; measurement or analytics partners unless you give us permission	We do not share information that personally identifies you (personally identifiable information is information be used to contact you or identifies who you are) with advertising; measurement or analytics partners

163	facebook	We may provide these partners with information about the reach and effectiveness of their advertising without providing information that personally identifies you; or if we have aggregated the information so that it does not personally identify you	We may provide these partners ; or if we have aggregated the information
166	facebook	You can adjust your ad preferences if you want to control and manage your ad experience on Facebook	You can adjust your ad preferences
168	facebook	We transfer information to vendors; service providers; and other partners who globally support our business; such as providing technical infrastructure services; analyzing how our Services are used; measuring the effectiveness of ads and services; providing customer service; facilitating payments; or conducting academic research and surveys	We transfer information to vendors; service providers; and other partners who globally support our business; such as providing technical infrastructure services; analyzing how our Services are used; measuring the effectiveness ; providing customer service; facilitating payments; or conducting academic research and surveys
171	facebook	You can manage the content and information you share when you use Facebook through the Activity Log tool	You can manage the content and information you share
172	facebook	You can also download information associated with your Facebook account through our Download Your Information tool	You can also download information associated
173	facebook	We store data for as long as it is necessary to provide products and services to you and others; including those described above	We store data ;
174	facebook	Information associated with your account will be kept until your account is deleted; unless we no longer need the data to provide products and services	Information associated will be kept ;
175	facebook	You can delete your account any time	You can delete your account any time
176	facebook	When you delete your account; we delete things you have posted; such as your photos and status updates	; we delete things you have posted; status updates
177	facebook	If you do not want to delete your account; but want to temporarily stop using Facebook; you may deactivate your account instead	If you do not want to delete your account; but want to temporarily stop using Facebook; you may deactivate your account instead

179	facebook	Keep in mind that information that others have shared about you is not part of your account and will not be deleted when you delete your account	Keep that information that others have shared you is not part and will not be deleted
181	facebook	We may access, preserve and share your information in response to a legal request (like a search warrant, court order or subpoena) if we have a good faith belief that the law requires us to do so	We may access, preserve and share your information in response to a legal request (like a search warrant, court order or subpoena)
183	facebook	We may also access; preserve and share information when we have a good faith belief it is necessary to: detect; prevent and address fraud and other illegal activity; to protect ourselves; you and others; including as part of investigations; or to prevent death or imminent bodily harm	We may also access; preserve and share information when we have a good faith belief it is necessary to: detect; prevent and address fraud and other illegal activity; ; you and others; including ; or to prevent death or imminent bodily harm
184	facebook	For example; we may provide information to third-party partners about the reliability of your account to prevent fraud and abuse on and off of our Services	; we may provide information to third-party partners
185	facebook	Information we receive about you; including financial transaction data related to purchases made with Facebook; may be accessed; processed and retained for an extended period of time when it is the subject of a legal request or obligation; governmental investigation; or investigations concerning possible violations of our terms or policies; or otherwise to prevent harm	Information we receive ; ; may be accessed; processed and retained ; governmental investigation; or investigations concerning possible violations ; or
186	facebook	We also may retain information from accounts disabled for violations of our terms for at least a year to prevent repeat abuse or other violations of our terms	We also may retain information from accounts
188	facebook	Facebook may share information internally within our family of companies or with third parties for purposes described in this policy	Facebook may share information internally

189	facebook	Information collected within the European Economic Area ('EEA') may; for example; be transferred to countries outside of the EEA for the purposes as described in this policy	Information collected ('EEA') may; ; be transferred
191	facebook	You can contact us using the information provided below with questions or concerns	You can contact us using the information provided below
192	facebook	We also may resolve disputes you have with us in connection with our privacy policies and practices through TRUSTe	We also may resolve disputes you have
196	google	We collect information to provide better services to all of our users from figuring out basic stuff like which language you speak, to more complex things like which ads you will find most useful, the people who matter most to you online, or which YouTube videos you might like	We collect information to provide better services from figuring out basic stuff which language you speak, which ads you will find most useful, the people who matter most online, or which YouTube videos you might
197	google	We collect information in the following ways	We collect information
199	google	When you do, we will ask for personal information, like your name, email address, telephone number or credit card to store with your account	, we will ask for personal information, like your name, email address, telephone number or credit card to store
200	google	If you want to take full advantage of the sharing features we offer, we might also ask you to create a publicly visible Google Profile, which may include your name and photo	features we offer, we might also ask you to create a publicly visible Google Profile, which may include your name and photo
202	google	We collect information about the services that you use and how you use them, like when you watch a video on YouTube, visit a website that uses our advertising services, or view and interact with our ads and content	We collect information about the services that you use and how you use them, like , visit a website that uses our advertising services, or view and interact
203	google	We collect device-specific information (such as your hardware model, operating system version, unique device identifiers, and mobile network information including phone number)	We collect device-specific information (such as your hardware model, operating system version, unique device identifiers, and mobile network information including phone number)
204	google	Google may associate your device identifiers or phone number with your Google Account	Google may associate your device identifiers or phone number

205	google	When you use our services or view content provided by Google, we automatically collect and store certain information in server logs	, we automatically collect and store certain information logs
210	google	When you use Google services, we may collect and process information about your actual location	, we may collect and process information
211	google	We use various technologies to determine location, including IP address, GPS, and other sensors that may, for example, provide Google with information on nearby devices, Wi-Fi access points and cell towers	We use various technologies to determine location, including IP address, GPS, and other sensors that may, , provide Google , Wi-Fi access points and cell towers
213	google	This number and information about your installation (for example, the operating system type and application version number) may be sent to Google when you install or uninstall that service or when that service periodically contacts our servers, such as for automatic updates	This number and information about your installation (, the operating system type and application version number) may be sent when you install or uninstall that service or when that service periodically contacts our servers,
214	google	We may collect and store information (including personal information) locally on your device using mechanisms such as browser web storage (including HTML 5) and application data caches	We may collect and store information () locally using mechanisms such as browser web storage (including HTML 5) and application data caches
217	google	Our Google Analytics product helps businesses and site owners analyze the traffic to their websites and apps	Our Google Analytics product helps businesses and site owners analyze the traffic
219	google	Information we collect when you are signed in to Google, in addition to information we obtain about you from partners, may be associated with your Google Account	Information we collect when you are signed , we obtain , may be associated
220	google	When information is associated with your Google Account, we treat it as personal information	, we treat it
222	google	We use the information we collect from all of our services to provide, maintain, protect and improve them, to develop new ones, and to protect Google and our users	We use the information we collect , maintain, protect and improve them, to develop new ones, and to protect Google and our users
223	google	We also use this information to offer you tailored content, like giving you more relevant search results and ads	We also use this information to offer you tailored content,

224	google	We may use the name you provide for your Google Profile across all of the services we offer that require a Google Account	We may use the name you provide
225	google	In addition, we may replace past names associated with your Google Account so that you are represented consistently across all our services	, we may replace past names associated
226	google	If other users already have your email, or other information that identifies you, we may show them your publicly visible Google Profile information, such as your name and photo	If other users already have your email, or other information that identifies you, we may show them your publicly visible Google Profile information,
227	google	If you have a Google Account, we may display your Profile name, Profile photo, and actions you take on Google or on third-party applications connected to your Google Account (such as +1ś, reviews you write and comments you post) in our services, including displaying in ads and other commercial contexts	, we may display your Profile name, Profile photo, and actions you take on Google or on third-party applications connected (+1ś, reviews you write and comments you post) ,
229	google	When you contact Google, we keep a record of your communication to help solve any issues you might be facing	, we keep a record to help solve any issues you might be facing
230	google	We may use your email address to inform you about our services, such as letting you know about upcoming changes or improvements	We may use your email address to inform you about our services,
231	google	We use information collected from cookies and other technologies, like pixel tags, to improve your user experience and the overall quality of our services	We use information collected from cookies and other technologies, ,
234	google	When showing you tailored ads, we will not associate an identifier from cookies or similar technologies with sensitive categories, such as those based on race, religion, sexual orientation or health	, we will not associate an identifier from cookies or similar technologies with sensitive categories, such as those based , religion, sexual orientation or health
235	google	Our automated systems analyze your content (including emails) to provide you personally relevant product features, such as customized search results, tailored advertising, and spam and malware detection	Our automated systems analyze your content () to provide you personally relevant product features, such as customized search results, tailored advertising, and spam and malware detection

236	google	We may combine personal information from one service with information, including personal information, from other Google services, for example to make it easier to share things with people you know	We may combine personal information , including personal information, ,
237	google	Depending on your account settings, your activity on other sites and apps may be associated with your personal information in order to improve Googleś services and the ads delivered by Google	, your activity may be associated in order to improve Googleś services and the ads delivered by Google
238	google	We will ask for your consent before using information for a purpose other than those that are set out in this Privacy Policy	We will ask
239	google	Google processes personal information on our servers in many countries around the world	Google processes personal information
240	google	We may process your personal information on a server located outside the country where you live	We may process your personal information
247	google	You can also visit that page to opt out of certain Google advertising services	You can also visit that page to opt
252	google	You may also set your browser to block all cookies, including cookies associated with our services, or to indicate when a cookie is being set by us	You may also set your browser to block all cookies, including cookies associated , or to indicate
254	google	For example, we may not remember your language preferences	, we may not remember your language preferences
257	google	Our services provide you with different options on sharing and removing your content	Our services provide you
261	google	When updating your personal information, we may ask you to verify your identity before we can act on your request	, we may ask you to verify your identity
265	google	Because of this, after you delete information from our services, we may not immediately delete residual copies from our active servers and may not remove information from our backup systems	, , we may not immediately delete residual copies and may not remove information

266	google	We do not share personal information with companies, organizations and individuals outside of Google unless one of the following circumstances applies	We do not share personal information with companies, organizations and individuals
267	google	We will share personal information with companies, organizations or individuals outside of Google when we have your consent to do so	We will share personal information with companies, organizations or individuals
268	google	We require opt-in consent for the sharing of any sensitive personal information	We require opt-
277	google	We provide personal information to our affiliates or other trusted businesses or persons to process it for us, based on our instructions and in compliance with our Privacy Policy and any other appropriate confidentiality and security measures	We provide personal information to our affiliates or other trusted businesses or persons , based
278	google	We will share personal information with companies, organizations or individuals outside of Google if we have a good-faith belief that access, use, preservation or disclosure of the information is reasonably necessary to	We will share personal information with companies, organizations or individuals if we have a good-faith belief that access, use, preservation or disclosure is reasonably necessary
283	google	We may share non-personally identifiable information publicly and with our partners' like publishers, advertisers or connected sites	We may share non-personally identifiable information publicly and with our partners' like publishers, advertisers or connected sites
284	google	For example, we may share information publicly to show trends about the general use of our services	, we may share information publicly
285	google	If Google is involved in a merger, acquisition or asset sale, we will continue to ensure the confidentiality of any personal information and give affected users notice before personal information is transferred or becomes subject to a different privacy policy	If Google is involved in a merger, acquisition or asset sale, we will continue to ensure the confidentiality and give affected users notice
289	google	We review our information collection, storage and processing practices, including physical security measures, to guard against unauthorized access to systems	We review our information collection, storage and processing practices, ,

298	google	When we receive formal written complaints, we will contact the person who made the complaint to follow up	, we will contact the person who made the complaint
301	google	We will not reduce your rights under this Privacy Policy without your explicit consent	We will not reduce your rights
302	google	We will post any privacy policy changes on this page and, if the changes are significant, we will provide a more prominent notice (including, for certain services, email notification of privacy policy changes)	We will post any privacy policy changes , , we will provide a more prominent notice (including, for certain services, email notification)
303	google	We will also keep prior versions of this Privacy Policy in an archive for your review	We will also keep prior versions
314	lufthansa	Our websites may contain links to other providersẃebsites that are not covered by this privacy statement	Our websites may contain links to other providersẃebsites that are not covered by this privacy statement
315	lufthansa	Lufthansa stores your personal information if you provide it to us yourself, for example, during registration, as part of a survey or competition, or during contractual transactions (for example booking a flight)	Lufthansa stores your personal information if you provide it yourself, , during registration, as part of a survey or competition, or during contractual transactions (booking a flight)
316	lufthansa	Lufthansa uses cookies to track visitor preferences and to improve the design of its websites accordingly	Lufthansa uses cookies to track visitor preferences and to improve the design accordingly
317	lufthansa	When a rating is submitted for a website , anonymised and non-personal data will be forwarded to a third party supplier	, anonymised and non-personal data will be forwarded
318	lufthansa	Lufthansa will process and use your personal data only in connection with services related to the website LH.com and the Lufthansa newsletter	Lufthansa will process and use your personal data only in connection with services related to the website LH.com and the Lufthansa newsletter
319	lufthansa	We only utilize as much information as is necessary	We only utilize
320	lufthansa	We will not share your personal data with third parties	We will not share your personal data
321	lufthansa	After your inquiry is closed, this information will not be retained without requesting your permission to do so	, this information will not be retained

322	lufthansa	If Government agencies or authorities ask us to collect or share personal data, we will do so only within the appropriate legal frameworks	, we will do so
323	lufthansa	We require our employees, suppliers and partners to maintain confidentiality and data secrecy in accordance with Article 5 of the German Federal Data Protection Act	We require our employees, suppliers and partners to maintain confidentiality and data secrecy in accordance with Article 5
324	lufthansa	We would like to use your data to tell you about our products and services and those of our partners, and also to conduct occasional surveys	We would like to use your data to tell you about our products and services and those , and
325	lufthansa	If you change your mind, you can alter your preferences at any time in the Newsletter and SMS services' section under 'My Account' > 'My Profile'	, you can alter your preferences at any time in the 'Newsletter and SMS services' section under 'My Account' > 'My Profile'
326	lufthansa	If requested, Lufthansa will confirm whether we store personal data about you and what this data is	, Lufthansa will confirm whether we store personal data what this data is
327	lufthansa	Lufthansa uses technical and organizational security measures to protect the data we hold about you against accidental or deliberate manipulation, loss, deletion or unauthorized access	Lufthansa uses technical and organizational security measures to protect the data we hold about you against accidental or deliberate manipulation, loss, deletion or unauthorized access
328	instagram	We collect the following types of information	We collect the following types
329	instagram	For example, we may send you Service-related emails (for example, account verification, changes/updates to features of the Service, technical and security notices)	, we may send you Service-related emails (, account verification, changes/updates to features of the Service, technical and security notices)
330	instagram	If you choose, you can use our "Find friends" feature to locate other people with Instagram accounts either through (i) your contacts list, (ii) third-party social media sites or (iii) through a search of names and usernames on Instagram	, you can use our "Find friends" feature to locate other people accounts either through (i) your contacts list, (ii) third-party social media sites or (iii)
331	instagram	If you choose to find your friends through (i) your deviceś contacts list, then Instagram will access your contacts list to determine whether or not someone associated with your contact is using Instagram	If you choose to find your friends (i) your deviceś contacts list, then Instagram will access your contacts list

