

Bootstrapping the CRISP-DM Process

MASTER THESIS

By: Marcin Kais, 4289684
First supervisor: dr Marco Spruit
Second supervisor: Vincent Menger



Universiteit Utrecht

Table of Contents

1 Research Plan.....	8
1.1 Problem Statement.....	8
1.2 Research question.....	9
1.3 Chosen Research Methods.....	10
1.4 Literature Review Protocol.....	11
Objectives of the Review.....	11
Search Strategy.....	11
Inclusion and Exclusion Criteria.....	13
1.5 Milestones for the Key Phases of the Project.....	13
Literature review.....	13
CRISP-DM guide analysis.....	13
Automated data science software analysis.....	14
Natural Language Processing – analysis of applications.....	14
A survey.....	14
Design science research:.....	14
2 Literature Review.....	16
2.1 Data Mining, Text Mining, and Information Extraction.....	16
2.2 Cross-Industry Standard Process for Data Mining.....	18
2.3 Semi-Automated Data Analysis Initiatives.....	20
2.4 Natural Language Processing.....	22
Phonetic/phonological level.....	22
Morphological level.....	23
Lexical level.....	23
Syntactic level.....	23
Semantic level.....	23
Discourse level.....	24
Pragmatic level.....	24
Current state of the art.....	24
Common applications.....	24

3	Setting Business Goals in Practice.....	25
3.1	Questionnaire overview	26
3.2	Background data of the respondents' organizations	29
3.3	Methods and processes used	31
3.3	Business understanding and its problems in practice.....	32
3.4	Tools used at initial stages of data mining.....	33
3.5	Conclusions of the survey	34
4	Proposed Solutions.....	35
4.1	Standardizing the interviews with stakeholders	35
4.1.1	SMART Goals.....	35
4.1.2	Affinity Diagramming and the KJ Method	36
4.2	Bootstrapping the review of corporate documentation & assessment of the project background.....	37
4.3	Integrated solution for specifying project goals	41
4.4	Success Criteria	41
5	Development of the tool aiding with business goal specification	43
5.1	Method.....	43
5.2	Tool.....	45
5.2.1	Technologies Used	47
5.2.2	Explanation of the key components.....	50
	Sentence Extraction Component	50
	Natural Language Processing Component	51
	Goal Formatting Component	53
	Filtering of the results	54
	Interview and Evaluation Component.....	55
5.2.3	Sample Results.....	55
6	Evaluation	61
	Measures	61
	Data Retrieved.....	62
	Results	71
7	Conclusions.....	76

7.1 Research Questions.....	76
7.2 Limitations.....	77
7.3 Future Research	77
References.....	79
Appendix A.....	83
Full set of answers to the questionnaire.....	83
Appendix B	88
Process-Deliverable Diagram.....	88

List of Figures

Figure 1: Research Plan.....	15
Figure 2: CRISP-DM process stages	18
Figure 3: Questionnaire as a Google Form.....	26
Figure 4: Results of the first question of the survey	30
Figure 5: Respondents' choice of DM processes	31
Figure 6: DM processes divided by groups.....	31
Figure 7: Most difficult steps of business understanding phase.....	32
Figure 8: Difficulty of the BU tasks, as assessed by the respondents.....	33
Figure 9: Stanford CoreNLP system architecture	47
Figure 10: PDD of extracting sentence fragments with goal-related keywords.....	51
Figure 11: PDD of extracting goals.....	52
Figure 12: PDD of formatting goals	54
Figure 13: Precision and recall illustrated	61

List of Tables

Table 1: Feature comparison of automatic data analysis software	21
Table 2: Interpretation levels used by different types of natural language processing systems...	25
Table 3: Questions and choice of answers in the questionnaire.....	29
Table 4: Types of success criteria for IT/IS projects.....	41
Table 5: Penn Treebank tags.....	49
Table 6: Sample results of the application of the tool	60
Table 7: Results of the application of the tool to Enron CEO's "sent" folder	71

List of Abbreviations

Abbreviation	Meaning
BU	Business Understanding
CRISP-DM	Cross-Industry Standard Process for Data Mining
DM	Data Mining
IE	Information Extraction
IEEE	Institute of Electrical and Electronics Engineers
IT	Information Technology
IS	Information Systems
JEE	Java Platform, Enterprise Edition
KDD	Knowledge Discovery in Databases
KJ	Kawakita Jiro, author of the KJ method
NLP	Natural Language Processing
NP	Noun Phrase
PDD	Process-Deliverable Diagram
POS	Part of Speech
PP	Prepositional Phrase
RQ	Research Question
RSQ	Research Sub-question
SDLC	Software Development Life Cycle
SEMMA	Sample, Explore, Modify, Model, and Assess
SMART	Specific, Measurable, Achievable, Relevant and Time-sensitive
VP	Verb Phrase

1 Research Plan

The following section of this document introduces the research for this thesis, starting with the identification of a gap in practice and the problem statement, closely followed by the defined research question, and the research plan with the justification of chosen methods, and planned project milestones. The protocol of the literature review is also included.

1.1 Problem Statement

The Cross-Industry Standard Process for Data Mining (CRISP-DM), the most commonly used data mining methodology (KDnuggets, 2014), is applied in a wide range of industries all around the world as a de facto worldwide standard for structuring the data mining process. However, it is often criticized by academics due to a perceived immaturity of some of its stages, lack of tools, or standard guidelines (refer to section 2. of this document, *Literature Review*).

My research project concerns the lack of reusable, standard tools or guidelines at the first stage of the process, namely the *Business Understanding* phase. *Business Understanding* concerns the identification of business objectives of the data mining effort, and their further translation to data mining goals along with their success criteria, risk analyses, as well as cost/benefit analyses. The identification of business objectives is usually done by gathering the background information about the current situation, conducting interviews and involving all the key stakeholders in these discussions, documenting the results and producing a list for further translation to data mining goals. Then, available data and personnel are introduced to the equation, as well as the risks the project carries and the contingency plans for all of them. The requirements are the next step of the process; a requirement in this sense is the business goal produced earlier, but it also considers:

- Security and legal restrictions on the project,
- Project scheduling along the personnel,
- Target deployment method of the results.

The requirements are further subjected to a cost/benefit analysis, and translated to data mining goals and their success criteria. The project plan is produced at the end of the business understanding phase, along with the assessment of the tools and techniques.

Due to the aforementioned lack of tools and guidelines, the initial stage of the process proves to be especially challenging for small and medium enterprises, which have a hard time defining the business goals of their data mining undertaking. This is problematic for them, especially in case of only introducing data mining as a viable technique for laying the foundations for the growth of the enterprise.

The research project is designed to help with this exact situation. The expected outcome of the project is the creation of a method and/or a prototype, preferably to be embedded within the

existing Business Understanding phase of the CRISP-DM model, which could facilitate the identification of the business objectives for small and medium enterprises, by the means of using corporate data and other resources, already existing within the organization. Specifically, these corporate data could be extracted and analyzed in order to discover a general direction for the data mining effort needed at the company.

As the structured data, if stored by a business, is often analyzed, or at least is feasible to analyze with abundance of tools and services available on the market, the research project I am conducting intends to explore the possibilities of using unstructured data to help define the business goals. These could be later translated to data mining goals. For this reason natural language processing is one of the leading themes of this research.

1.2 Research question

As introduced in the previous subsection, my research project concerns the facilitation of implementing the *Business Understanding* phase of CRISP-DM in practice. The focus lies in small and medium enterprises, due to their frequent inability to deal with the initial stages of their data mining efforts. It is, however, not restricted to them.

As a result of the literature review, and the exploration of related fields, tools and services, I have formulated the following research question:

RQ: To what extent can enterprises semi-automatically uncover the definition of their data mining business goals using text analytics?

This is set to investigate the possibility of (semi-) automatic corporate data extraction and subjecting it to thorough analyses, in order to facilitate the formulation of business goals for data mining. Since CRISP-DM is the most prominent data mining methodology, the research question is going to be investigated within the frames of its context. To aid with particular steps of my research, I have also formulated the following sub-questions to the main research question:

RSQ1: Are there any theoretical models designed to help with data mining business goals definition?

This question is set the initial direction for the literature review, and to investigate the theoretical models and frameworks, which were designed to obviate the problems encountered at the first stages of data mining efforts.

RSQ2: What do enterprises do in practice to define their data mining business goals?

The second sub-question refers to the same idea, however this time to its practical implementations in the field instead of academic research. The survey distributed among professionals will play a vital role in investigating this problem.

RSQ3: *Which textual resources could be helpful to bootstrap the business understanding phase of data mining efforts in practice, and in what way?*

The third of the sub-questions is formulated to investigate the usability, as well as availability of the resources, which could be used to help small and medium enterprises define their business goals. This will mainly concern the relevance of corporate data of various types and origins, and the feasibility of using them. The focus will lie on unstructured textual data, as it is quite often not explored, due to lack of feasible tools and services, as well as no specific methods to deal with this type of resources.

Generally, the investigation is also concerned with the exploration of the possibilities of creating a generalizable method of bootstrapping the *Business Understanding* phase of the CRISP-DM process.