332	instagram	If you choose to find your friends through a (ii) third-party social media site, then you will be prompted to set up a link to the third-party service and you understand that any information that such service may provide to us will be governed by this Privacy Policy	If you choose to find your friends through a (ii) third-party social media site, then you will be prompted to set up a link to the third-party service and you understand that any information that such service may provide will be governed by this Privacy Policy
333	instagram	If you choose to find your friends (iii) through a search of names or usernames on Instagram then simply type a name to search and we will perform a search on our Service	If you choose to find your friends (iii) then simply type a name
334	instagram	If you choose to invite someone to the Service through our "Invite friends" feature, you may select a person directly from the contacts list on your device and send a text or email from your personal account	If you choose to invite someone to the Service through our "Invite friends" feature, you may select a person directly and send a text or email
335	instagram	We use third-party analytics tools to help us measure traffic and usage trends for the Service	We use third-party analytics tools to help us measure traffic and usage trends
336	instagram	We collect and use this analytics information with analytics information from other Users so that it cannot reasonably be used to identify any particular individual User	We collect and use this analytics information so that it cannot reasonably be used to identify any particular individual User
337	instagram	When you visit the Service, we may use cookies and similar technologies like pixels, web beacons, and local storage to collect information about how you use Instagram and provide features to you	, we may use cookies and similar technologies like pixels, web beacons, and local storage to collect information
338	instagram	We may ask advertisers or other partners to serve ads or services to your devices, which may use cookies or similar technologies placed by us or the third party	We may ask advertisers or other partners to serve ads or services to your devices, which may use cookies or similar technologies placed by us or the third party
339	instagram	We may also collect similar information from emails sent to our Users which then help us track which emails are opened and which links are clicked by recipients	We may also collect similar information

340	instagram	When you use a mobile device like a tablet or phone to access our Service, we may access, collect, monitor, store on your device, and/or remotely store one or more “device identifiers”	, we may access, collect, monitor, store , and/or remotely store one or more “device identifiers”
341	instagram	If you geotag your photo or tag your photo using otherś APIs then, your latitude and longitude will be stored with the photo and searchable (for example, through a location or map feature) if your photo is made public by you in accordance with your privacy settings	If you geotag your photo or tag your photo using otherś APIs then, your latitude and longitude will be stored with the photo and searchable (,)
342	instagram	In addition to some of the specific uses of information we describe in this Privacy Policy, we may use information that we receive to	, we may use information that we receive
343	instagram	We will not rent or sell your information to third parties outside Instagram (or the group of companies of which Instagram is a part) without your consent, except as noted in this Policy	We will not rent or sell your information to third parties outside Instagram (or the group of which Instagram is a part) ,
344	instagram	We may share User Content and your information (including but not limited to, information from cookies, log files, device identifiers, location data, and usage data) with businesses that are legally part of the same group of companies that Instagram is part of, or that become part of that group ("Affiliates")	We may share User Content and your information (including but not limited to, information from cookies, log files, device identifiers, location data, and usage data) with businesses that are legally part that Instagram is part of, or that become part of that group ("Affiliates")
345	instagram	Affiliates may use this information to help provide, understand, and improve the Service (including by providing analytics) and Affiliatesówn services (including by providing you with better and more relevant experiences)	Affiliates may use this information to help provide, understand, and improve the Service (including by providing analytics) and Affiliatesówn services (including)
346	instagram	We also may share your information as well as information from tools like cookies, log files, and device identifiers and location data, with third-party organizations that help us provide the Service to you ("Service Providers")	We also may share your information as well as information from tools like cookies, log files, and device identifiers and location data, with third-party organizations that help us provide the Service to you ("Service Providers")

347	instagram	Our Service Providers will be given access to your information as is reasonably necessary to provide the Service under reasonable confidentiality terms	Our Service Providers will be given access
348	instagram	We may also share certain information such as cookie data with third-party advertising partners	We may also share certain information such as cookie data with third-party advertising partners
349	instagram	We may remove parts of data that can identify you and share anonymized data with other parties	We may remove parts
350	instagram	We may also combine your information with other information in a way that it is no longer associated with you and share that aggregated information	We may also combine your information share that aggregated information
351	instagram	If we sell or otherwise transfer part or the whole of Instagram or our assets to another organization (for example, in the course of a transaction like a merger, acquisition, bankruptcy, dissolution, liquidation), your information such as name and email address, User Content and any other information collected through the Service may be among the items sold or transferred	If we sell or otherwise transfer part or the whole to another organization (, in the course of a transaction like a merger, acquisition, bankruptcy, dissolution, liquidation), your information such as name and email address, User Content and any other information collected may be
352	instagram	We may access, preserve and share your information in response to a legal request (like a search warrant, court order or subpoena) if we have a good faith belief that the law requires us to do so	We may access, preserve and share your information in response to a legal request (like a search warrant, court order or subpoena)
353	instagram	We may also access, preserve and share information when we have a good faith belief it is necessary to detect, prevent and address fraud and other illegal activity; to protect ourselves, you and others, including as part of investigations; and to prevent death or imminent bodily harm	We may also access, preserve and share information when we have a good faith belief it is necessary to detect, prevent and address fraud and other illegal activity; to protect ourselves, you and others, including ; and to prevent death or imminent bodily harm

354	instagram	Information we receive about you may be accessed, processed and retained for an extended period of time when it is the subject of a legal request or obligation, governmental investigation, or investigations concerning possible violations of our terms or policies, or otherwise to prevent harm	Information we receive may be accessed, processed and retained for an extended period of time when it is the subject of a legal request or obligation, governmental investigation, or investigations concerning possible violations of our terms or policies, or otherwise to prevent harm
355	instagram	Your information collected through the Service may be stored and processed in the United States or any other country in which Instagram, its Affiliates or Service Providers maintain facilities	Your information collected may be stored and processed , its Affiliates or Service Providers maintain facilities
356	instagram	Instagram, its Affiliates, or Service Providers may transfer information that we collect about you, including personal information across borders and from your country or jurisdiction to other countries or jurisdictions around the world	Instagram, its Affiliates, or Service Providers may transfer information that we collect ,
357	instagram	We will usually conduct our communication via email; should we require any other information, we will contact you at the email address you have provided in your request	We will usually conduct our communication ; should we require any other information, we will contact you
358	instagram	We may provide you additional forms of notice of modifications or updates as appropriate under the circumstances	We may provide you additional forms

C.3 Content Analysis Protocol

Introduction
For my thesis project I subtract components from privacy requirements. Privacy requirements are extracted from privacy statements. I try to develop an algorithm that subtract those components automatically from the text. However, in order to test how well this algorithm performs it needs to be tested. The test entails a comparison of the labeled output from the algorithm and a test set. The test set contains the same type of results as the output from the algorithm (i.e. labeled components), but in the test set the components are labeled manually by the participants.

You, as a participant, are one of the test subjects that will label different components from a list of privacy requirements.

A table was created containing a list of requirements (sentences) together with a ID (i.e.

the number of requirements in the list) and the source (i.e. the source/owner of the original privacy statement). The following columns were added afterwards: valid (i.e. if the tagger considers the sentences to be a requirement, based on the presence of mandatory components) and a column for each components that needs to be tagged.

Resources
The test set contains [nr requirements] privacy requirements from three different privacy statements:

- Instagram
- Lufthansa
- Bitcoin

The tagging takes about an hour and 15 minutes.

Guidelines
Attached is an excel file that contain the column explained in Table 4.2. The task of the participant is to fill the empty cells if applicable. The components that need to be labeled in the text are the following (see also Table 4.2):

- Controller
- Requirement type
- Processing activity
- Personal data
- Data subject
- Purpose
- Restriction
- Data refinement

Table 4.2 shows the guidelines in as a table with the column names and a description.

General remarks in addition to Table 4.2:

- Components are mutually exclusive and exhaustive. This means that one word of group of words can not be labeled more than once.
  - The exception is within the data subject. It might be that the data subject is part of the personal data.
- Make sure no components are tagged within text pieces that do not form the main requirement;
  - The purpose, data refinement, and restrictions/conditions are not part of the main requirement.
  - Example: When you are business travelers, we also collect information relating to your company such as company name and business location.

  - The first and last part of the sentences are not part of the main requirement, so do not tag components, ’controller’, ’model verb’, ’personal data’, and ’data subject’, in the part of the sentence that indicate the ’data refinement’, ’purpose’, and ’restriction’.
  - If you are in doubt try to select the words were the other components are also represented.

- If multiple words or groups of words are applicable choose the group of words that is within the part of the sentences that contains most other components.
- A processing activity can consist of multiple activities (i.e. “collect and share” or “collect, access, and process” is allowed).

C.4 Text Chunking

Table C.2: Common found NLTK tag set abbreviations

POS abbreviation	**meaning**	**example**
PRP	personal pronoun	you
PRP$	possessive pronoun	your
VB	verb, base form	record
VBP	verb, non 3rd person singular present	collect
VBG	verb, gerund or present participle	relating
VBD	verb, past tense	visited
MD	modal verb	may
NN	noun singular	information
NNS	noun plural	registrations
DT	determiner	the
JJ	adjective or numeral, ordinal	different
IN	preposition	for
RB	adverbs	also
CC	coordinating conjunction	and
WRB	wh-adverb	when
TO	to	to
NP	noun phrase, consist of DT, JJ, and NN	a company name

Appendix D

Jupyter Notebooks

D.1 Running the Jupyter Notebooks

README Jupyter Notebook Installation Guide for Windows

One of the deliverable artifacts of the project is code that combines the text mining techniques to process privacy statements and extract components from privacy requirements.

Getting Started

To run the Jupyter notebooks an anaconda environment needs to be installed.

Install Anaconda

Anaconda is a package manager containing a wide range of pre-installed open source packages. The main reason for using Anaconda are the predefined internal and external dependency for various Python packages. All packages used are installed in a virtual environment.

- Install Anaconda with Python 2.7

Create Anaconda Environment

- Save the spec-file.txt somewhere on your computer where you can access it.
- Open Anaconda Prompt
- Create an Anaconda environment based on the text file (spec-file.txt) provided link:

```
conda create --name re_env --file spec-file.txt
```

- Activate the new environment

```
activate re_env
```

- Verify that the environment and its packages are installed correctly

```
conda list
```

Running the Jupyter Notebook

- The Jupyter Notebook with:

```
jupyter notebook
```

- In the Jupyter environment navigate to the directory where the notebooks are stored.
- Run the code in the notebook by clicking on Cell > Run Cells.

D.2 Preparing the Data Sets

Tagging_Preparation

October 7, 2017

1 Preparing Requirement Statements for Manually Tagging

```
In [1]: import pandas as pd
        import nltk
        import re
```

1.0.1 Steps data preparation

1. Read the policies from a text file

- Split the text on newlines
- Split the output strings of step 2 on dots

1.1 Training

Facebook

```
In [2]: # step 1A
        fname = 'policies/social/fb_policy2.txt'
        fb_policy = open(fname, 'r')
        sentences = fb_policy.read()
        sentences = sentences.replace('e.g.','for example').replace(':','').replace
        sentences = [x for x in map(str.strip, sentences.split('.')) if x]
        sentences = [s for s in sentences if len(s) > 30]

        print "number of requirements:", len(sentences)

        # step 1B
        df_fb = pd.DataFrame(sentences, columns=["requirement"])
        df_fb.head()

        # step 1C
        # store sentences/requirements as csv
        # df_fb.to_csv('trainingset/manual_labeling_fb.csv', sep=';')
```

```
number of requirements: 108
```

```
Out[2]:                                          requirement
        0  Depending on which Services you use, we collec...
        1          Things you do and information you provide
        2  We collect the content and other information y...
        3  This can include information in or about the c...
        4  We also collect information about how you use ...
```

KLM

```
In [3]: # step 1A
        fname = 'policies/webshop/klm_policy.txt'
        klm_policy = open(fname, 'r')
        sentences = klm_policy.read()
        sentences = sentences.replace('e.g.','for example').replace(':','').replace
        sentences = [x for x in map(str.strip, sentences.split('.')) if x]
        sentences = [s for s in sentences if len(s) > 30]

        print "number of requirements:", len(sentences)
        sentences

        # step 1B
        df_klm = pd.DataFrame(sentences, columns=["requirement"])
        df_klm.head()

        # step 1C
        # store sentences/requirements as csv
        # df_klm.to_csv('trainingset/manual_labeling_klm.csv')
```

```
number of requirements: 79
```

```
Out[3]:                                          requirement
        0  We may collect and process the following categ...
        1  For example, we may record your name, title, g...
        2  Your contact details may include your address,...
        3  When you create a personal account or register...
        4  For business travellers, we also collect infor...
```

Google+

```
In [4]: # step 1A
        fname = 'policies/social/google_policy.txt'
        google_policy = open(fname, 'r')
        sentences = google_policy.read()
        sentences = sentences.replace('e.g.','for example').replace(':','').replace
        sentences = [x for x in map(str.strip, sentences.split('.')) if x]
        sentences = [s for s in sentences if len(s) > 30]
        print "number of requirements:", len(sentences)
        sentences

        # step 1B
        df_google = pd.DataFrame(sentences, columns=["requirement"])
        df_google.head()

        # step 1C
        # store sentences/requirements as csv
        # df_google.to_csv('trainingset/manual_labeling_google.csv')
```

```
number of requirements: 121
```

```
Out[4]:                                              requirement
        0  We collect information to provide better servi...
        1        We collect information in the following ways
        2  For example, many of our services require you ...
        3  When you do, we'll ask for personal informatio...
        4  If you want to take full advantage of the shar...
```