1.3 Chosen Research Methods

The research implies the creation of new knowledge, which could be later used in the field, to design solutions to the problems within that field. The result of the research is going to be a new, knowledge-containing artefact, whether it turns out to be a method, or a tool prototype - generally a new artefact, which is an innovative product. For this reason, a design science research approach is needed as the main research method during this project.

In order to reach the artefact development stage, the research has to start with thorough literature reviews, confirming the existence of a gap in knowledge, as well as the exploration of already published, related works. The literature review will also concern the fields which play a great role in exploring unstructured data, introduce the basics, and set a direction for using them within the development stage of the research.

In pursuance of gaining knowledge of what is commonly implemented in the field to bypass the aforementioned gap in knowledge/practice, I am going to publish a survey about this problem. The results, which these two approaches will uncover, will have a great impact on the next steps of the research, by building on the approaches which are already used in practice.

As already mentioned, the expected result of the research project is a knowledge-containing artefact, which could be a generalizable solution to the problem of lack of tools or guidelines towards the implementation of the *Business Understanding* stage within the context of Cross-Industry Standard Process for Data Mining. This artefact will be evaluated on the base of a case study with one or multiple cases, provided either by professionals working in the field, open-source data sets, or academic literature about the subject.

As suggested by Prat, Comyn-Wattiau, & Akoka (2014) in their paper on design science artefact evaluation, this evaluation will concern the adherence to the goals of the artefact (its efficacy, validity and generality), as well as consistency with the environment (people, organizations,

technology), its structure (completeness, clarity, level of detail, etc.), activity (performance, accuracy, efficiency), and possible evolution and robustness.

1.4 Literature Review Protocol

This subsection of the document concerns the protocol of the literature review done during the work on this stage of the project, as well as the future work on the next steps. The objectives of the review are provided, closely followed by the explanation of the search strategy, and inclusion and exclusion criteria for the relevant articles.

Objectives of the Review

The literature review is an integral part of this research and one of the first steps which shape the main part of the investigation. The objective of this review is to gather all the relevant knowledge about data mining, traditionally available data, Cross-Industry Standard Process for Data Mining, its initial stages, as well as their shortcomings, identified by the academics.

The review further explores related works conducted in the field to bypass these shortcomings. This is covered in order to delve into the current standards, as well as the recent ideas of enhancing the CRISP-DM process. The results should shape the design science research in such a way, that it does not repeat any of the already proposed solution, but builds on them to propose a new artefact, which could be a valuable solution to the existing gap in knowledge/practice.

Aside from the Cross-Industry Standard Process for Data Mining, and its identifiable shortcomings, the literature review explores Natural Language Processing (NLP), as it is a mean to explore more data than it is traditionally done within businesses.

Search Strategy

By virtue of virtually unrestricted access to relevant academic papers, books, or journals, provided by the university library, the searching process of the articles was automated. The process was conducted through Google Scholar search engine, using the Utrecht University library proxy (scholar.google.com.proxy.library.uu.nl).

Then, search queries were entered. For instance, to uncover as much information about the current standards of data mining, the following search queries were used:

- "Data Mining"
- "Data Mining Algorithm(s)"
- "Tabular Data Mining"
- "Multirelational Data Mining"
- "Multi-relational Data Mining"
- "Relational Data Mining"

- "Propositional Data Mining"
- "Structured Data Mining"
- "Semi-structured Data Mining"
- "Data Mining Efficiency"
- "Data Mining Scalability"
- "Inductive Logic Programming"
- "ILP"
- "Data Mining Optimization"
- "Data Mining Databases"

To maximize the amount of the relevant search results, in all of the above queries, "Data Mining" was used interchangeably with "Pattern Recognition", "Knowledge Discovery", and "Information Extraction".

Also, to analyze CRISP-DM, and other data mining standards, as well as to uncover their well-known deficiencies, the following search queries were used:

- "CRISP-DM"
- "CRISP-DM Automation"
- "CRISP-DM Business Understanding"
- "CRISP-DM Business Application"
- "CRISP-DM Implementation"
- "CRISP-DM Preparation"
- "CRISP-DM Model"
- "CRISP-DM Process Model"
- "CRISP-DM Requirements"
- "Data Mining Automation"
- "Data Mining Business Understanding"
- "Data Mining Business Application"
- "Data Mining Business Intelligence"
- "Data Mining Framework"
- "Data Mining Implementation"
- "Data Mining Knowledge Engineering"
- "Data Mining Knowledge Management"
- "Data Mining Machine Learning"
- "Data Mining Methods"
- "Data Mining Methodologies"
- "Data Mining Model"
- "Data Mining Objectives"
- "Data Mining Preparation"
- "Data Mining Process Model"
- "Data Mining Requirements"
- "Data Mining Requirements Elicitation"
- "Data Mining Standards"

In this case, "Data Mining" was also used interchangeably with "Pattern Recognition", "Knowledge Discovery", and "Information Extraction", while "CRISP-DM" with its full name, "Cross-Industry Standard Process for Data Mining".

The time period covered by the literature review was restricted to the years 2000 - 2016, however the ancillary use of snowballing method also returned relevant results from before the year 2000. To accompany the literature review, an exploration of the latest automatic data analysis initiatives was also conducted. This was done on the base of an automated data science software list published on the KDnuggets website, as well as additional searches through Google search engine.

As the decision to make use of natural language processing was taken, the literature review was expanded to include this subject.

Inclusion and Exclusion Criteria

All types of scientific studies, as well as grey literature (e.g. IBM CRISP-DM whitepaper) are eligible for inclusion in the study review. Newspaper articles and other types of popular media, as well as Wikipedia articles, and content published on websites with no references to reliable sources, are read, but excluded from the review, due to the lack of possibility of their evaluation at a scientific angle.

1.5 Milestones for the Key Phases of the Project

The plan of this research project consists of a number of approaches, done either sequentially, or in parallel. The techniques used up to now, as well as envisioned for the remaining stages of the project are the following:

Literature review

The literature review, which has started at the beginning on the work on the thesis concerns the investigation about general ideas and developments in the fields of data and text mining. It is also conducted in order to pinpoint the gap in the literature and practice, which the thesis project is supposed to counter. This component of the research is set to be continued further, until the work on the artefact starts.

CRISP-DM guide analysis

The analysis of the official CRISP-DM whitepaper was conducted to get an understanding about all the stages of the process, as well as to investigate the official guidelines when it comes to the slightly problematic business understanding phase of it.

Automated data science software analysis

The automated data science software initiatives provide an idea of what is possible in this field and how it could be used to bootstrap the CRISP-DM process.

Natural Language Processing – analysis of applications

As this research project intends to explore the possibilities of applying natural language processing to the business understanding phase of CRISP-DM, a thorough investigation of NLP methods and levels is required before using it within the artefact design process.

A survey

The survey was conducted to gain deeper understanding of how data mining is conducted in the field, if the professionals have issues deploying the business understanding phase, and how they deal with them.

Design science research:

Definition of success criteria,

This stage will sum up everything learned in the previous phases, in order to define the success criteria for the artefact to be developed.

Development of the artefact

The artefact will be developed on the basis of the success criteria, defined on the foundation of literature review and practical analyses.

Evaluation of the results

The evaluation will concern the adherence to the goals of the artefact, as well as consistency with the environment, its structure, activity, and possible evolution and robustness.

All this is to be done in parallel with the written part, which consists of two deliverables: this document, and the final paper. The envisioned workflow for this research can be presented in form of following charts:

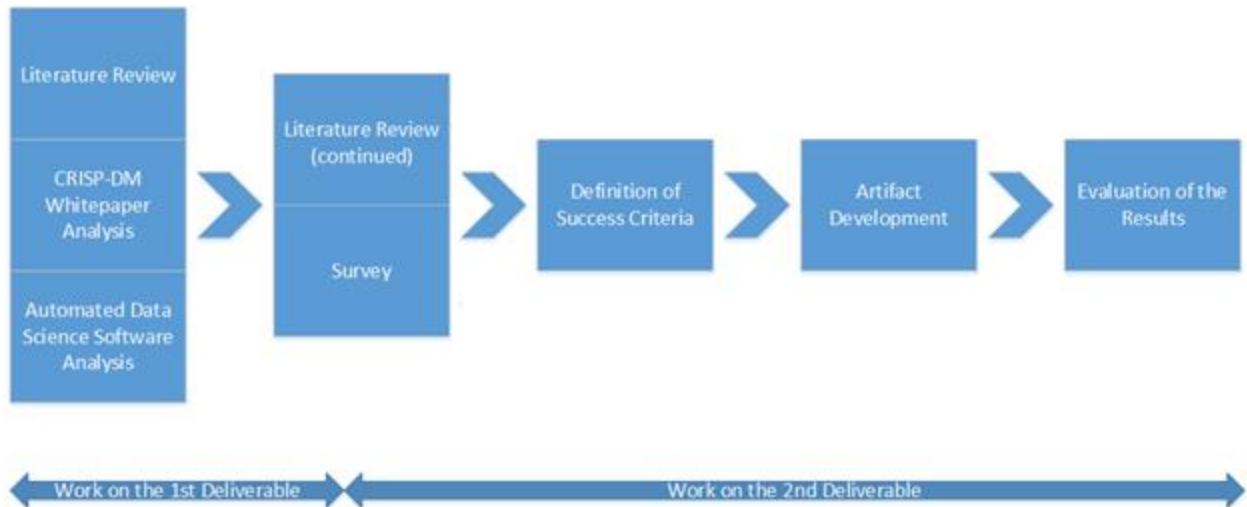


Figure 1: Research Plan

The research plan depicted above was a subject to change, especially in case of uncovering some interesting information, which could lead to another, not yet envisioned, stage in the process. In any case however, the main phases and milestones of the project remained unchanged.

2 Literature Review

This section of the document presents the literature review conducted to gain knowledge about data mining, text mining, information extraction, and other, related notions. Their types and commercial, as well as non-commercial applications are briefly analyzed.

Further, the CRISP-DM process is examined and its previously identified deficiencies are provided. Additionally, this section touches upon some recent initiatives in semi-automated data analysis and explores the potential of bootstrapping the CRISP-DM process with this type of innovation.

As the research is intended to explore the possibilities of applying Natural Language Processing to the *Business Understanding* phase of CRISP-DM, the review introduces that notion, possible applications, as well as the idea of different levels of NLP.

2.1 Data Mining, Text Mining, and Information Extraction

In the recent years, our capabilities of creating and storing data grew rapidly. A new field has emerged to satisfy the needs of businesses to make use of these data. This field, concerning the notion of discovering interesting patterns in data has been given a variety of names, including - but not limited to - data mining, knowledge discovery, knowledge extraction, information discovery, information harvesting, data archaeology, pattern recognition, or data pattern processing (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

The most widely used terms, *Knowledge Discovery (in Databases)*, and *Data Mining* are usually used interchangeably, however some researches argue that data mining is only the algorithmic step within the overall process of discovering useful knowledge in data (Fayyad et al., 1996). *Knowledge discovery* was preferred by these researchers as the term describing the whole process, due to the inclusion of the results of the process in the name itself; in spite of that, it did not gain that much popularity, and for instance such methodologies of knowledge discovery as CRISP-DM use the term *data mining* in their names. Since this exact method is the subject of this thesis, the entire process will be referred to as *data mining* within this paper.

Data mining algorithms look for patterns in data. The data (Džeroski, 2003) can be for instance tabular (just a data table), multi-relational (relational database with multiple tables, related to each other), semi-structured (e.g. XML), or even stored within text written in natural language. For all these data structures, different approaches exist - propositional, looking for patterns in a single data table; relational, used for multiple, interrelated data tables; or information extraction for semi-structured or unstructured data.

For instance, propositional data mining assumes that each individual has its own set of attributes, or characteristics. Thus, the individuals can be represented as sets of attribute-value pairs, and the database structure is reduced to a table with records representing said individuals, and columns

corresponding to attributes. The algorithms then build models by working on subgroups defined by constraints on propositional data (Madhusudhan, & Rao, 2015). Most of the traditional data mining algorithms operate on these single relations, or tables (Domingos, 2003). This is contrary to their real-world application, where most of the databases are multi-relational, with multiple tables. This issue occurred due to the complexity of multi-relational data mining context, which turned out to be a massive test for the efficiency and scalability of the algorithms and implementations (Blockeel & Sebag, 2003).

There are multiple ways studied to facilitate multi-relational data mining; for instance Inductive Logic Programming (ILP), which is the most widely used set of approaches towards dealing with multi-relational classification (Madhusudhan et al., 2015). The principles of ILP used for multi-relational data mining can be summed up as automatic hypothesis deduction by the means of analyzing given background situation, positive and negative examples (Mooney, Melville, Tang, Shavlik, Castro Dutro, Page, & Costa, 2002). Another interesting approach is extending local pattern extraction to the multi-relational context, while at the same time reducing the noise from further relations (Cerf, Besson, Nguyen, & Boulicaut, 2013). Other work, applicable in this context, however not necessarily concerning this exact subject, has been conducted as well. For instance, one of the propositions concerned representing the relational databases as graphs (Angles & Gutierrez, 2008). The focus here, however, lied in database representation, and querying, not optimizing multi-relational data mining.

As the amount of data stored increases rapidly, and a great part of it is not stored in any structured way, there is a need for different algorithms, not only these mining tabular or relational databases. Text mining, information retrieval and information extraction form an answer to this problem. Where information retrieval is query-based (Manning, Raghavan, & Schütze, 2008), information extraction is all about extracting and structuring the unstructured information found in textual documents (Aggarwal & Zhai, 2012). An information extraction system can be used as a front-end for precise information retrieval, but also for text routing, or even to create input for an intelligent agent, which requires understanding of the information conveyed by the text (Soderland, 1999).

Text mining is a broader term, a key part of which is the aforementioned information extraction (Nahm & Mooney, 2002). One of the definitions of text mining, provided by Hearst (1999), explains text mining as the automatic discovery of new, previously unknown, information from unstructured textual data. Three major tasks of text mining are then seen as (Ananiadou & McNaught, 2006) information retrieval (or gathering of the relevant pieces of text), information extraction (extraction of the interesting information from previously gathered documents), and data mining (here - discovering associations and patterns in the previously extracted information).

Natural Language Processing (NLP) is loosely tied to text mining, however it has more applications than that. It generally refers to computer systems, which analyze, attempt to understand, manipulate, and/or produce natural human language, such as e.g. English or Chinese (Chowdhury, 2003). Other functions, aside from aiding with text mining, include translations, speech

recognition, generating summaries, or generating a dialogue with the user as a part of the user interface of an application. The future of NLP is firmly linked to the future of artificial intelligence, as it is an AI-complete problem, which means that its ultimate solution requires a synthesis of a human-level intelligence as a precondition (Bergmair, 2004).

2.2 Cross-Industry Standard Process for Data Mining

Cross-Industry Standard Process for Data Mining (CRISP-DM) is the most commonly used data mining methodology (Sharma, Osei-Bryson, & Kasper, 2012). According to a relevant online poll, 42% of respondents choose it as their main method (KDNuggets, 2014). The process is divided into a number of phases, namely - (1) business understanding, (2) data understanding, (3) data preparation, (4) modelling, (5) evaluation, and (6) deployment. This literature review concerns the academic papers written about the process (and its respective counterparts), and focuses on its maturity and detailedness, especially when it comes to its initial stage - the business understanding phase.

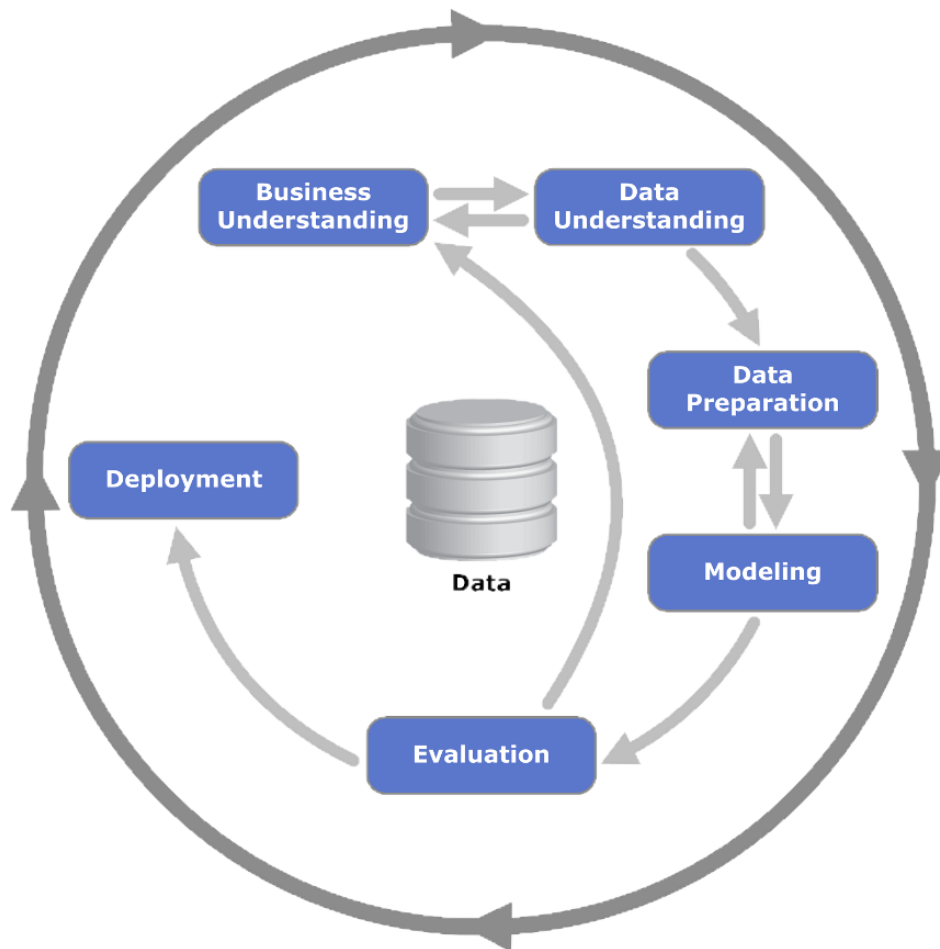


Figure 2: CRISP-DM process stages (IBM SPSS Modeler CRISP-DM Guide, 2011)

Data mining has applications in many domains. Kohavi & Provost (2001) for instance focus on the more conspicuous ones, such as web merchandising or electronic commerce. They state that even though these are the perfect applications for data mining efforts, the current state of practice in these areas is still not mature enough, supporting these claims by addressing the issues of how manual (with the exception of DM algorithms) the data mining processes still are. They suggest the need of performing more research, especially in the areas of knowledge discovery other than the algorithmic phase.