1.2 Test

Lufthansa

```
In [5]: # step 1A
        fname = 'policies/webshop/lufthansa.txt'
        lufthansa_policy = open(fname, 'r')
        sentences = lufthansa_policy.read()
        sentences = sentences.replace('LH.com','Lufthansa').replace('e.g.','for exa
        sentences = [x for x in map(str.strip, sentences.split('.')) if x]
        sentences = [s for s in sentences if len(s) > 30]

        print "number of requirements:", len(sentences)

        # step 1B
        df_lufthansa = pd.DataFrame(sentences, columns=["requirement"])
        df_lufthansa.head()

        # step 1C
        # store sentences/requirements as csv
        # df_lufthansa.to_csv('testset/manual_labeling_lufthansa.csv', sep=';')
```

```
number of requirements: 47
```

```
Out[5]:                                              requirement
        0  Any personal information collected during visi...
        1  Our data protection policy also complies with ...
```

```
2  Our websites may contain links to other provid...
3           Lufthansa respects your personal privacy
4         Collection and processing of personal data
```

Instagram

```
In [6]: # step 1A
        fname = 'policies/social/instagram_policy.txt'
        insta_policy = open(fname, 'r')
        sentences = insta_policy.read()
        sentences = sentences.replace('e.g.','for example').replace(':','').replace
        sentences = [x for x in map(str.strip, sentences.split('.')) if x]
        sentences = [s for s in sentences if len(s) > 30]

        print "number of requirements:", len(sentences)

        # step 1B
        df_insta = pd.DataFrame(sentences, columns=["requirement"])
        df_insta.head()

        # step 1C
        # store sentences/requirements as csv
        # df_insta.to_csv('trainingset/manual_labeling_insta.csv', sep=';')
```

```
number of requirements: 113
```

```
Out[6]:                                       requirement
        0      We collect the following types of information
        1                Information you provide us directly
        2  Your username, password and e-mail address whe...
        3  Profile information that you provide for your ...
        4  This information allows us to help you or othe...
```

D.3 Modeling the Algorithm

Text_Chunking_Dependencies

October 7, 2017

1 Import packages

```
In [1]: # for including images
        from IPython.display import Image
        import pandas as pd
        # for preprocessing text and text chunking
        import nltk
        from nltk import Tree
        from nltk.tag import UnigramTagger
        # wordnet
        from nltk.corpus import wordnet as wn
        # for calculating performance scores
        import sklearn
        from sklearn.metrics import precision_recall_fscore_support, confusion_matr
        import re
        import numpy as np
        # spacy
        import spacy
        nlp = spacy.load('en')
        # timestamp
        import time
        import datetime
        # plotting
        import matplotlib.pyplot as plt
        import matplotlib
        from matplotlib import cm # colormap
        %matplotlib inline
        import seaborn as sns
```

```
In [2]: print "The following packages and versions are used:"
        print "pandas: ", pd.__version__
        print "NLTK: ", nltk.__version__
        print "Matplotlib: ", matplotlib.__version__
        print "sklearn: ", sklearn.__version__
        print "spacy: ",spacy.__version__
```

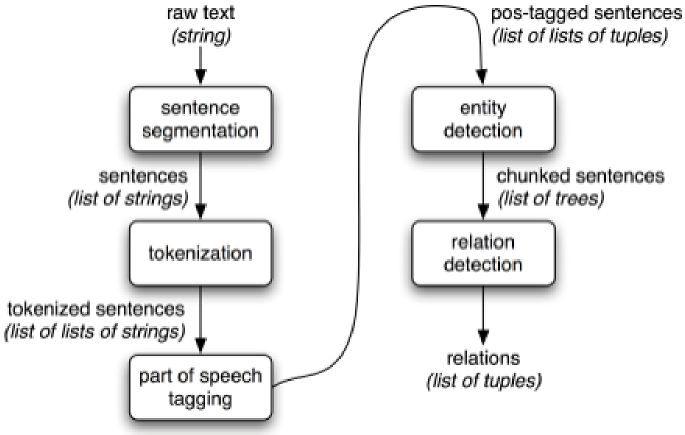
```
The following packages and versions are used:
pandas:  0.20.3
```

```
NLTK:  3.2.4
Matplotlib:  2.0.2
sklearn:  0.18.2
spacy:  1.9.0
```

2 Component Identification with Text Chunking

2.0.1 Steps

1. General (see notebook Manually Tagging Preparation)
2. Split the privacy statement (.txt) in separate requirements on "." (dots)
3. Create a dataframe
4. Store the dataframe locally as .csv file
5. *Manually tagged all the elements for each requirement in excel (based on the .csv)*
6. Element specific actions
7. Load the file with manually tagged elements in a pandas dataframe
8. Select the needed element ('data subject' in this case)
9. POS-tagging (see notebook POS tagging)
10. Create a list of alle requirements
11. Get only the requirements that are considered to be valid requirements
12. A. Fixed chunk
13. Take only the requirements that have a manually tagged data subject
14. Change manually tagged data subject 'your' to 'you', assuming that they are the same person
15. Set the chunk fixed on 'you', this means that we assume that the data subject is always categorized as 'you'
16. B. Text Chunking (see notebook Text Chunking)
17. Apply several preprocessing techniques and POS-tagging to each individual requirement
18. Determine grammar (PRP)
19. Apply text chunking to each individual requirement
20. Merge test and training set
21. Calculate the the precision, recall, and f-score by comparing the manually tagged controllers with the controllers labeled by text chunking



source: NLTK handbook - Extracting Information from Text

3 Preparing Data Set

```
In [3]: # step 2A
        df_fb = pd.DataFrame.from_csv('trainingset/manual_labeling_fb_csv.csv', sep
        df_fb['source'] = 'facebook'
        df_klm = pd.DataFrame.from_csv('trainingset/manual_labeling_klm_csv.csv', s
        df_google = pd.DataFrame.from_csv('trainingset/manual_labeling_google_csv.c
        df_golden = pd.DataFrame.from_csv('testset/golden_test_set_csv.csv', sep=';
        df_golden['valid'] = 1
        df_golden['set'] = 'test'
        df_fb['set'], df_google['set'], df_klm['set'] = 'training', 'training', 'tr
        columns = ['source','set','valid','requirement','controller','requirement_t
        df_klm = df_klm[columns]
        df_fb = df_fb[columns]
        df_google = df_google[columns]

        df_golden = df_golden[columns]
        frames = [df_klm, df_fb, df_google,df_golden]
        df = pd.concat(frames, ignore_index=True)
        df = df.rename(columns={'requirement_type': 'modality'})
        # only use the valid requirements (valid = 1)
        df_valid = df.loc[df['valid'] == 1]
        print "We will be working with", len(df_valid), "valid requirements."
```

```
We will be working with 194 valid requirements.
```

4 Grammar Development

```
In [4]: # step 4A: determine the chunk pattern
        grammar0 =  r"""
```

```
    controller: {<PRP.*>}                   # controller
    NP: {<DT|JJ|NN.*>+}          # Chunk sequences of DT, JJ, NN
  """

grammar1 = r"""
    controller: {<PRP.*>}                   # controller
    process: {<MD>?<RB>?<VB.*><RP>?<CC>?<VB.*>?<RP>?}  # activity: a option
    NP: {<DT|JJ|NN.*>+}          # Chunk sequences of DT, JJ, NN
    data: {<NP>?<CC>?<NP>}
  """

grammar2 = r"""
    controller: {<PRP.*>}                   # controller
    process: {<MD>?<RB>?<VB.*><RP>?<CC>?<VB.*>?<RP>?}  # activity: a option
    activity: {<controller><process>}
    NP: {<DT|JJ|NN.*>+}          # Chunk sequences of DT, JJ, NN
    data: {<NP>?<CC>?<NP>}
    restriction: {<WRB><activity>?<controller>?<data>?}
  """
grammar3 = r"""
    # controller
    controller: {<PRP>}

    # processing activity
    process: {<MD>?<RB>?<VB.*><RP>?<CC>?<VB.*>?<RP>?}  # activity: a option
    activity: {<controller><process>}

    # data
    data_subject: {<PRP\$>}       # data object/owner
    NP: {<DT|JJ|NN.*>+}          # Chunk sequences of DT, JJ, NN
    data: {<NP>?<CC>?<NP>}
    personal_data: {<data_subject>?<data><IN>?<data>?}

    # restriction
    restriction: {<WRB><activity>?<controller>?<personal_data>?}
  """
grammar4 = r"""
    # controller
    controller: {<PRP>}

    # processing activity
    modal_verb: {<MD><RB>?<RB>?}
    process: {<VB.*><RP>?<CC>?<VB.*>?<RP>?}  # activity: a optional MD and
    activity: {<controller><modal_verb>?<process><RB>?}

    # data
    data_subject: {<PRP\$>}      # data object/owner
    NP: {<DT|JJ|NN.*>+}          # Chunk sequences of DT, JJ, NN
    data: {<NP>?<CC>?<NP>}
    personal_data: {<data_subject>?<data><IN>?<data>?}

    # restriction
    restriction: {<WRB><activity>?<controller>?<personal_data>?}
  """

grammar5 = r"""
    # controller
    controller: {<PRP>}

    # processing activity
    modal_verb: {<MD><RB>?<RB>?}
    process: {<VB.*>+<RP>?<CC>?<VB.*>?<RP>?}  # activity: a optional MD and
    activity: {<controller><modal_verb>?<process><RB>?}

    # data
    data_subject: {<PRP\$>}      # data object/owner
    NP: {<DT|JJ|NN.*>+}          # Chunk sequences of DT, JJ, NN
    data: {<NP>?<CC>?<NP>}
    personal_data: {<data_subject>?<data><IN>?<data>?}

    # restriction
    restriction: {<WRB><activity>?<controller>?<personal_data>?}
  """

grammar5spacy = r"""
    # controller
    controller: {<PRON>}

    # processing activity
    modal_verb: {<VERB><ADV>?<ADV>?}
    process: {<VERB.*>+<ADP>?<CCONJ>?<VERB.*>?<ADP>?}  # activity: a optior
    activity: {<controller><modal_verb>?<process>}

    # data
    data_subject: {<PRON>}      # data object/owner
    NP: {<DET|ADJ|NOUN>+}          # Chunk sequences of DT, JJ, NN
    data: {<NP>?<CCONJ>?<NP>}
    personal_data: {<data_subject>?<data><ADP>?<data>?}
  """

grammar6 = r"""
    # controller
    controller: {<PRP>}

    # processing activity
    modal_verb: {<MD><RB>?<RB>?}
    process: {<VB.*><RP>?<CC|,|:>?<VB.*>?<RP>?<,|:>?<CC>?<VB.*>?<RP>?}  # a
    activity: {<controller><modal_verb>?<process><RB>?}

    # data
    data_subject: {<PRP\$>}      # data object/owner
    NP: {<DT|JJ|NN.*>+}          # Chunk sequences of DT, JJ, NN
    data: {<NP>?<CC>?<NP>}
    personal_data: {<data_subject>?<data><IN>?<data>?}
  """

grammar7 = r"""
    # controller
    controller: {<PRP>}

    # processing activity
    modal_verb: {<MD><RB>?<RB>?}
    process: {<VB.*><RP>?<CC|,|:>?<VB.*>?<RP>?<,|:>?<CC>?<VB.*>?<RP>?}  # a
    activity: {<controller><modal_verb>?<process><RB>?}

    # data
    data_subject: {<PRP\$>}      # data object/owner
    NP: {<DT|JJ|NN.*|,>+}          # Chunk sequences of DT, JJ, NN
    personal_data: {<NP>?<CC>?<NP><IN>?<NP>?<CC>?<NP>?}
    data: {<data_subject><personal_data>}
  """

grammar8 = r"""
    # controller
    controller: {<PRP>}

    # processing activity
    modal_verb: {<MD><RB>?<RB>?}
        # activity: a optional MD and RB, then a VB, followed by an optiona
    process: {<VB.*><RP>?<CC|,|:>?<VB.*>?<RP>?<,|:>?<CC>?<VB.*>?<RP>?}
    activity: {<controller><modal_verb>?<process><RB>?}

    # data
        # data object/owner
    data_subject: {<PRP\$>}
        # Chunk sequences of DT, JJ, NN
    NP: {<DT|JJ|NN.*|,>+}
    personal_data: {<NP>?<CC>?<NP><IN>?<NP>?<CC>?<NP>?}
    data: {<data_subject><personal_data>}

    # purpose
    purpose: {<TO><process><controller|modal_verb|process|activity|NP|data_
    """
```

```
grammar9 = r"""
    # controller
    controller: {<PRP>}

    # processing activity
    modal_verb: {<MD><RB>?<RB>?}
    process: {<VB.*><RP>?<CC|,|:>?<VB.*>?<RP>?<,|:>?<CC>?<VB.*>?<RP>?}
        # activity: a optional MD and RB, then a VB, followed by an optiona
    activity: {<controller><modal_verb>?<process><RB>?}

    # data
        # data object/owner
    data_subject: {<PRP\$>}
        # Chunk sequences of DT, JJ, NN
    NP: {<DT|JJ|NN.*|,>+}

    personal_data: {<NP>?<CC>?<NP><IN>?<NP>?<CC>?<NP>?<process>?}
    data: {<data_subject|personal_data>+}

    # purpose
    purpose: {<TO><process><controller|modal_verb|process|activity|NP|data_
    """
grammar10 = r"""
    # controller
    controller: {<PRP>}

    # processing activity
    modal_verb: {<MD><RB>?<RB>?}
        # activity: a optional MD and RB, then a VB, followed by an optiona
    process: {<VB.*><RP>?<CC|,|:>?<VB.*>?<RP>?<,|:>?<CC>?<VB.*>?<RP>?}
    activity: {<controller><modal_verb>?<process><RB>?}

    # data
        # data object/owner
    data_subject: {<PRP\$>}
        # Chunk sequences of DT, JJ, NN
    NP: {<DT|JJ|NN.*|,>+}
    personal_data: {<NP>?<CC>?<NP><IN>?<NP>?<CC>?<NP>?<process>?}
    data_t: {<data_subject|personal_data>+}
    data: {<activity><data_t>}

    # purpose
    purpose: {<TO><process><controller|modal_verb|process|activity|NP|data_
    """

grammarbhatia = r"""
    HKO: {<JJ|RB|NN.*>?<POS.*>*<NN.*>?}
    """
```

```
        grammar = grammar8

In [5]: df_text_chunking = df_valid.copy()
        df_text_chunking.head(2)

Out[5]:   source       set  valid                                        requiremen
        0    klm  training      1  We may collect and process the following categ..
        3    klm  training      1  When you create a personal account or register..

          controller    modality keyword  processing_activity  \
        0         we  permission     may  collect and process
        3         we  permission     may               record

                                   personal_data data_subject  \
        0     categories of personal information          NaN
        3  sign in details and other information         your

                                                 restriction purpose data_refinemen
        0                                                NaN     NaN               Na
        3  When you create a personal account or register...     NaN               Na

In [6]: print len(df_text_chunking['data_subject'])
        df_top_words = df_text_chunking.groupby(['data_subject'])['requirement'].nu
        df_top_words = df_top_words.reset_index()
        df_top_words['percentage'] = df_top_words['requirement']/len(df_text_chunki
        df_top_words

194


Out[6]:            data_subject  requirement  percentage
        0                   our            1    0.515464
        1  personal information            1    0.515464
        2                    us            1    0.515464
        3                   who            1    0.515464
        4                   you           49   25.257732
        5                  your           56   28.865979
```