As identified by Wang & Wang (2008), the first step of any data mining process - not only CRISP-DM - is understanding of the problem owner's (business insider's) concerns. The concerns are then translated to data mining concepts, and the data mining goals are defined. Considering this research focuses on the CRISP-DM process, these activities can be mapped to its initial stage, business understanding, and its sub-tasks (determine business objectives, assess situation, determine data mining objectives, produce project plan). Especially the first of these sub-tasks carries a certain weight, and is crucial to the project as a whole, for the reason of affecting every single task in the remainder of the data mining project. However, is there any formal method of implementing it? Sharma and Osei-Bryson (2009) write about how little detailed advice for data miners there is, advice on how to actually carry out a given step. They identify the business understanding phase as the stage, where this issue is particularly dominant. They also discuss how vital the determination of business objectives is, and support it by giving examples of dependencies between the phases, as well as all the extra tasks, which have to be carried out if the business goals happen not to be defined appropriately. They go as far as considering the business goals definition as a standalone phase of the project, not just the first step of the first stage, and argue that CRISP-DM does not highlight this fact enough. Vleugel, Spruit, & van Daal (2010) also elaborate on this problem, by introducing hypotheses as a new approach of retrieval of source-independent data, in the first phase of what they call "A Three-Phases Model".

The importance of the initial stages of data mining effort, and the prominent lack of support for them is a point of view shared with other researchers, for instance Becher, Berkhin, & Freeman (2000) introduce their paper by stating how critical of a step the selection of most appropriate attributes is, and support this claim by listing the dependencies between this task and further phases of the data mining effort. Berry & Lindoff (2000) in their case study review, present the fact of how there is a lack of unified framework for implementing the business understanding phase, and show that this crucial step is often implemented in an ad-hoc manner. Sharma et al. (2012) explain this phenomenon by describing the general lack of support towards how this implementation should be performed. The only official guideline for the BU phase is the suggestion to use an organizational chart to identify divisions, managers, their responsibilities etc. (CRISP-DM, 2011). This lack of support contrasts with the importance of this stage of the data mining process. Marbán, Segovia, Menasalvas, & Fernández-Baizán (2008) state that even though CRISP-DM is an improvement when compared to the previous standards in the data mining community, it is still not mature enough to deal with the complex problems it needs to address,

one of the examples of this situation, given by them, yet again concerns the business understanding phase and the lack of business modelling procedures, formal tools, or methods.

How can this problem be tackled? Several researchers address the need of building an integrated process model (Brachman & Anand, 1996; Kurgan & Musilek, 2006; Sharma et al., 2012). This model could be subsequently used for automation, or semi-automation of some of the tasks within the process, not necessarily only these referring to the modelling stage. Britos, Dieste, & García-Martínez (2008) focus on identifying concepts related to the requirements specifications of data mining processes, which partially overlaps with the business understanding phase of CRISP-DM. Their article proposes a data mining project requirements elicitation process, and its documentation method, through a template set.

Another solution for making current data mining methods more mature, proposed by Marbán et al. (2008) is looking at data mining from an engineering perspective, and reusing the experiences gained from the software engineering field throughout over forty years of its existence. The authors recommend reusing ideas and concepts from IEEE Std 1074 and ISO 12207 software engineering model processes, as a way to enhance CRISP-DM and make it a data mining engineering standard. The most interesting ideas concerning the improvements to the standard CRISP-DM process, especially the ones considering the process management side, encompass life cycle selection, project iterations identification, allocation of resources, definition of data mining metrics, as well as step-by-step evaluations instead of a holistic approach. The authors also point out the lack of guidelines towards the form of process documentation, which could be as well modelled, having IEEE Std 1074 as a basis.

2.3 Semi-Automated Data Analysis Initiatives

Recent years marked a rise in the development of (semi-)automated initiatives of data identification and analysis, or even creation of natural language narratives for the results. Products, or prototypes, such as The Automatic Statistician, Data Science Machine, Autodiscovery (Butler Scientifics), or Wordsmith (Automated Insights) require raw data as an input, in order to turn it into information by the means of statistical analysis. The output varies, from identification of interesting correlations within the dataset, to automatically generated natural-language articles presenting the information in an easiest to ingest way for the reader. A simple feature comparison of these tools and prototypes is presented below, in Table 1.

	The Automatic Statistician	Data Science Machine	AutoDiscovery	Wordsmith
Structure of the Input Data	Multidimensional	Multidimensional	Tabular	Tabular

Feature Set Construction	Yes	Yes	No	No
Data Consolidation	Yes	Yes	Yes	No
Post-Analysis of Result Relevance	Yes	Yes	Yes	No
Natural Language Description of Results	Automatic	No	No	Automatic
Development Stage	Early Prototype	Early Prototype	Fully functional	Beta
Availability to Public	Not available	Not available	Free trial, paid full version	Free demo

Table 1: Feature comparison of automatic data analysis software

As presented in the table above, the current prototypes, as well as fully-functional programs, use different approaches towards data analysis, a noteworthy example is automatic feature set construction done by e.g. The Data Science Machine to explore the data set further. While most of the analyzed prototypes perform an “exclusiveness analysis” to determine the scientific relevance of the results, other pieces of software, like Wordsmith in that case, focus on creating articles out of raw data. Wordsmith does so by including a person, who is responsible for creating narratives, in the process. The remaining part of the article - result description - is done automatically. A similar procedure is applied by The Automatic Statistician. The generated output in that case consists of graphs depicting the relationships deemed interesting by the program, and their natural language descriptions.

The initiatives described above prove the potential usefulness of automatic data analyses, while the studies described earlier confirm the necessity of enhancing the CRISP-DM process. Specifically here - the business understanding phase of it, which is considered by numerous researchers to be one of the most important steps (or even the most crucial), and at the same time to be not detailed enough. It severely lacks guidelines, tools and standardization. As described by some of the aforementioned studies, even the (semi-)automation of this stage of CRISP-DM is possible with a standardized process model.

2.4 Natural Language Processing

Businesses store more and more textual data, either with an intention of using it for something data-mining-related, or without one. Here for example, we would like to mention all records of communication between employees, or between businesses. It is absolutely common to keep these transcripts indefinitely, or at least for a couple of years (Fisher, Brush, Gleave, & Smith, 2006). These e-mails, or instant messaging records are not stored in order to computationally process them, however they can hold enough information for this to be worthy of an effort, especially in context of identifying business needs and goals. This angle was explored considering the potential relevance of natural language processing to this project.

Natural Language Processing (also referred to as just NLP) is an area of research and application of theoretically motivated range of computational techniques for analyzing, understanding, and manipulating naturally occurring texts at one or more levels of linguistic analysis for the purpose of performing and achieving desired tasks with human-like language processing accuracy (Liddy, 2001).

This definition is broad for a number of reasons. The range of computational techniques is unspecified and, for a definition, quite imprecise, due to a vast number of methods and techniques which can be used to accomplish a desired type of language analysis. As for the “naturally occurring texts”, this signifies that the text for the analysis should not be constructed having the purposes of analysis specifically in mind. Quite the opposite, it should be constructed with the aim of communication between humans, in one of the languages used for that. These are virtually the only restrictions of that point, for instance the way of recording the text (e.g. audio recording or a written transcript) does not matter.

Humans interpret the meaning of text on multiple levels, Feldman (1999) writes about at least seven of them. As different NLP systems use these levels selectively, depending on their objectives, it is important to understand them, be able to distinguish between them, and know which ones need to be applied in which cases.

Phonetic/phonological level

This level is used only in spoken language and voice recognition systems, as it deals with interpreting the sounds, both within words and across them. This is done according to three types of rules, namely: (1) phonetic rules – used for interpreting the sounds within words; (2) phonemic rules – used to determine the sounds and their variation when the words are spoken together; and (3) prosodic rules – which deal with the stress and intonation within sentences.

Morphological level

In linguistics, a morpheme is a smallest fragment of a word, which contains meaning. An example of a morpheme would be the word "human", which can be a stem for words like humanity, humans, or inhumane. Compound words can be divided into multiple morphemes, e.g. the word transformation consists of a prefix "trans", a root "forma", and a suffix "tion". Morphemes keep their meaning across words, which makes it easy for a human (or in this case, an NLP system) to break down an unknown word into them, in order to deduct the meaning.

Lexical level

This level of text interpretation considers the individual meaning of particular words. Taking into account that a word may naturally function as more than one parts of speech, its use within a sentence is a crucial step before determining its meaning in this particular context (which is still not completely achieved on this level). Aside from that, NLP systems often use lexicons on this level. These lexicons range from simple ones, which match the words with their respective part-of-speech tags, to more complex ones, which may for example contain semantic information about the word, as well as context templates for different usage patterns of the word.

Syntactic level

In linguistics, syntax is a set of rules and processes, which govern the sentence structures of natural languages. Such attributes as word order can drastically change a meaning of a sentence. To provide an example the sentences: "Jenny turned on the lights" and "The lights turned Jenny on" differ only in terms of syntax, more particularly word order, but as a result they convey a completely different meaning. The NLP systems use grammar and a parser to analyze the syntax and establish dependencies across the words in sentences.

Semantic level

Semantic level of text interpretation considers the meaning of a sentence by focusing on interactions between the words within it. While a complex lexicon on lexical level can determine the meaning of a word by exploring the patterns in a sentence and comparing it to the pattern resources to identify the correct part-of-speech tag, it does not explore the semantic context to identify the meaning. For example in the sentence "The score is fifteen love", the word "love" – normally referring to a strong feeling of affection – means a score of zero in the game of tennis. This is where semantics plays a greater role in decrypting the meaning – the word "score" used in the same sentence can help on a semantic level, but not on the lexical.

Discourse level

The analyzed text normally would consist of something more than a sentence-long unit. The meaning is not only conveyed within the sentences, but also between them, as well as through the structure of the piece. Discourse level is used on the text as a whole, to add to the identified meaning. For instance, replacing the semantically empty pronouns with the entities they refer to is done on the discourse level. Another aid here would be text structure recognition, for example identifying the information within the abstract of a scientific paper, or within the headline and the first paragraph of a newspaper article, and then applying it to the rest of the text.

Pragmatic level

This level requires much world knowledge and is practically impossible to be fully implemented within an NLP system. Pragmatic level refers to understanding what the text conveys without this being encoded in said text. This creates a massive barrier in NLP, as fully implementing this level within a system would require that system having access to a complete body of knowledge of what we, as humans know about the world.

Current state of the art

The levels used in particular natural language processing projects vary, depending on the nature of the task, available resources etc. Most tools however use only the lower levels, since their purposes mostly do not require interpretation at higher levels. Aside from that, the lower levels deal with more rule-based units of analysis, such as sentences, words, and morphemes, which is definitely more feasible to implement than a tool working on a bigger unit, whose rules are often relative and ungoverned.

Common applications

Natural language processing has a number of applications, such as:

- Machine translation,
- Spelling correction,
- Grammar checking,
- Spam filtering,
- Information extraction,
- Automatic summarization,
- Text classification,
- Sentiment analysis,
- Speech recognition,
- Human-computer interaction,
- And many others.

The table below presents the most common interpretation levels these types of natural language processing systems use:

	Phonological	Morphological	Lexical	Syntactic	Semantic	Discourse	Pragmatic
Machine translation	optional	X	X	X	X	optional	-
Spelling correction	-	X	X	-	-	-	-
Grammar checking	-	X	X	X	optional	-	-
Spam filtering	-	X	X	X	optional	-	-
Information extraction	optional	X	X	X	optional	optional	-
Text classification	-	X	X	X	optional	optional	-
Sentiment analysis	-	X	X	X	optional	optional	-
Human-computer interaction	optional	X	X	X	X	optional	-

Table 2: Interpretation levels used by different types of natural language processing systems

As presented here, the most common applications of NLP use only the lower levels of interpretation, having the phonological level only when working on spoken text in addition to the written one. Semantic level gets mandatorily included when the text needs to be not only analyzed, but also responded to, in form of a translation (e.g. Google Translate) or an answer (e.g. Siri or other voice assistants). It can however also be applied to enhance precision of other tasks, such as text classification. The highest level, which is possible to implement is discourse. It is especially useful to have an NLP artefact work on discourse level, when it is often applied to larger chunks of text. For instance, machine translation can largely benefit from being able to interpret the text on discourse level, when translating full articles and being able to exchange the knowledge between sentences.

3 Setting Business Goals in Practice

To answer the second research sub-question (i.e. "RSQ2"), a questionnaire was created and deployed as a Google form. We promoted the questionnaire using inter alia LinkedIn groups and our public LinkedIn profiles.

DM Questionnaire

QUESTIONS RESPONSES 46

Does your organization have a specialized Data Mining division? *

Yes

No

Which methodology do you use the most often for data mining? *

CRISP-DM

SEMMA

KDD Process

Your own

Your organizations'

Other

Which of these activities do you perform at the early stages of the project? *

Assess background to the project

Review corporate documentation

Conduct interviews with stakeholders

Assess risks

Define the success criteria

Figure 3: Questionnaire as a Google Form

The response rate was initially low, however it picked up and reached 48 responses from various data mining professionals. 46 of these responses were detailed enough (answered more questions than just the ones about demographical details) to be usable in the research.

3.1 Questionnaire overview

The questionnaire consisted of 14 questions regarding the demographical data of the respondents, and the current state of data mining (or, specifically, business understanding)

practice in their organization. The questions, as well as the choice of answers, are listed in the table below:

	Question	Choice of answers	Explanation
1.	What is the size of the organization you work for?	<ul style="list-style-type: none"> • 1-10 employees (micro) • 11-50 employees (small) • 51-100 employees (medium) • 101-250 employees (medium) • 250+ employees (large) 	The purpose of this question was to be able to compare the answers of the respondents working for organizations of various sizes; for instance to investigate if problems with deploying the business understanding stage of CRISP-DM (or any other methodology) are more prevalent in smaller organizations.
2.	Does your organization collect data to support decision making?	<ul style="list-style-type: none"> • Yes • No 	The intention behind this question was to estimate the maturity level of the organization when it comes to data mining.
3.	Does your organization have a specialized Data Mining division?	<ul style="list-style-type: none"> • Yes • No 	The intention behind this question was to estimate the maturity level of the organization when it comes to data mining.
4.	Which methodology do you use the most often for data mining?	<ul style="list-style-type: none"> • CRISP-DM • SEMMA • KDD Process • Your own • Your organizations' • Other 	Most popular methodologies, and other options to choose from, which may turn out to be just as popular.
5.	Which of these activities do you perform at the early stages of the project?	<ul style="list-style-type: none"> • Assess background to the project • Review corporate documentation • Conduct interviews with stakeholders • Assess risks • Define the success criteria • Write a project plan • Other 	Steps of initial stages defined in the whitepapers/manuals of the most popular methodologies.
6.	How do you rate the feasibility of the activities from the	1 (Easy) - 5 (Difficult)	Likert scale, one to five, in order to determine the perceived feasibility of the initial stage of data mining.

	previous question under the chosen methodology?		
7.	Which of these activities prove to be most difficult, or time-consuming?	Open question	Key question of the survey, the results, if conclusive, should shape further stages of the research.
8.	Do you use any tools and / or services to help define the business goals of the project?	<ul style="list-style-type: none"> • Yes • No 	In case of a lot of respondents answering “yes” to this question, an investigation of the tools and services aiding with the projects would be necessary for the research.
9.	If you do, please mention which ones.	Open question	Specification of the tools and services needed to be investigated.
10.	Do you think some steps of the initial phase (i.e. Business Understanding in case of CRISP-DM) can be (semi-) automatized? If yes, which ones?	Open question	Collecting suggestions from the professionals working in the field.
11.	At which point do you know that you know enough to define the business goals of the project?	Open question	A so-called “tipping point”, which would signify the end of phase one of the project.
12.	How would you envision the steps which should be made to define the data mining business goals at your organization?	Open question	Suggestions about the possible changes to the business understanding phase, when applied to a specific organization.
13.	Do you think any of these steps could be performed (semi-) automatically? If yes, please mention which ones.	Open question	Collecting suggestions from the professionals working in the field.

14.	Do you know of any tools and / or services, which could help with defining the data mining business goals? If yes, please mention them by name.	Open question	Specification of the tools and services needed to be investigated.
-----	---	---------------	--

Table 3: Questions and choice of answers in the questionnaire

The combined first three questions of the survey were written with intention of classifying the organization by their size and their maturity when it comes to data mining. Further, a more specific set of questions about the processes and techniques used at respondents' organizations are used (questions 4 and 5). This is done both to assess the popularity of particular processes, as well as to see if they are followed by the book.

Questions 6 and 7 deal with the problems within the business understanding phase, as perceived by professionals using it in the field. Especially question 7 is vital to the next stages of the research, because the result – if conclusive – should shape up further work on the design of the artefact, which would actually be helpful when deployed alongside CRISP-DM.

Questions asked in points 8, 9 and 14 were set to investigate if any tools and/or services are prominently used to aid with setting business goals of the project. In case of a high number of positive answers, these tools are to be subject of the research, with the intention of determining their offers, as well as what they lack and where they could be improved.