5 Implementing the Text Mining Techniques

5.1 Preparation Functions

```
In [8]: # POS-tagger
        def tokenize(doc):
            tokens = nltk.word_tokenize(sentence)
            return tokens


        def simplify(text):
            simplified = []
            for word in text:
                if word.isalpha():
                    simplified.append(word.lower())
            return simplified


        def pos_tagging(doc):
            tokens = nltk.word_tokenize(doc)
            tokens_simplified = simplify(tokens)
            tags = nltk.pos_tag(tokens_simplified)
            return tags
```

5.1.1 spaCy: Remove Dependency Tree

spaCy dependency parser provides token properties to navigate the generated dependency parse tree. Using the dep attribute gives the syntactic dependency relationship between the head token and its child token. The syntactic dependency scheme is used from the ClearNLP. The generated parse tree follows all the properties of a tree and each child token has only one head token although a head token can have multiple children. We can obtain the head token with the token.head property and its children by the token.children property. A subtree of a token can also be extracted using the token.subtree property. Similarly, ancestors for a token can be obtained with token.ancestors. To obtain the rightmost and leftmost token of a token's syntactic descendants the token.right_edge and token.left_edge can be used. It is also worth mentioning that to extract the neighboring token we can use token.nbor. spaCy doesn't provide an inbuilt tree representation although you can use the NLTK's tree representation.

https://shirishkadam.com/2016/12/23/dependency-parsing-in-nlp/

```
In [9]: def get_lookup(requirement):
            lookup = {}
            for i, word in enumerate(requirement):
                if word.head is word:
                    head_idx = 0
                else:
                    head_idx = requirement[i].head.i+1
                lookup[i+1] = (requirement[i],requirement[i].dep_,requirement[i].ta
            return lookup

        def get_root(lk):
            for i in lk:
                if lk[i][1] == 'ROOT':
                    return i

        def get_dependency_parent(lk, modifier):
            # get the ROOT of the dependency subtree (note: not the ROOT of the rec
            root_i = get_root(lk)
            dependency_parents = []
            for i in lk:
```

```
        if lk[i][1] == modifier: # and lk[i][3] == root_i
            dependency_parents.append([i])
    return dependency_parents

def get_dependency_children(dependency_parents, lk):
    # get the children for all the depedency parents
    for dep_i in dependency_parents:
        for lk_i in lk:
            if dep_i == lk[lk_i][3]:
                dependency_parents.append(lk_i)
    return dependency_parents

def remove_dependency(requirement, modifier):
    # create a lookup object that maps all the words in the requirement wit
    lk = get_lookup(requirement)
    dependency_sentences = []
    # get all the parents that have the dependency type that we are looking
    parent = get_dependency_parent(lk, modifier)
    # find the children of the dependency parents
    for dep_parent in parent:
        dependencies = get_dependency_children(dep_parent, lk)
        # clean the original requirement from the dependency
        dependency = ' '.join([str(lk[i][0]) for i in lk if i in dependenci
        dependency_sentences.append(str(dependency))
        # replace the original sentence with the found dependency
        requirement = str(requirement).replace(dependency,'')
    return dependency_sentences, str(requirement).rstrip().lstrip()
```

5.1.2 Data Refinement with Regular Expressions

```
In [10]: def get_data_refinement(requirement):
             refinement = ""
             if "(" not in requirement:
                 refinement = None
             else:
                 ref = re.findall(r'\((.*?)\)',requirement)
                 if len(ref) > 1:
                     for r in ref:
                         refinement += r + " "
             return refinement
```

5.1.3 Modalities according to Breaux

```
In [11]: modalities = pd.read_csv('samples/modalities.csv', delimiter=";")
         modalities
```

```
Out[11]:                  modal verbs         modality
         0   does not have a right to      prohibition
```

```
1                  does not      prohibition
2                    do not      prohibition
3             has a right to      permission
4          is not required to  anti-obligation
5                        may      permission
6                   may deny      permission
7                    may not      obligation
8             may not require      obligation
9                 may require      permission
10                      must      obligation
11                 must deny      obligation
12               must permit      obligation
13              must request      obligation
14       retains the right to      permission
15                     could      permission
16                       can      permission
17                  will not      prohibition
```

```
In [12]: def get_modality(modal_verb):
             modalities = pd.read_csv('samples/modalities.csv', delimiter=";").to_c
             for i in modalities['modal verbs']:
                 if modalities['modal verbs'][i] == modal_verb:
                     return modalities['modality'][i]
```

5.2 Text Chunking Functions

```
In [13]: def text_chunker(grammar, tags):
             labelset = []
             cp = nltk.RegexpParser(grammar)
             tree = cp.parse(tags)
             return tree


         # data object, personal data
         def get_data_values(df, tree, component):
             data_subject = ""
             personal_data = ""
             # get the subtree for each component
             for stree in tree.subtrees(filter=lambda x: x.label() == component):
                 if component == "data_subject":
                     if stree[0][0] == "your":
                         data_subject = "you"
                     else:
                         data_subject = stree[0][0].rstrip().lstrip()
         #           else:
         #               print "data subject does not exists: ", stree[0][0]
         #               data_subject = "you" # experiment!!!
                 if component == "personal_data":
```

```
            data_tmp = ""
            for el in stree.leaves():
                data_tmp += el[0] + ' '

            if personal_data == "":
                personal_data += data_tmp

    # fill database
    if component == 'data_subject':
#         if not data_subject: # not a better preformance
#             data_subject = 'you'
        if data_subject != 'you': # better performance
            data_subject = np.nan
        df_text_chunking.loc[index,'data_subject_tc'] = data_subject
    if component == 'personal_data':
#         print personal_data
        df_text_chunking.loc[index,'personal_data_tc'] = personal_data.rst

# controller, modality, processing activity
def get_activity_values(df, tree, component):
    controller = ""
    processes = ""
    modal_verb = ""
    for stree in tree.subtrees(filter=lambda x: x.label() == component):
        if component == 'controller':
            if (controller == "" or controller == "you"):
                controller = stree[0][0]
        if component == 'process':
            if len(stree[0][0]) > 2:
                if (processes == "" and len(stree) > 2):
                    for el in stree:
                        processes += (el[0] + ' ')
                if processes == "" or len(processes) < 3:
                    processes = stree[0][0]
        if component == 'modal_verb':
            for el in stree:
                if (el[1] == 'MD' or el[0] == "not"):
                    modal_verb += el[0] + ' '
    # fill database
    if component == 'controller':
        df_text_chunking.loc[index,'controller_tc'] = controller
    if component == 'process':
        df_text_chunking.loc[index,'processes_tc'] = processes.rstrip()
    if component == 'modal_verb':
        df_text_chunking.loc[index,'modal_verb_tc'] = modal_verb.rstrip()
        # find right modality
        # remove the whitespace at the end of the line
        mod = get_modality(modal_verb.rstrip())
        if mod is None:
            mod = "permission"
#         else:
        df_text_chunking.loc[index, 'modality_tc'] = mod


# data object, personal data
def get_purpose_values(df, tree, component):
    purpose = ""
    # get the subtree for each component
    for stree in tree.subtrees(filter=lambda x: x.label() == component):
        if component == "purpose":
            purpose_tmp = ""
            for el in stree.leaves():
                purpose_tmp += el[0] + ' '
            if purpose == "":
                purpose += purpose_tmp
    # fill database
    if component == 'purpose':
#         print purpose
        df_text_chunking.loc[index,'purpose_tc'] = purpose.rstrip().lstrip


def remove_subtree(tree,requirement, component):
    sub = []
    for subtree in tree.subtrees(filter=lambda t: t.label() == component):
        sub.append(" ".join([a for (a,b) in subtree.leaves()]))
    subtree_str = sub
    if subtree_str:
        cleaned_requirement = requirement.lower().replace(sub[0], '')
        cleaned_tags = pos_tagging(cleaned_requirement)
        pruned_tree = text_chunker(grammar, cleaned_tags)
        return pruned_tree
    else:
        return tree

def text_chunking():
    if (row["requirement"]):
        cleaned_requirement = row["requirement"]
        # RESTRICTION: SPLITTING DEPENDENCY (ADVERBIAL CLAUSE MODIFIER)
        requirement = nlp(unicode(cleaned_requirement, 'unicode-escape'))
        restriction, cleaned_requirement = remove_dependency(requirement,
        if len(restriction) > 0:
            df_text_chunking.loc[index,'restriction_spacy'] = restriction
        else:
            df_text_chunking.loc[index,'restriction_spacy'] = np.nan

        # DATA REFINEMENT: SPLITTING DEPENDENCY(PREPOSITIONAL MODIFIER)
        requirement = nlp(unicode(cleaned_requirement, 'unicode-escape'))
        refinement, cleaned_requirement = remove_dependency(requirement, '
#    print refinement
        if len(refinement) > 0:
            df_text_chunking.loc[index,'data_refinement_spacy'] = refineme
        else:
            df_text_chunking.loc[index,'data_refinement_spacy'] = np.nan
        # based on Text Chunking
        data_refinement = get_data_refinement(cleaned_requirement)
        if data_refinement is not None:
            df_text_chunking.loc[index,'data_refinement_re'] = data_refine

#    print cleaned_requirement
        # set new cleaned requirement in the dataframe
        df_text_chunking.loc[index,'cleaned_requirement'] = cleaned_requi

#    print cleaned_requirement
        # POS-tagging
        tags = pos_tagging(cleaned_requirement)
        if (tags):
            tree = text_chunker(grammar, tags)
            # loop through the different components
            activity_components = ['controller','modal_verb','process']
            for component in activity_components:
                get_activity_values(df, tree, component)
            data_components = ['data_subject','personal_data']
            for component in data_components:
                get_data_values(df, tree, component)

            purpose_components = ['purpose']
            for component in purpose_components:
                get_purpose_values(df, tree, component)


# ,('restriction','restriction_spacy')
comp_names = [('controller','controller_tc'),
              ('keyword','modal_verb_tc'),
              ('modality','modality_tc'),
              ('processing_activity','processes_tc'),
              ('data_subject','data_subject_tc'),
              ('personal_data','personal_data_tc'),
              ('data_refinement','data_refinement_re'),
              ('restriction','restriction_spacy'),
              ('purpose','purpose_tc')]
for index, row in df_text_chunking.iterrows():
    # loop through all the requirements and apply POS-tagging
    text_chunking()
```

```
        # (try 3) change 'us' with 'we'
        # df_text_chunking["data_subject"] = df_text_chunking["data_subject"].rep
        # (try 3) change 'our' with 'we'
        # df_text_chunking["data_subject"] = df_text_chunking["data_subject"].rep
        # (try 3) change 'your' with 'you'
        df_text_chunking["data_subject"] = df_text_chunking["data_subject"].replac
        # clean personal data from punctuation
        # df_text_chunking['personal_data'] = df_text_chunking['personal_data'].st
        # df_text_chunking['personal_data'] = df_text_chunking['personal_data'].st

        df_text_chunking = df_text_chunking.replace("", np.nan)
```

6 Calculating the Results

6.1 Requirement cleaning

Compare the difference between the cleaned requirement after the dependency parsing and the original requirement

```
In [43]: df_original_clean = df_text_chunking[['requirement','cleaned_requirement']
         df_original_clean.head()

         df_original_clean['diff_len'] = df_original_clean['requirement'].str.len()
         df_original_clean['len'] = df_original_clean['requirement'].str.len()
         diff_sum = df_original_clean['diff_len'].sum()
         len_sum = df_original_clean['len'].sum()
         print diff_sum, len_sum
         percentage_diff = float(diff_sum)/float(len_sum)*100
         print "On average, the dependency parser reduces the lenght of the origina
         print round(percentage_diff,2), "percent."
```

```
C:\Users\LBrakenhoff\AppData\Local\Continuum\Anaconda2\envs\thesisEnv\lib\site-pack
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/i
  after removing the cwd from sys.path.
C:\Users\LBrakenhoff\AppData\Local\Continuum\Anaconda2\envs\thesisEnv\lib\site-pack
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/i
  """


11734 29353
On average, the dependency parser reduces the lenght of the original requirement wi
39.98 percent.
```

6.2 Perfomance Measurements

6.2.1 Precision, Recall, F-score

```
In [16]: def get_performance_report(df, column_true, column_pred):
             y_true = df[column_true].astype(str)
             y_pred = df[column_pred].astype(str)
             report = classification_report(y_true, y_pred)
             return report


         def calculate_performance(df, column_true, column_pred):
             y_true = df[column_true].astype(str)
             y_pred = df[column_pred].astype(str)
             preformance = precision_recall_fscore_support(y_true, y_pred,
                                                           average='weighted')
             return preformance


         def get_confusion_matrix(df, column_true, column_pred):
             y_true = df[column_true].astype(str)
             y_pred = df[column_pred].astype(str)
             matrix = confusion_matrix(y_true, y_pred)
             return matrix


         def get_matthews_CC(df, column_true, column_pred):
             y_true = df[column_true].astype(str)
             y_pred = df[column_pred].astype(str)
             cc = matthews_corrcoef(y_true, y_pred)
             return cc


         def get_scores_table(comp_names, df):
             scores = []
             indexes = []
             for comp_name in comp_names:
                 score = calculate_performance(df,comp_name[0],comp_name[1])
                 indexes.append(comp_name[0])
                 scores.append(score[:-1])
             df_scores = pd.DataFrame(scores,
                                      columns=['precision','recall','f-score'],
                                      index=indexes)
             return df_scores
```

```
In [ ]:
```

```
In [17]: sources = list(df_text_chunking.source.unique())
         result_tables = []
         df_results_total = get_scores_table(comp_names, df_text_chunking)
         for source in sources:
             if (source != 'total'):
                 df = df_text_chunking.loc[df_text_chunking['source'] == source]
                 df_results = get_scores_table(comp_names, df)
             result_tables.append(df_results)
         df_results = pd.concat(result_tables,
                                keys=sources,
                                names=['source', 'component'])
         # save result as csv
         # df_results.to_csv('samples/result_with_punct_20082017.csv')
```

```
C:\Users\LBrakenhoff\AppData\Local\Continuum\Anaconda2\envs\thesisEnv\lib\site-pack
  'precision', 'predicted', average, warn_for)
C:\Users\LBrakenhoff\AppData\Local\Continuum\Anaconda2\envs\thesisEnv\lib\site-pack
  'recall', 'true', average, warn_for)
```

```
In [18]: sets = list(df_text_chunking.set.unique())
         # for gather the results for each set (test and training)
         for s in sets:
             df = df_text_chunking.loc[df_text_chunking['set'] == s]
             df_results2 = get_scores_table(comp_names, df)
             result_tables.append(df_results2)
         df_results_sources = pd.concat(result_tables, keys=sets, names=['set','con
         df_results_sources
```

```
Out[18]:                               precision    recall   f-score
         set      component
         training controller            0.918239  0.924528  0.918762
                  keyword               0.881551  0.905660  0.890658
                  modality              0.981509  0.981132  0.979880
                  processing_activity   0.853774  0.830189  0.839473
                  data_subject          0.807383  0.792453  0.783827
                  personal_data         0.342767  0.301887  0.314825
                  data_refinement       0.609942  0.773585  0.682086
                  restriction           0.618711  0.603774  0.611090
                  purpose               0.419932  0.528302  0.467925
         test     controller            0.868530  0.869565  0.862719
                  keyword               0.937888  0.956522  0.946488
                  modality              0.936715  0.956522  0.942523
                  processing_activity   0.671739  0.673913  0.665985
                  data_subject          0.816496  0.500000  0.502541
                  personal_data         0.184783  0.239130  0.206377
                  data_refinement       0.682420  0.826087  0.747412
                  restriction           0.560641  0.543478  0.551839
                  purpose               0.480549  0.608696  0.537084
```

```
In [21]: df_component = df_results.mean(level='component')
         save = df_component.sort_values(by=['precision'], ascending=False) # .to_
         save.mean()
         # save.to_csv('samples/with_depdendency_parsing.csv')

Out[21]: precision    0.657998
         recall       0.664808
         f-score      0.649514
         dtype: float64

In [22]: df_mean_save = pd.DataFrame.from_csv('samples/without_depdendency_parsing.
         df_mean_save

Out[22]:                      precision    recall   f-score
         component
         modality              0.897392  0.934959  0.912151
         keyword               0.804798  0.799978  0.794134
         data_subject          0.732863  0.695779  0.694920
         controller            0.703984  0.743888  0.712589
         data_refinement       0.685328  0.804321  0.737309
         restriction           0.648130  0.554962  0.593087
         purpose               0.612791  0.543886  0.573672
         processing_activity   0.558018  0.492247  0.508267
         personal_data         0.202417  0.135291  0.150620

In [23]: df_source = df_results.mean(level='source')
         df_source

Out[23]:            precision    recall   f-score
         source
         klm         0.714868  0.737945  0.720947
         facebook    0.682196  0.685990  0.662552
         google      0.680419  0.688889  0.671536
         lufthansa   0.609203  0.626984  0.608906
         instagram   0.603304  0.584229  0.583631
```

Personal Data

```
In [24]: # get_performance_report(df_text_chunking,column_true_pred[0], column_true
         df_personal_data = df_text_chunking[['personal_data','personal_data_tc']]
         df_personal_data = df_personal_data[['personal_data', 'personal_data_tc']]

         df_pd_false = df_personal_data.loc[df_personal_data['not_equal'] == True]
         print "Number of wrongly identified personal data objects:", len(df_pd_fal
         print "That are only", len(df_personal_data)-len(df_pd_false),"correclty t
         # df_personal_data.head(26)
         # df_pd_false[['personal_data','personal_data_tc']]

Number of wrongly identified personal data objects: 143 / 194 .
That are only 51 correclty tagged personal data objects.
```

```
In [25]: def check_in(true, pred):
         #     if isinstance(true, str) and isinstance(pred, str):
             try:
                 if pred in true:
                     return True
                 else:
                     return False
             except:
                 return False
```

6.3 Correlations

6.3.1 Training vs. Test

```
In [28]: df_sd = df_results_sources.T.std()
         df_sd = df_sd.reset_index()
         df_sd.columns = ['set', 'component','std']
         df_mean = df_results_sources.T.mean()
         df_mean = df_mean.reset_index()
         df_mean.columns = ['set','component','mean']
         df_std_mean = pd.merge(df_mean, df_sd,  how='left',
                                  left_on=['set','component'],
                                 right_on = ['set','component'])

         df_std_mean.set_index(['set', 'component'], inplace=False) #.to_latex()
```

```
Out[28]:                                    mean       std
         set      component
         training controller          0.920510  0.003490
                  keyword             0.892623  0.012174
                  modality            0.980840  0.000853
                  processing_activity 0.841145  0.011881
                  data_subject        0.794554  0.011918
                  personal_data       0.319826  0.020894
                  data_refinement     0.688537  0.082012
                  restriction         0.611191  0.007469
                  purpose             0.472053  0.054303
         test     controller          0.866938  0.003690
                  keyword             0.946966  0.009326
                  modality            0.945253  0.010182
                  processing_activity 0.670546  0.004097
                  data_subject        0.606346  0.182000
                  personal_data       0.210097  0.027364
                  data_refinement     0.751973  0.071942
                  restriction         0.551986  0.008582
                  purpose             0.542110  0.064221
```

```
In [41]: # calculate the correlation between the test and training set.
         print "The Pearson Correlation Coefficient for the relationship between"
         print "the performance of the text mining techniques on the training and i
         df_results_sources.unstack(level=1).T.corr()

The Pearson Correlation Coefficient for the relationship between
the performance of the text mining techniques on the training and is:


Out[41]: set        training      test
         set
         training  1.000000  0.891902
         test      0.891902  1.000000

In [30]: df_results_sources.unstack(level=1).T.std()

Out[30]: set
         training    0.214729
         test        0.233462
         dtype: float64
```

6.3.2 Readability vs. Performance Measurements

```
In [42]: df_info = pd.DataFrame.from_csv('samples/text_info.csv')
         df_merge = pd.merge(df_results.mean(level=['source']), df_info,
                             left_index=True, right_on='source', how='left', sort=F
         df_merge['ASL'] = df_merge['Words']/df_merge['Sentences']
         df_merge['ASW'] = df_merge['Syllables']/(df_merge['Words']-df_merge['num_r
         df_merge = df_merge.set_index(df_merge['source'])
         df_merge

Out[42]:            precision    recall   f-score       name     source  \
         source
         klm         0.714868  0.737945  0.720947        KLM        klm
         facebook    0.682196  0.685990  0.662552   Facebook   facebook
         google      0.680419  0.688889  0.671536     Google     google
         lufthansa   0.609203  0.626984  0.608906  Lufthansa  lufthansa
         instagram   0.603304  0.584229  0.583631  Instagram  instagram

                                                path   purpose  Sentences   Wor
         source
         klm             policies/webshop/klm_policy.txt  training       53.0  1117
         facebook         policies/social/fb_policy2.txt  training       46.0  1250
         google       policies/social/google_policy.txt  training       50.0  1200
         lufthansa        policies/webshop/lufthansa.txt      test       14.0   287
         instagram  policies/social/instagram_policy.txt      test       31.0   975

                    Syllables  UniqueWords      FRE    ARI  FKGL    DCRI       SMO
         source
         klm           1722.0        356.0  102.950   7.91  4.40  1.0416   3.129100
         facebook      1889.0        396.0   97.875  15.62  6.74  1.3392   3.129100
```

```
google        1863.0        438.0  100.920  14.12  5.57  1.1904   3.129100
lufthansa      452.0        169.0  103.965  12.12  4.01  0.9920  14.167174
instagram     1377.0        370.0   94.830  17.62  8.30  1.5376   3.129100

           num_not_found_words        ASL       ASW
source
klm                       81.0  21.075472  1.662162
facebook                 103.0  27.173913  1.646905
google                   133.0  24.000000  1.746017
lufthansa                 31.0  20.500000  1.765625
instagram                114.0  31.451613  1.599303
```

```
In [32]: perf_read = df_merge[['precision','recall','f-score','FRE','ARI','FKGL','
         perf_read
```

```
Out[32]:            precision    recall   f-score      FRE    ARI  FKGL    DCRI
         source
         klm         0.714868  0.737945  0.720947  102.950   7.91  4.40  1.0416
         facebook    0.682196  0.685990  0.662552   97.875  15.62  6.74  1.3392
         google      0.680419  0.688889  0.671536  100.920  14.12  5.57  1.1904
         lufthansa   0.609203  0.626984  0.608906  103.965  12.12  4.01  0.9920
         instagram   0.603304  0.584229  0.583631   94.830  17.62  8.30  1.5376
```

```
In [33]: df_perf_read = perf_read.corr()
         df_perf_read
```

```
Out[33]:            precision    recall   f-score       FRE       ARI      FKGL  \
         precision   1.000000  0.973648  0.979827  0.305842 -0.562425 -0.321997
         recall      0.973648  1.000000  0.993378  0.511398 -0.708963 -0.527663
         f-score     0.979827  0.993378  1.000000  0.470160 -0.713358 -0.485259
         FRE         0.305842  0.511398  0.470160  1.000000 -0.846752 -0.998403
         ARI        -0.562425 -0.708963 -0.713358 -0.846752  1.000000  0.861228
         FKGL       -0.321997 -0.527663 -0.485259 -0.998403  0.861228  1.000000
         DCRI       -0.321997 -0.527663 -0.485259 -0.998403  0.861228  1.000000

                        DCRI
         precision -0.321997
         recall    -0.527663
         f-score   -0.485259
         FRE       -0.998403
         ARI        0.861228
         FKGL       1.000000
         DCRI       1.000000
```

7 Visualizations

```
In [34]: def visualize_results(df, label, save):
             ttl = label +' score per component for each source'
```

```
    df.cumsum();
    fig = plt.figure();
    # colors: klm, fb, google, lufthansa, instagram
    color_list = ['#00a1de', '#3b5998', '#ff002b', '#ffc233', '#d10869']
    ax = df.unstack(level=0).plot(kind="bar",
                                  width=0.7,
                                  color=color_list,
                                  title=ttl,
                                  figsize=(10,5),
                                  fontsize=12,rot=45)
    ax.set_axis_bgcolor('white')

    for p in ax.patches:
        ax.annotate(str(round(p.get_height(),2)), (p.get_x() * 1.005, p.ge

    ax.set_ylim([0,1.1])
    lgd = ax.legend(bbox_to_anchor=(1, 1), loc='upper left', ncol=1)
    plt.ylabel(label + ' score',fontsize = 14)
    plt.xlabel('',fontsize = 14)
    plt.grid(False)
    plt.show()
    if save == 'save':
        ts = time.time()
        st = datetime.datetime.fromtimestamp(ts).strftime('%Y-%m-%d-%H-%M-
        fig = ax.get_figure()
        fig.tight_layout()
        # save figure
#         fig.savefig('charts/result_'+label+'.png',
#                     bbox_extra_artists=(lgd,),
#                     bbox_inches='tight') # +'_'+st+


def create_correlation_heatmap(df,size):
    # calculate the correlation matrix
    corr = df.corr()
    fig, ax = plt.subplots(figsize=(size))         # Sample figsize in inc
    sns.set(font_scale=1.4)
    # plot the heatmap
    sns_heatmap = sns.heatmap(corr,
            xticklabels=corr.columns,
            yticklabels=corr.columns,
            cmap="coolwarm",
            annot=True,
            vmin=-1, vmax=1)
    sns_heatmap.set_yticklabels(sns_heatmap.get_yticklabels(),
                                rotation = 0,
                                fontsize = 14)
    sns_heatmap.set_xticklabels(sns_heatmap.get_xticklabels(),
                                      rotation = 45,
                                      fontsize = 14)
            sns_heatmap.hlines([3, 6], *ax.get_xlim(), color='w')
            sns_heatmap.vlines([3, 6], *ax.get_ylim(), color='w')
```