The tenth question of the survey is set to collect suggestions about possible automatization, which could be done to generally aid with CRISP-DM's initial phase. Questions 11, 12, and 13 are more organization-specific, asking about personal experiences with setting business goals at the respondents' businesses, the end-point to business understanding, as well as suggestions for possible automatization when it comes to the exact data mining processes used within said organizations.

3.2 Background data of the respondents' organizations

The only background data that was collected, referred to the organizations the respondents were working for. As mentioned above, the combined first three questions of the survey were written with intention of classifying the organization by their size and their maturity when it comes to data mining. The percentages of the respondents working for organizations of different sizes look as follows:

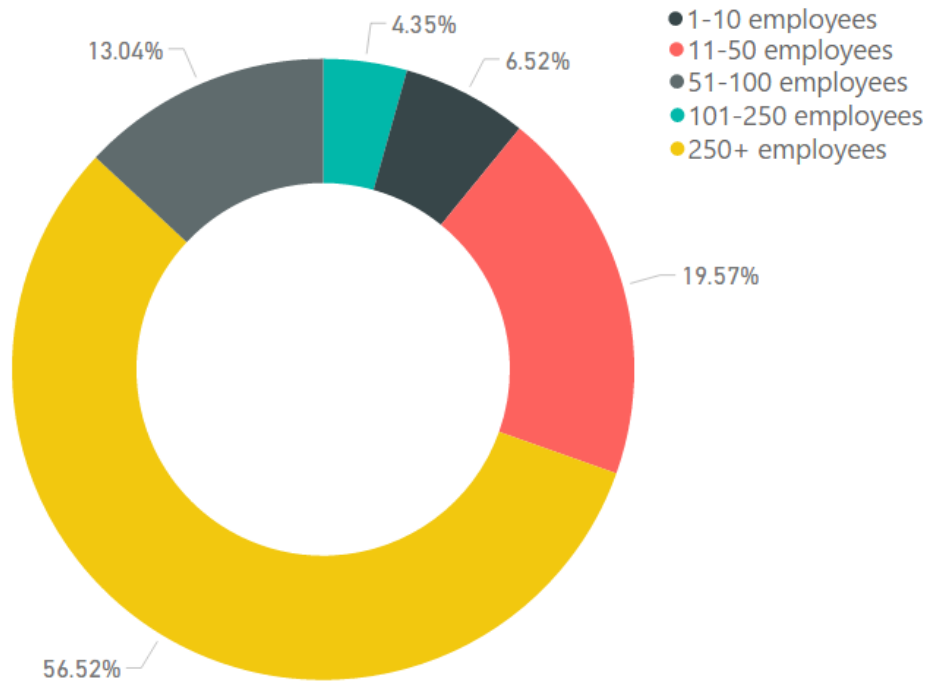


Figure 4: Results of the first question of the survey (What is the size of the organization you work for?)

Most of the respondents work for large organizations, employing more than 250 people. Since the subject of the thesis concerns small and medium organizations, this presented an opportunity of comparing the state of practice of data mining between the bigger organizations, and the smaller ones, which, for this purpose, due to the number of responses, were grouped into one cluster – businesses employing less than 250 people.

93.5 percent of respondents are employed by businesses, which collect data for further use, however only 51.2 percent of these businesses have a specialized data mining division. Since the answers provided by people working for organizations which did not collect any data lack substance when it comes to personal experience (all of them failed to answer questions 3 to 11), the following groups were created to compare the business understanding phases of data mining under different organizational conditions:

- Group 1 (more than 250 employees, specialized data mining division),
- Group 2 (more than 250 employees, no specialized data mining division),
- Group 3 (fewer than 250 employees, specialized data mining division),
- Group 4 (fewer than 250 employees, no specialized data mining division),

3.3 Methods and processes used

● CRISP-DM ● Other ● Your organizations' ● Your own

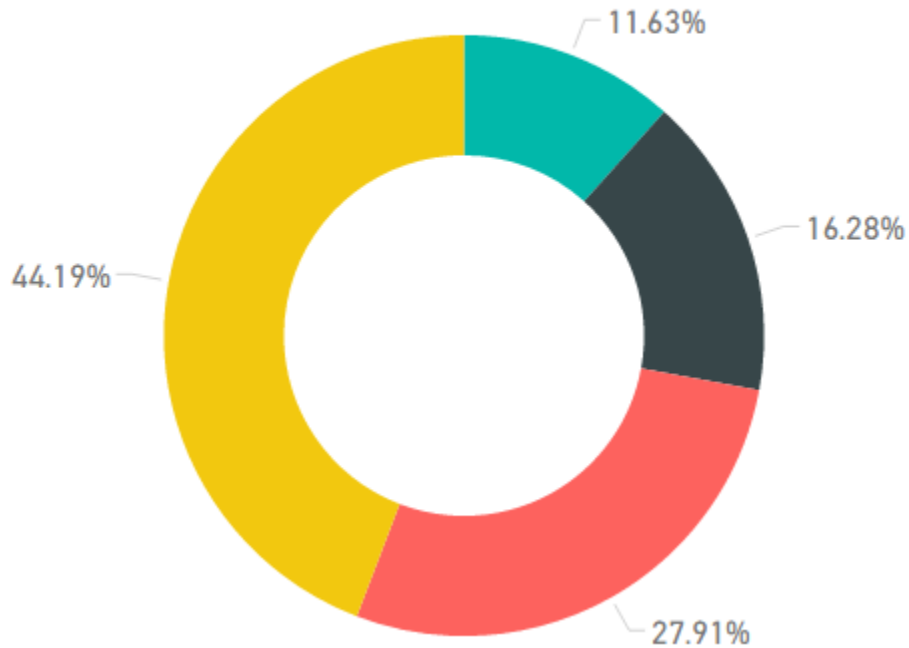


Figure 5: Respondents' choice of DM processes

The majority of respondents (72.1%) use non-standard methods, devised either by them, or by the organization they work for. Out of the standard methodologies, which have been considered to be most prominent, only CRISP-DM has a significant following at 11.63%. No respondents picked SEMMA, or Knowledge Discovery in Databases Process. There is no significant variation between four groups.

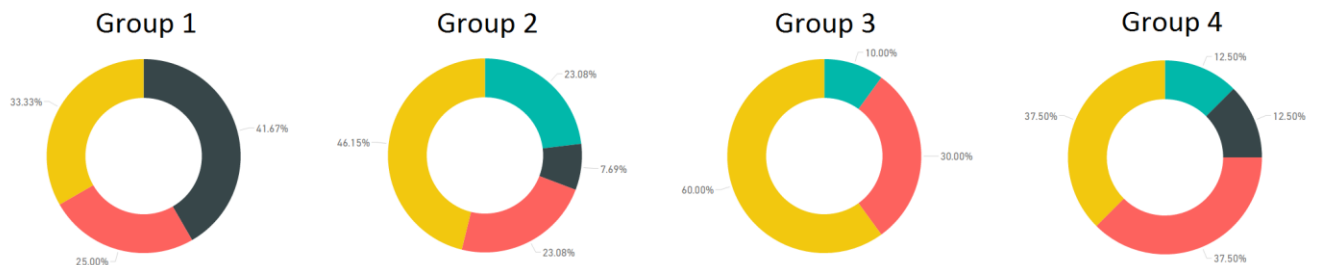


Figure 6: DM processes divided by groups

3.3 Business understanding and its problems in practice

As for the most prominent activities during the early stage of their data mining undertaking, the participants of the survey identified the following:

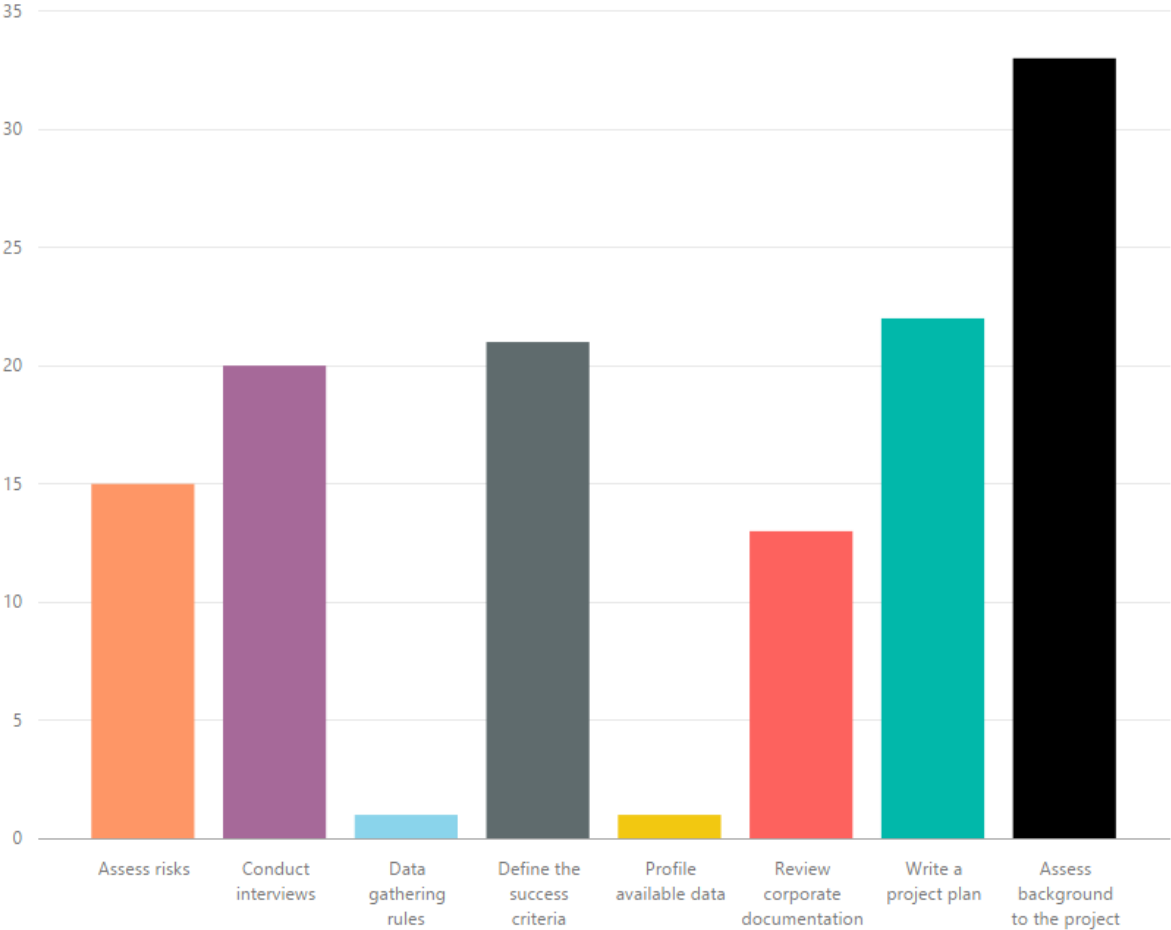


Figure 7: Most difficult steps of business understanding phase, as answered by the professionals

These activities can be further divided into managerial tasks (assess risks, consult data gathering rules & regulations, define success criteria, profile available data, write a project plan), and content-decisive tasks (conduct interviews with stakeholders, review corporate documentation, assess background to the project).

The next questions referred to the feasibility of these tasks. Overwhelming majority rated them at 3 points out of 5 on the difficulty scale, with no significant difference between people working for larger and smaller businesses.

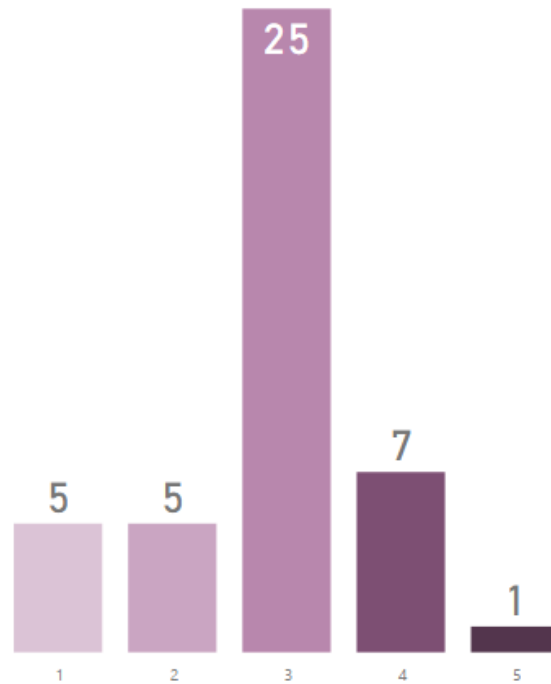


Figure 8: Difficulty of the BU tasks, as assessed by the respondents

When asked about the most difficult of these tasks, 16% of the respondents chose interviews with stakeholders, 14% picked reviewing corporate documentation (or, as some of them put it, “non-existent corporate documentation”), while another 14% chose assessing background to the project. Another set of respondents identified preparation and integration of sources from different platforms as the most problematic endeavor.

3.4 Tools used at initial stages of data mining

Only 25.6% of the respondents admit to using tools/services to help with the initial stages of data mining, however this number grows to 38.5% in second group (large enterprises with no data mining division). The following tools have been identified:

- KJ method (in the responses also referred to as “KJ analysis” and “affinity diagrams”),
- Excel,
- Flowcharts,
- SMART goals,
- Corporate KPIs,
- SAS,
- SDLC,
- In-house tools.

3.5 Conclusions of the survey

In the end, the survey provided some expected answers, but at the same time it did not confirm the discrepancy between small / medium enterprises, and large ones, when it comes to the difficulties they have while setting up goals for data mining. According to the results, these difficulties are widespread in both types of businesses.

What has been confirmed, on the other hand, is the apparent lack of tools and services helping with this type of efforts. If the professionals admit to use anything like that, it is usually Excel (which is not really a tool specific for this type of projects), KJ method (a method for structuring the interviews, not a tool), or an "in-house tool", which would suggest that due to no widely recognized standards, the businesses had to develop their own ways of working around the problem.

The identification of the most problematic parts of the business understanding step of the respondents' data mining endeavors suggests that it's three crucial tasks, as well as their integration, which need to be bootstrapped. The content-decisive tasks, as we called them earlier, are the ones which data mining professionals would like to see standardized, or at least integrated somehow.

Multiple respondents' choice of affinity diagrams/KJ method as their preferred way of conducting interviews also suggests that there is a basis, which could be further developed to create an industry standard in identifying business goals for data mining projects.

4 Proposed Solutions

Through examining the results of the survey, we have pinpointed the most problematic steps of business understanding to: interviews with stakeholders, review of corporate documentation, assessment of the project background, as well as lack of any platform for integrating information or suggestions from different sources. The ways the respondents of the survey deal with these problems generally concern deploying various management strategies, such as SMART goals, or affinity diagrams.

They do not however agree on using any specific tools available on the market (in some cases, they admit to using in-house tools for this stage of the project). This would suggest a need of creating one, which could help standardize the platform for identifying business goals, while applying as much of the suggested strategies as possible, and at the same time trying to solve the other, yet unresolved issues.

4.1 Standardizing the interviews with stakeholders

The management strategies suggested by the respondents are applicable for standardization of this step of the business understanding stage of CRISP-DM. The interviews with stakeholders normally conducted during this phase can, and should be structured well enough to be replicable. Replicable in the sense of having the stakeholders understand the issue well enough to suggest the same ideas, no matter who is conducting the interview. To achieve that, they should be familiar with the SMART goal setting technique before taking part in the interview.

4.1.1 SMART Goals

To aid the stakeholders with setting data mining goals, we propose using the technique, most commonly referenced as SMART (Shahin & Mahbod, 2007). The acronym stands for the requirements the goals need to meet, in order to be of desired quality. These are respectively: Specific, Measurable, Attainable, Realistic, and Time-sensitive. Short summary of practical implications of using these attributes is presented below:

- Specific - the goals need not to be vague, or too broad, they should also be as detailed as it is possible. Otherwise misunderstandings may occur, and the project may fail with no one accountable for it.
- Measurable - the measurement can be qualitative or quantitative in nature, however it needs to be clear enough to determine if the objectives have been achieved.
- Attainable - the goals need to be within the reach of the organization, otherwise they would be hardly ever met.

- Realistic - it is possible that a goal can be attainable, but not realistic within the working environment. The availability of resources for the project needs to be a factor in the goal-picking stage, so that it is known in case the goals are not realistic.
- Time-sensitive - a time frame for completion of the goal is helpful in monitoring the progress of the project and intermediate success measurements, as well as setting a realistic action plan.

4.1.2 Affinity Diagramming and the KJ Method

The interview itself would take place with the use of affinity diagrams, which help having all the stakeholders involved in both suggestion, and evaluation of the business goals. As source integration was also identified as an issue, the affinity diagrams, or - more specifically - their evaluation stage, would apply to the next step of the process as well, thus creating a unified platform for the definition of business goals. More on this unification can be found in section 4.3. of this document, titled "Integrated tool for specification of the project goals".

Affinity diagrams are generally used to make sense of, and organize large volumes of unstructured, dissimilar quantitative data. (Lucero, 2015). The KJ method, proposed by a Japanese anthropologist Jiro Kawakita (Scupin, 1997) is a first source, on which the affinity diagrams were based. It consists of four steps:

- label making,
- label grouping,
- chart making,
- explanation.

The first step consists of participants writing down their ideas on separate pieces of paper/post-it notes. Afterwards, in the second step, the ideas are grouped into clusters. The so-called lone wolves, are set aside and left for later use. Then, a chart is created and annotated to describe the relationships and dependencies between the groups. Finally, a verbal or written explanation is devised. The affinity diagrams add another step at the end, which is evaluation of the resulting clusters and groups.

Since this method is considered to be a powerful tool for collaborative data preparation and analysis (Spool, 2004), it fits well for adding a general structure to the interviews. To add to that, it enables all the participants to take part not only in the idea generation, but also in clustering of similar ideas, and their further evaluation. This evaluation could be extended to account for the automatically generated data, which is further explained in the next section of this document.