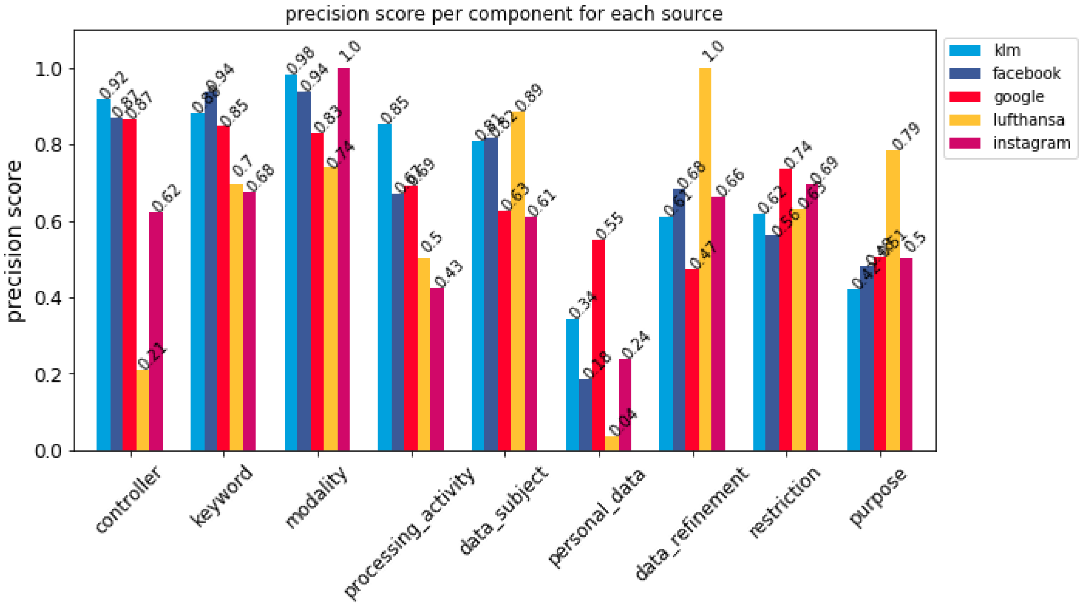
7.1 Performance Measurements

7.1.1 Precision, Recall, and F-score

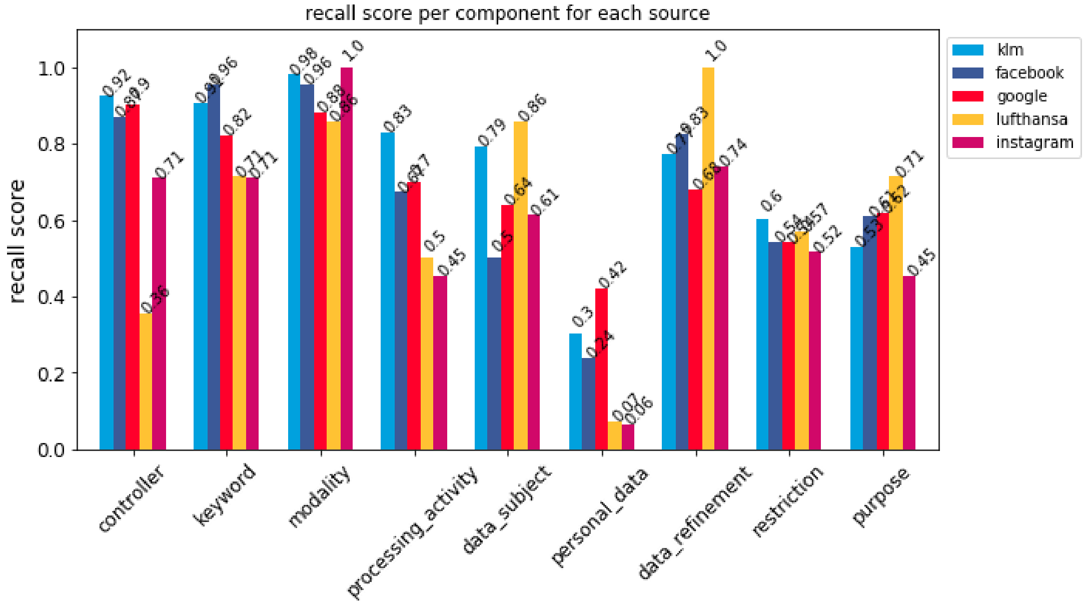
```
In [35]: # Call visualizations
         visualize_results(df_results['precision'], 'precision', 'save')
         visualize_results(df_results['recall'], 'recall', 'save')
         visualize_results(df_results['f-score'], 'f-score', 'save')
```

```
C:\Users\LBrakenhoff\AppData\Local\Continuum\Anaconda2\envs\thesisEnv\lib\site-pack
```

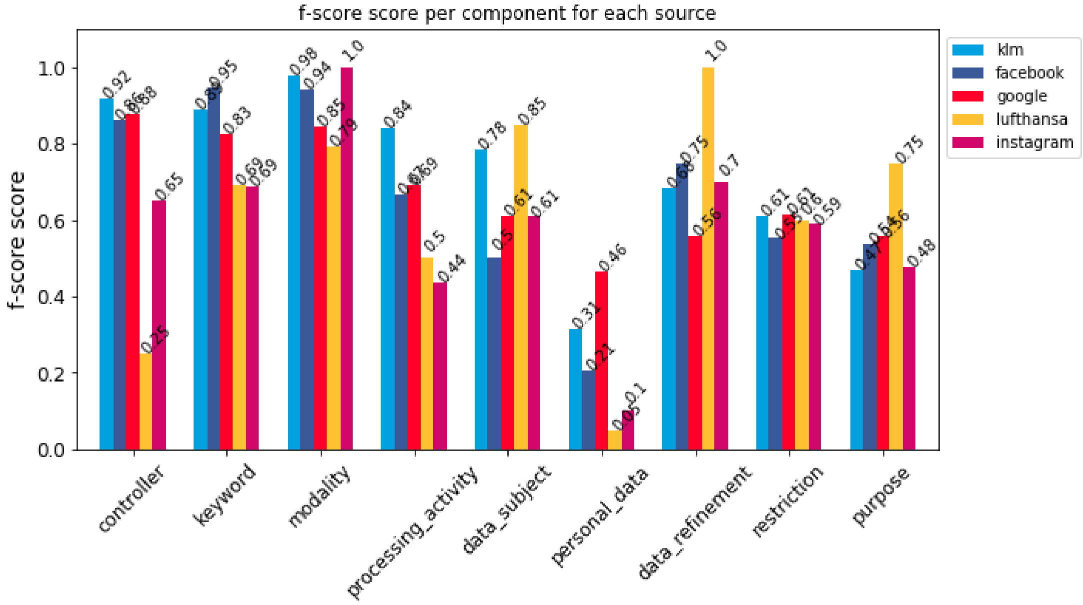
```
<matplotlib.figure.Figure at 0x51e0e5f8>
```



```
<matplotlib.figure.Figure at 0x4ef98c88>
```



```
<matplotlib.figure.Figure at 0x5206d9e8>
```



7.1.2 Correlations

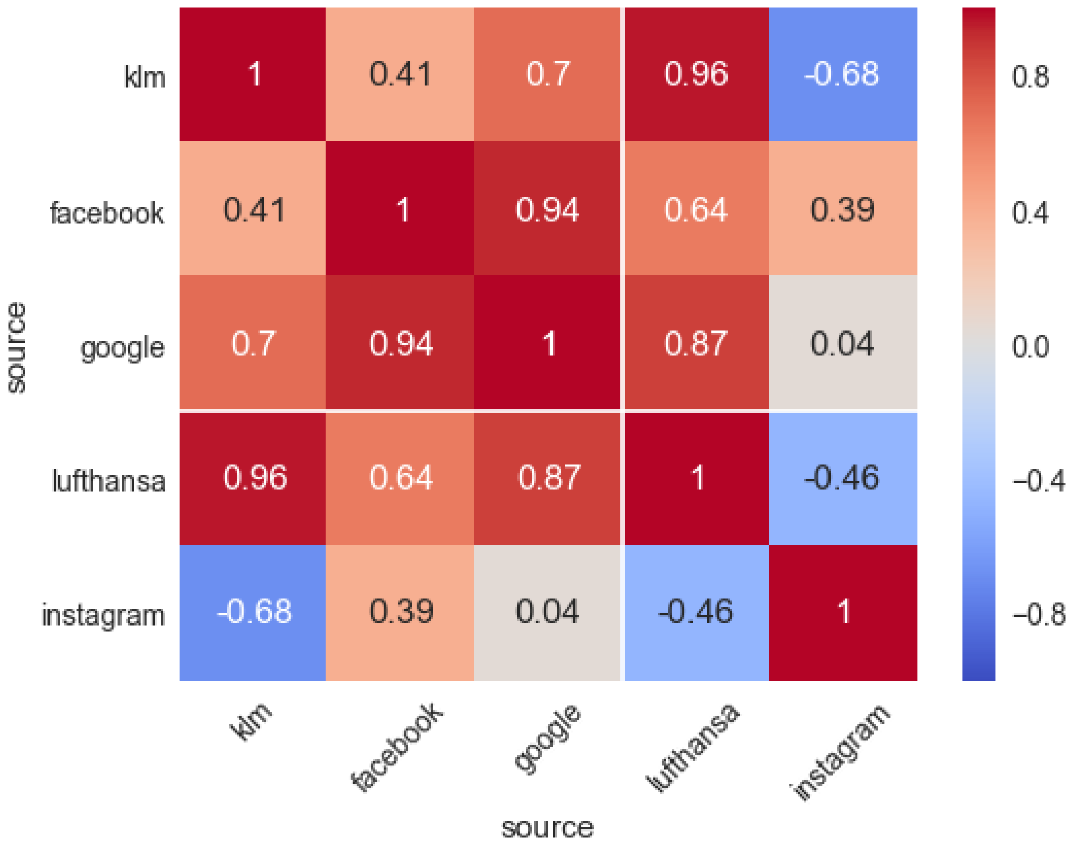
Exactly –1. A perfect downhill (negative) linear relationship

–0.70. A strong downhill (negative) linear relationship
–0.50. A moderate downhill (negative) relationship
–0.30. A weak downhill (negative) linear relationship

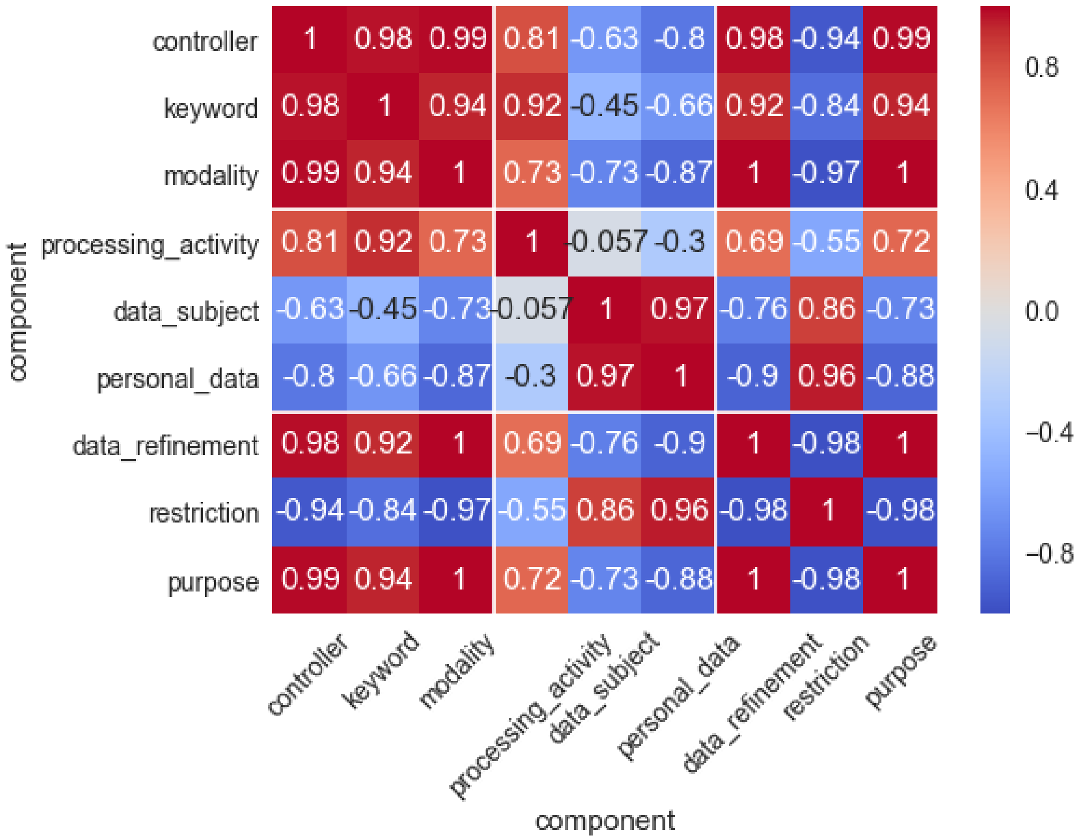
0. No linear relationship

+0.30. A weak uphill (positive) linear relationship
+0.50. A moderate uphill (positive) relationship
+0.70. A strong uphill (positive) linear relationship
Exactly +1. A perfect uphill (positive) linear relationship

```
In [44]: correlation_sources = create_correlation_heatmap(df_source.T, (8,6))
         # correlation_sources.savefig("charts/correlation_heatmap_sources.png")
         correlation_sources
```



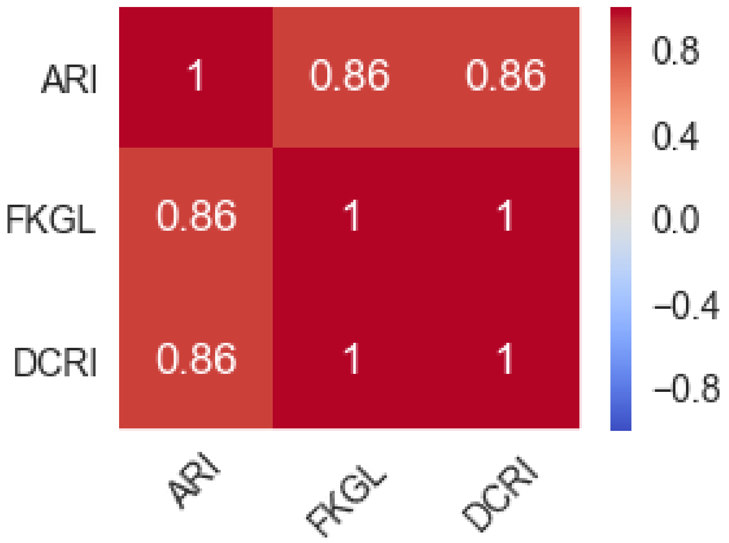
```
In [38]: correlation_components = create_correlation_heatmap(df_component.T,(8,6))
         # correlation_sources.savefig("charts/correlation_heatmap_sources.png")
         correlation_components
```



	controller	keyword	modality	processing_activity	data_subject	personal_data	data_refinement	restriction	purpose
controller	1	0.98	0.99	0.81	-0.63	-0.8	0.98	-0.94	0.99
keyword	0.98	1	0.94	0.92	-0.45	-0.66	0.92	-0.84	0.94
modality	0.99	0.94	1	0.73	-0.73	-0.87	1	-0.97	1
processing_activity	0.81	0.92	0.73	1	-0.057	-0.3	0.69	-0.55	0.72
data_subject	-0.63	-0.45	-0.73	-0.057	1	0.97	-0.76	0.86	-0.73
personal_data	-0.8	-0.66	-0.87	-0.3	0.97	1	-0.9	0.96	-0.88
data_refinement	0.98	0.92	1	0.69	-0.76	-0.9	1	-0.98	1
restriction	-0.94	-0.84	-0.97	-0.55	0.86	0.96	-0.98	1	-0.98
purpose	0.99	0.94	1	0.72	-0.73	-0.88	1	-0.98	1

7.2 Readability

```
In [39]: create_correlation_heatmap(df_merge[['ARI','FKGL','DCRI']],(4,3))
```



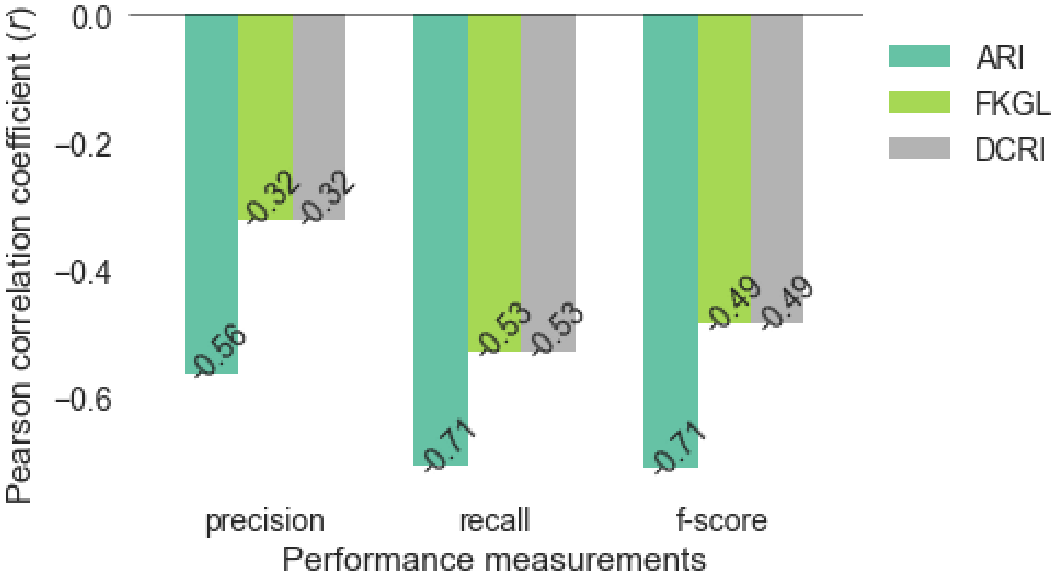
	ARI	FKGL	DCRI
ARI	1	0.86	0.86
FKGL	0.86	1	1
DCRI	0.86	1	1

7.3 Performance vs Readability

```
In [40]: fig = plt.figure()
         title = "Comparing the observations of the three participants."
         ax = df_perf_read.iloc[:3].plot(y=['ARI','FKGL','DCRI'],
                                         kind="bar",
                                         width=0.7,
                                         cmap='Set2',
                                         fontsize=14) # title=title,
         for p in ax.patches:
             ax.annotate(str(round(p.get_height(),2)), (p.get_x() * 1.005, p.get_he
         # ax.set_ylim([-.8,.8])
         ax.set_axis_bgcolor('white')
         lgd = ax.legend(loc="upper right", bbox_to_anchor=(1.3, 1))

         plt.xticks(rotation=0)
         plt.axhline(0, color='black',linewidth=.5)
         plt.ylabel('Pearson correlation coefficient ($r$)')
         plt.xlabel('Performance measurements')
         plt.show()
         fig = ax.get_figure()
         fig.tight_layout()
         # save figure
         # fig.savefig('charts/pearson_performance_readability.png',
         #             bbox_extra_artists=(lgd,),
         #             bbox_inches='tight')
```

```
C:\Users\LBrakenhoff\AppData\Local\Continuum\Anaconda2\envs\thesisEnv\lib\site-pack
  import sys
```

```
<matplotlib.figure.Figure at 0x5251ccf8>
```


Pearson correlation coefficient (r)
0.0
−0.2
−0.4
−0.6
-0.56
-0.32
-0.32
-0.71
-0.53
-0.53
-0.71
-0.49
-0.49
precision
recall
f-score
Performance measurements
ARI
FKGL
DCRI

D.4 Calculating the Observers Agreement with Cohen's Kappa

Cohens_Kappa

October 7, 2017

1 Cohen's Kappa

Cohen's Kappa coefficient is a statistical measure of inter-rater reliability which many researchers regard as more useful than the percentage agreement figure, since it takes into account the amount of agreement that could be expected to occur through chance.

If the two users are in complete agreement about which content of the source should be coded at the node, then the Kappa coefficient is 1. If there is no agreement between the two users (other than what could be expected by chance), the Kappa coefficient is ≤ 0. A value between 0 and 1 indicates partial agreement.

- Calculate the expected frequency by which the agreement between users could have occurred by chance (ΣEF), by summing:
- The number of units of the source's content coded at the node by user A, multiplied by the number of units coded at the node by user B, divided by the total number of units in the source (EF1)
- The number of units of the source's content not coded at the node by user A, multiplied by the number of units not coded at the node by user B, divided by the total number of units in the source (EF2)
- Expected frequency (EF) of the agreement occurring by chance = EF1 + EF2
- Calculate the Kappa coefficient (K) as equal to:
- Total units of agreement between the two users (TA) minus the expected frequency (ΣEF) of the agreement occurring by chance, divided by the total units (TU) within the source minus the expected frequency (ΣEF) of the agreement occurring by chance: K = (TA – ΣEF) ÷ (TU – ΣEF)
- In the case where both users are in complete agreement as to how the source's content should be coded at the node, then the value of Kappa will equal 1

```
In [1]: import pandas as pd
        from sklearn.metrics import cohen_kappa_score, confusion_matrix
        # plotting
        import matplotlib.pyplot as plt
        import matplotlib
        from matplotlib import cm # colormap
        %matplotlib inline
```

1.0.1 Load Tagged Files of the Observers

```
In [2]: # df1 = pd.DataFrame.from_csv('testset/tagged/testset_p1_csv.csv', sep=';',
        df1 = pd.DataFrame.from_csv('testset/tagged/testset_p1_csv.csv', sep=';') #
        df2 = pd.DataFrame.from_csv('testset/tagged/testset_p2_csv.csv', sep=';') #
        df3 = pd.DataFrame.from_csv('testset/tagged/testset_p3_csv.csv', sep=';') #
        columns=['source','requirement','controller','requirement_type','keyword','
```

```
C:\Users\LBrakenhoff\AppData\Local\Continuum\Anaconda2\envs\thesisEnv\lib\site-pack
  if self.run_code(code, result):
```

```
In [3]: df1 = df1[columns]
        print len(df1)
        df1.head(2)
```

```
48
```

```
Out[3]:        source                                        requirement controller
        id
        1   lufthansa  Our websites may contain links to other provid...        NaN
        2   lufthansa  Lufthansa stores your personal information if ...  Lufthansa

           requirement_type keyword processing_activity         personal_data  \
        id
        1        Permission     may             contain                links
        2        permission     NaN              stores  personal information

           data_subject                       restriction purpose data_refinement
        id
        1           NaN                               NaN     NaN             NaN
        2          your  if you provide it to us yourself     NaN             NaN
```

```
In [4]: df2 = df2[columns]
        print len(df2)
        df2.head(2)
```

```
48
```

```
Out[4]:        source                                        requirement controller
        id
        1   lufthansa  Our websites may contain links to other provid...        NaN
        2   lufthansa  Lufthansa stores your personal information if ...  Lufthansa

           requirement_type keyword processing_activity         personal_data  \
        id
        1               NaN     NaN                 NaN                   NaN
        2        Permission     NaN              Stores  Personal information

           data_subject                                        restriction  purpos
        id
        1           NaN                                                NaN      Na
        2          Your  If you provide it to us yourself, for example ...      Na

            data_refinement
        id
        1                NaN
        2                NaN
```

```
In [5]: df3 = df3[columns]
        print len(df3)
        df3.head(2)
```

```
48
```

```
Out[5]:        source                                        requirement controlle
        id
        1   lufthansa  Our websites may contain links to other provid...       Na
        2   lufthansa  Lufthansa stores your personal information if ... Lufthansa

           requirement_type keyword processing_activity          personal_data  \
        id
        1        permission     may                 NaN                    NaN
        2        permission     NaN              stores  personal information

           data_subject                                        restriction purpose
        id
        1           NaN                                                NaN     NaN
        2          your  if you provide it to us yourself, for example,...     NaN

           data_refinement
        id
        1               NaN
        2               NaN
```

```
In [6]: # test answers
        y1 = list(df1['purpose'].str.strip().str.lower())
        y2 = list(df3['purpose'].str.strip().str.lower())
```

1.0.2 Calculating Cohen's Kappa

```
In [7]: def cohens_kappa(observer1, observer2):
            kappa = cohen_kappa_score(observer1, observer2)
            return kappa
```

```
In [8]: # ,'purpose','data_refinement'
        components = ['controller','requirement_type','keyword','processing_activit
        pairs = [('Observer 1 vs 2', df1, df2),('Observer 1 vs 3', df1, df3),('Obse
        sim_scores = {}

        def get_document_similarity(components):
            for pair in pairs:
        #         print pair[0]
                sim_scores[pair[0]] = {}
                for component in components:
                    try:
                        y1 = list(pair[1][component].str.lower().str.strip())
                        y2 = list(pair[2][component].str.lower().str.strip())
                        ks = cohens_kappa(y1, y2)
                        sim_scores[pair[0]][component] = ks
                    except:
                        sim_scores[pair[0]][component] = 0

            return sim_scores

        kappa_scores = get_document_similarity(components)
```

```
In [15]: df_kappas = pd.DataFrame.from_dict(kappa_scores)
         df_kappas
```

```
In [24]: df_kappas['mean'] = df_kappas.mean(axis=1)
         df_kappas.loc['mean'] = df_kappas.mean()
         df_kappas
         # df_kappas.to_csv('samples/result_cohens_kappa_19092017.csv')
```

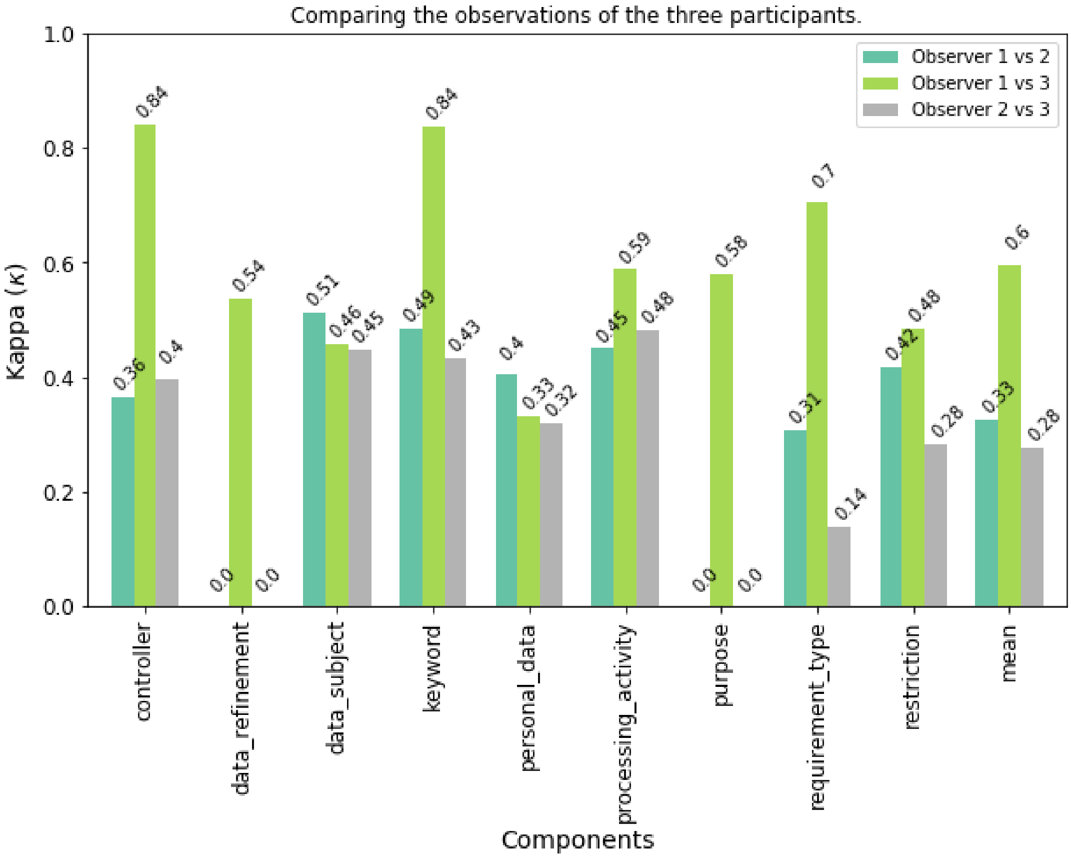
```
Out[24]:                      Observer 1 vs 2  Observer 1 vs 3  Observer 2 vs 3  \
         controller                  0.364839         0.842001         0.396985
         data_refinement             0.000000         0.538462         0.000000
         data_subject                0.513514         0.457627         0.448276
         keyword                     0.486326         0.837563         0.431953
         personal_data               0.404035         0.331787         0.320644
         processing_activity         0.450046         0.588262         0.481690
         purpose                     0.000000         0.579685         0.000000
         requirement_type            0.308108         0.704918         0.138810
         restriction                 0.416667         0.484909         0.283582
         mean                        0.327059         0.596135         0.277993

                                  mean
         controller           0.534609
         data_refinement      0.179487
         data_subject         0.473139
         keyword              0.585281
         personal_data        0.352155
```

```
        processing_activity  0.506666
        purpose              0.193228
        requirement_type     0.383945
        restriction          0.395053
        mean                 0.400396
```

```
In [18]: # df_kappas.to_latex()
```

```
In [27]: fig = plt.figure()
         title = "Comparing the observations of the three participants."
         ax = df_kappas[['Observer 1 vs 2','Observer 1 vs 3','Observer 2 vs 3']].pl
         for p in ax.patches:
             ax.annotate(str(round(p.get_height(),2)), (p.get_x() * 1.005, p.get_he
         ax.set_ylim([0,1])
         # df_kappas['mean'].plot(ax=ax)
         plt.xticks(rotation=90)
         # plt.legend(bbox_to_anchor=(1, 1), loc='upper left', ncol=1)
         plt.ylabel(r'Kappa ($\kappa$)',fontsize = 14)
         plt.xlabel('Components',fontsize = 14)
         plt.show()
         fig = ax.get_figure()
         fig.tight_layout()
         fig.savefig('charts/kappa_scores.png')
```

```
<matplotlib.figure.Figure at 0xdd78438>
```


Comparing the observations of the three participants.
Kappa (κ)
1.0
0.8
0.6
0.4
0.2
0.0
Observer 1 vs 2
Observer 1 vs 3
Observer 2 vs 3
0.36
0.84
0.4
0.0
0.54
0.0
0.51
0.46
0.45
0.49
0.84
0.43
0.4
0.33
0.32
0.45
0.59
0.48
0.0
0.58
0.0
0.31
0.7
0.14
0.42
0.48
0.28
0.33
0.6
0.28
controller
data_refinement
data_subject
keyword
personal_data
processing_activity
purpose
requirement_type
restriction
mean
Components

```
In [ ]:
```

D.5 Calculate Readability Scores

Readability_Stats

October 7, 2017

```
In [1]: from __future__ import unicode_literals

        import pandas as pd
        import nltk
        # from textstat import textstat
        # to get syllables per word
        from nltk.corpus import cmudict
        from nltk.tokenize import sent_tokenize, word_tokenize
        d = cmudict.dict()
        # for the square root
        import math

        import scattertext as st
        import spacy
        # from __future__ import unicode_literals

        # plotting
        import matplotlib.pyplot as plt
        import matplotlib
        from matplotlib import cm # colormap
        %matplotlib inline
```

```
In [2]: # d.itervalues().next()
```

1 Get Text

```
In [3]: files = [('Facebook','facebook','policies/social/fb_policy2.txt', 'training
                  ('KLM','klm','policies/webshop/klm_policy.txt', 'training'),
                  ('Google','google','policies/social/google_policy.txt', 'training
                  ('Instagram','instagram','policies/social/instagram_policy.txt',
                  ('Lufthansa','lufthansa','policies/webshop/lufthansa.txt', 'test'
                ]
```

2 Get Info

- number of sentences

- average length sentences
- number of words
- number of unique words
- ratio
- average length words
- readability (FRI)

2.0.1 Create Text Info Table

```
In [4]: df_text_info = pd.DataFrame(files, columns=['name','source','path','purpose
        df_text_info.set_index(['name'])
```

```
Out[4]:               source                                  path   purpose
        name
        Facebook    facebook        policies/social/fb_policy2.txt  training
        KLM              klm       policies/webshop/klm_policy.txt  training
        Google        google     policies/social/google_policy.txt  training
        Instagram  instagram  policies/social/instagram_policy.txt      test
        Lufthansa  lufthansa        policies/webshop/lufthansa.txt      test
```

2.0.2 Get Text Info

text info

```
In [5]: def get_text(path):
            policy = open(path, 'r')
            text = policy.read()
            text = text.replace('\n', ' ').encode('utf-8')
            return text

        # get_text('policies/social/fb_policy2.txt')
```

```
In [6]: df_fb = pd.DataFrame.from_csv('trainingset/manual_labeling_fb_csv.csv', sep
        df_klm = pd.DataFrame.from_csv('trainingset/manual_labeling_klm_csv.csv', s
        df_google = pd.DataFrame.from_csv('trainingset/manual_labeling_google_csv.c
        df_instagram = pd.DataFrame.from_csv('testset/manual_labeling_instagram_csv
        df_lufthansa = pd.DataFrame.from_csv('testset/manual_labeling_lufthansa_csv
        columns = ['source','valid','requirement']
        df_klm = df_klm[columns]
        df_fb = df_fb[columns]
        df_fb['source'] = 'facebook'
        df_google = df_google[columns]
        df_instagram = df_instagram[columns]
        df_lufthansa['valid'] = 1
        df_lufthansa = df_lufthansa[columns]
        frames = [df_klm, df_fb, df_google, df_instagram, df_lufthansa]
        df = pd.concat(frames, ignore_index=True)
        print(len(df))
        df_valid = df.loc[df['valid'] == 1]
        print(len(df_valid))
        df_valid.head()

359
194
```

```
Out[6]:   source  valid                                        requirement
        0     klm      1  We may collect and process the following categ...
        3     klm      1  When you create a personal account or register...
        4     klm      1  For business travellers, we also collect infor...
        5     klm      1  When you make a reservation or book a flight w...
        7     klm      1  In addition hereto we process information in r...
```

```
In [7]: def number_syllables_per_word(word):
            num_syllables = [len(list(y for y in x if y[-1].isdigit())) for x in d
            return num_syllables[0]

        def flesch_reading_ease(total_num_words, total_num_syllables_words,total_nu
            fre = float(206.835)-(float(1.015)*(total_num_syllables_words/total_nur
        #     fkgl = float(0.39)*(total_words/total_sentences)+float(11.8)*(total_s
            return fre

        def flesch_kincaid_grade_level(total_num_words, total_num_syllables_words,t
            fkgl = float(0.39)*(total_num_words/total_num_sentences)+float(11.8)*(t
            return fkgl

        def smog_index(total_num_polysyllables, total_num_sentences):
            smog = float(1.0430)*math.sqrt(total_num_polysyllables*(30/total_num_se
            return smog

        def automated_readability_index(total_num_characters,total_num_words,total_
            ari = (float(4.71)*(total_num_characters/total_num_words))+(float(0.5)*
            return ari

        def dale_chall_readability_index(total_num_polysyllables,total_num_words,tc
            dcri = float(0.1579)*(total_num_polysyllables/total_num_words)+float(0.
            return dcri
```

```
In [8]: df_text_info
```

```
Out[8]:         name     source                                  path   purpose
        0   Facebook   facebook          policies/social/fb_policy2.txt  training
        1        KLM        klm        policies/webshop/klm_policy.txt  training
        2     Google     google      policies/social/google_policy.txt  training
        3  Instagram  instagram  policies/social/instagram_policy.txt      test
        4  Lufthansa  lufthansa          policies/webshop/lufthansa.txt      test
```

```
In [11]: for index, row in df_text_info.iterrows():
         #     print row['source']
```

```
    total_num_sentences = 0
    total_num_syllables = 0
    total_num_polysyllables = 0
    total_num_characters = 0
    not_found_words = []
    tokens = []
    tokens_syllables = []

    df_tmp = df_valid.loc[df_valid['source'] == row['source']]
    total_num_sentences = len(df_tmp)

    # create list with tokens for each source
    for i, r in df_tmp.iterrows():
        words = r['requirement'].split()
        for token in words:
            tokens.append(token)
            total_num_characters += len(token)
            try:
                num_syllables = number_syllables_per_word(token.lower())
                total_num_syllables += num_syllables
                if num_syllables > 2:
                    total_num_polysyllables += 1
                if num_syllables > 0:
                    tokens_syllables.append(token)
            except:
                not_found_words.append(token)
                continue

#     print row['source'], total_num_polysyllables, total_num_sentences
    total_num_words = len(tokens)
    total_num_syllables_words = len(tokens_syllables)
    total_num_unique_words = len(set(tokens))

#     print total_num_words,total_num_syllables_words, total_num_sentences

    fre = flesch_reading_ease(total_num_words,
                              total_num_syllables_words,
                              total_num_sentences,
                              total_num_syllables)
    fkgl = flesch_kincaid_grade_level(total_num_words,
                                      total_num_syllables_words,
                                      total_num_sentences,
                                      total_num_syllables)
#     smog = smog_index(total_num_polysyllables, total_num_sentences)
    ari = automated_readability_index(total_num_characters,
                                      total_num_words,
                                      total_num_sentences)
    dcri = dale_chall_readability_index(total_num_polysyllables,
                                        total_num_words,
                                        total_num_sentences)


    df_text_info.loc[index,'Sentences'] = total_num_sentences
    df_text_info.loc[index,'Words'] = total_num_words
    df_text_info.loc[index,'Syllables'] = total_num_syllables
    df_text_info.loc[index,'UniqueWords'] = total_num_unique_words
    df_text_info.loc[index,'FRE'] = fre
    df_text_info.loc[index,'ARI'] = ari
    df_text_info.loc[index,'FKGL'] = fkgl
    df_text_info.loc[index,'DCRI'] = dcri
#     df_text_info.loc[index,'SMOG'] = smog
    df_text_info.loc[index,'num_not_found_words'] = len(not_found_words)

    print fre, fkgl, ari, dcri
df_text_info
```

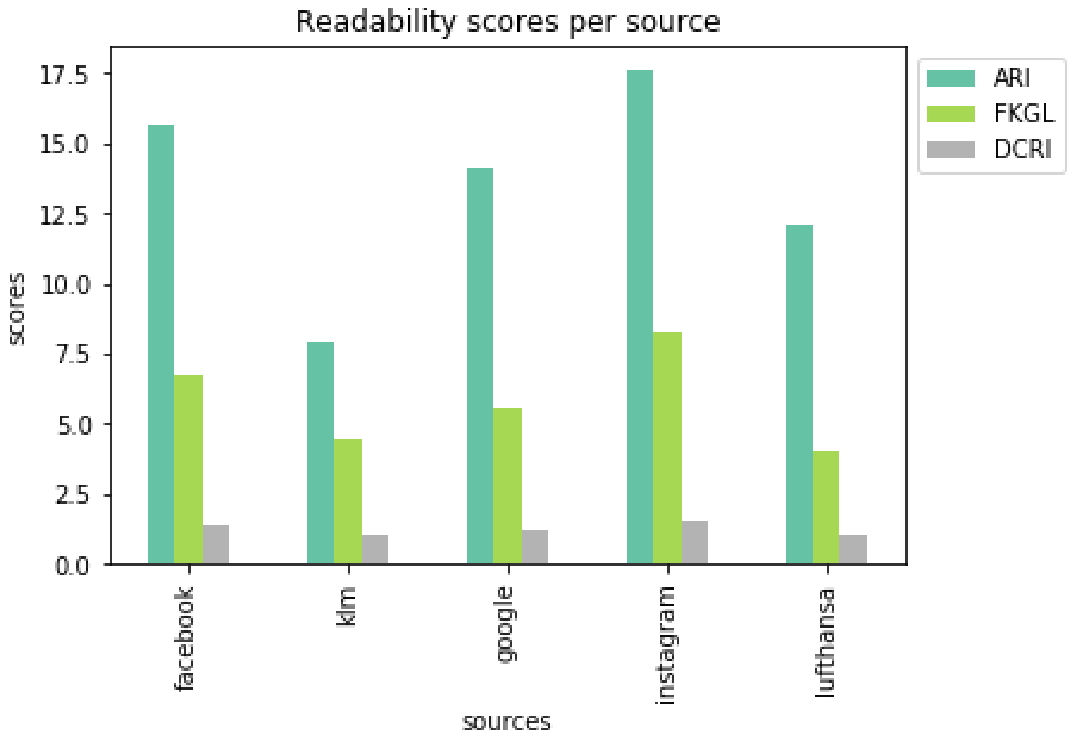
```
 97.875 6.74 15.62 1.3392
102.95 4.4 7.91 1.0416
100.92 5.57 14.12 1.1904
94.83 8.3 17.62 1.5376
103.965 4.01 12.12 0.992
```

```
Out[11]:         name     source                                  path   purpose  \
         0    Facebook   facebook        policies/social/fb_policy2.txt  training
         1         KLM        klm       policies/webshop/klm_policy.txt  training
         2      Google     google     policies/social/google_policy.txt  training
         3  Instagram  instagram  policies/social/instagram_policy.txt      test
         4  Lufthansa  lufthansa        policies/webshop/lufthansa.txt      test

            Sentences   Words  Syllables  UniqueWords      FRE    ARI  FKGL    DCRI
         0       46.0  1250.0     1889.0        396.0   97.875  15.62  6.74  1.3392
         1       53.0  1117.0     1722.0        356.0  102.950   7.91  4.40  1.0416
         2       50.0  1200.0     1863.0        438.0  100.920  14.12  5.57  1.1904
         3       31.0   975.0     1377.0        370.0   94.830  17.62  8.30  1.5376
         4       14.0   287.0      452.0        169.0  103.965  12.12  4.01  0.9920

            num_not_found_words
         0                103.0
         1                 81.0
         2                133.0
         3                114.0
         4                 31.0
```

```
In [12]: # df_text_info.to_csv('samples/text_info.csv')
```

```
In [13]: df_text = df_text_info.copy()
         df_text = df_text_info.set_index('source')
```

```
In [14]: readability_scores = df_text[['ARI','FKGL','DCRI']].T #.to_latex()
         readability_scores #.to_latex()

Out[14]: source  facebook     klm   google  instagram  lufthansa
         ARI      15.6200  7.9100  14.1200    17.6200     12.120
         FKGL      6.7400  4.4000   5.5700     8.3000      4.010
         DCRI      1.3392  1.0416   1.1904     1.5376      0.992

In [15]: ttl = 'Readability scores per source'
         df_text_info.cumsum();
         fig = plt.figure();
         # color_list = ['#3b5998','#00a1de',  '#ff002b'] # klm, fb, google
         ax = df_text[['ARI','FKGL','DCRI']].plot(kind='bar', title=ttl,  cmap='Set
         lgd = ax.legend(bbox_to_anchor=(1, 1), loc='upper left', ncol=1)
         plt.ylabel('scores')
         plt.xlabel('sources')

         # plt.grid(True)
         plt.show()
         fig = ax.get_figure()
         fig.tight_layout()
         # save figure
         # fig.savefig('charts/readability_scores.png', bbox_extra_artists=(lgd,),

<matplotlib.figure.Figure at 0x14689198>
```


Readability scores per source
scores
facebook
klm
google
instagram
lufthansa
sources
ARI
FKGL
DCRI