4.2 Bootstrapping the review of corporate documentation & assessment of the project background

This is where natural language processing comes in. Generally, the corporate data is not structured, and can include massive amounts of text from different sources. An example of a source would be official organizational documents, which may describe the background to the project by providing information about the business. A different type of corporate documentation would be all the written information exchanged between the employees. Corporate e-mail repository is a valuable source of expertise, with an easy-to-mine set of communications between people in the social network (Campbell, Maglio, Cozzi, & Dom, 2003). Merali & Davies (2001) in their paper *Knowledge capture and utilization in virtual communities* explore the so-called “weak ties” in businesses, stating that these semi-anonymous connections between people, who communicate through media such as e-mail, are crucial to the flow of knowledge within organizations. Grobelnik, Mladenec, & Fortuna (2009) mention how knowledge management within a business can be supported by capturing the employees’ communication records and subjecting them to further analysis when needed. This type of analysis should provide some insights in case of the definition of business goals for data mining efforts. These business goals might not be immediately thought of during the interviews, but were once considered by people working for the organization.

This is the reason why we propose a natural language processing analysis of all the potentially relevant textual data found within the organization, which would search for once written business goals, thus bootstrapping the business understanding stage of data mining endeavors. To conduct an experiment with this type of data, we decided to produce a textual formula for an explicit business goal, and then analyze sample texts at this angle, in order to extract the business goals without the need of having to manually go through hundreds of pages of text.

Academic literature was carefully reviewed to find out if any study has ever been conducted on the default syntax and structure of business goals. This yielded a set of results, which were further used for the creation of the artefact. Casagrande, Woldeamlak, Woon, Zeineldin, & Svetinovic (2014) in their paper “NLP-KAOS for Systems Goal Elicitation: Smart Metering System Case Study” writes about automated goal elicitation from research publications by integrating data mining and natural language processing techniques with the (Knowledge Acquisition in Automated Specification) goal-oriented requirements engineering method. His technique of goal extraction from text is highly relevant to this project and was partially adapted for its purposes. As suggested by Van Lamsweerde (2009), the process is based on so-called goal-specific keywords. The table below shows goal keywords, as identified by Casagrande and Lamsweerde, divided into two types:

Amelioration

- Improve
- Increase
- Decrease
- Reduce
- Enhance
- Enable
- Support
- Provide
- Make

Intentional

- Objective
- Aim
- Purpose
- Achieve
- Maintain
- Avoid
- Ensure
- Guarantee
- Want
- Wish
- Motivate

This list is further expanded to contain various forms of these keywords, such as "aims", "aiming", etc. for the word "aim". The author realizes that the goals may be hidden well enough within the natural language that no natural language processing system would be able to identify with perfect recall. His argument for his approach is that from a statistical perspective *using simple common patterns and applying to a large amount of textual data, while some goals will be lost, the most relevant goals will appear often in the data*. Thus a formula for identifying the goals, first by keywords, and then through natural language processing is created.

The triplet:

< predicate – object – prepositional phrase >

(With only one of object/prepositional phrase necessary to complete extraction of the phrase) is considered as a basic formula for goal identification. The predicate is a verb within the phrase, part of which matches the keyword. Object and prepositional phrase are extracted by exploring the subtree of the sentence and returning the deepest noun or adjective as an object and the following phrase matching a PP tag as prepositional phrase.

Another group of researches, working on a similar subject, Lezcano, Guzmán, & Alonso Gómez (2015) in their paper *Extraction of goals and their classification in the KAOS model using natural language processing* suggest a similar approach to automatic goal identification. It is however important to note, that this research was conducted in Spanish, using Spanish translations of the words below. They divided the goals into four types, namely:

To achieve:

- Form
- Improve
- Increase
- Promote
- Register
- Develop
- Formulate
- Make
- Insert
- Prepare
- Reduce

To maintain:

- Administer
- Guarantee
- Save
- Offer
- Prolong
- Endorse
- Manage
- Obtain
- Keep
- Know

To cease:

- Stop
- Finish
- End
- Interrupt
- Cease

To avoid:

- Avoid
- Block
- Prevent

They also consider modal periphrases as keywords identifying possible goals. The sentences with all these keywords are extracted from the text, and a natural language processing method is applied to them to determine if the sentence contains a goal, based on the following formulas:

noun phrase + modal periphrasis + verb + complement

verb phrase + 'that' + verb + complement

('that') + noun phrase + verb + complement

While the Spanish verbs had their English versions included in the research, these formulas were only applicable to Spanish, without any English versions. As some of the constructions might not be directly translatable, I have decided to test them in English, and only use the ones yielding positive results.

To translate the notions within these formulas into Casagrande's terms, the verb is a predicate, and the complement stands for the object and the prepositional phrase. As it is visible in here, Lezcano et al. added another element to the formula, which is the noun phrase, or the subject of the sentence. Casagrande however argues that the subject is not needed for automatic goal extraction, as it can be often implied, while also there are cases where it is not explicit in the sentence, e.g. defined in the previous sentence or substituted by pronouns. In that case, the NLP tool would require anaphora resolution, of a discourse level of text interpretation, to function. In case of a business goal, such as "The X department should increase their productivity by 20%", Casagrande's approach would yield:

<increase – productivity – by 20%>,

While Lezcano's et al. approach:

X department + should + increase + their productivity by 20%.

However, considering the semi-formal tone of some text documents, common use of bullet points etc., the noun phrase/subject might not always be present in the sentence, in which case Casagrande's approach would still give results, while Lezcano's et al. would not. A possible solution to this problem would be merging both techniques, but making the subject optional. The identification under these rules would be based on identifying either a pair <predicate – object>, a pair <predicate – prepositional phrase>, or a triplet <predicate – object – prepositional phrase>, but the extracted result would be of a form:

(subject) predicate (object) (prepositional phrase*)*

With the subject extracted only if found, and only one of object and prepositional phrase obligatory.

4.3 Integrated solution for specifying project goals

The combination of methods from sections 4.1 and 4.2 was planned as an integrated tool, which:

- collects all the ideas from interviews with stakeholders,
- enables the participants to cluster their ideas, to avoid unnecessary duplications,
- analyzes the provided textual data, searching for once written business goals,
- enables the participants to evaluate all the findings, both proposed by them and found within the documentation,
- lets the participants discard false positives from the analysis, as well as ill-fitting suggestions from the interviews,
- ranks the results based on the evaluations,
- presents the evaluated results ranked.

The development of the tool, and its further evaluation are described respectively in sections 5 and 6 of this document.

4.4 Success Criteria

Success criteria in IS/IT projects can be objective or subjective (Chan, Scott, & Lam, 2002) and concern a number of themes, presented in the table below:

Objective measures	Subjective measures
Time	Quality
Cost	Meeting technical performance specification
Financial performance	Goal attainment
Profitability	Completion
Health and safety	Functionality
	Efficiency
	Stakeholder satisfaction
	Stakeholder expectations
	Dispute resolution
	Absence of conflicts
	Professional image
	Aesthetics
	Professional aspects
	Environmental sustainability

Table 4: Types of success criteria for IT/IS projects

As this is an exploratory project within the frames of this research, a much thinner range of these themes applies to setting its success criteria. These are highlighted in yellow within the table. As the themes are taken from multiple papers analyzed by Chan et al., they may overlap and the particular success criteria may consider multiple themes at ones. The exact specification of success criteria for the tool is given below:

1. The tool allows users to input their business goals.
2. The tool analyzes given textual data and automatically retrieves business goals.
3. The users are able to rate the goals.
4. The tool presents a list of goals along with their ratings.
5. The tool eliminates goals with unsatisfactory ratings.
6. The tool is complete and functional.
7. The tool's information retrieval component achieves high precision and recall scores.
8. The tool is applicable within CRISP-DM's BU context.

5 Development of the tool aiding with business goal specification

All the ideas gathered from the research have culminated in the design of the method for identification of business goals, and the tool based on it. The method takes into account the official CRISP-DM guide, more specifically its section on business understanding, and adds the results of the research to that. The tool, currently unnamed, aids with the issues of:

- Text analysis,
- Integrating the results of interviews with text analysis,
- Evaluation of these two types of results.

5.1 Method

Since up to now, no unified method of dealing with business understanding has been developed, I have collected the suggested steps from the CRISP-DM whitepaper in order to present how they can work in parallel with the developed tool, and where exactly the tool should be applied.

The crucial steps of *business understanding* listed in the CRISP-DM guide are the following:

Determining Business Objectives:

- Compiling the Business Background
- Defining Business Objectives
- Business Success Criteria

Assessing the Situation:

- Resource Inventory
- Requirements, Assumptions, and Constraints
- Risks and Contingencies
- Cost/Benefit Analysis

Determining Data Mining Goals:

- Data Mining Goals
- Data Mining Success Criteria

Producing a Project Plan:

- Writing a Project Plan
- Assessing Tools and Techniques

This method and tool developed during this research concern the initial step, which is the determination of business objectives. As presented in the survey results section, to determine business goals, the professionals working in the field perform the following steps:

- Conduct interviews,
- Review corporate documentation,
- Assess background to the project.

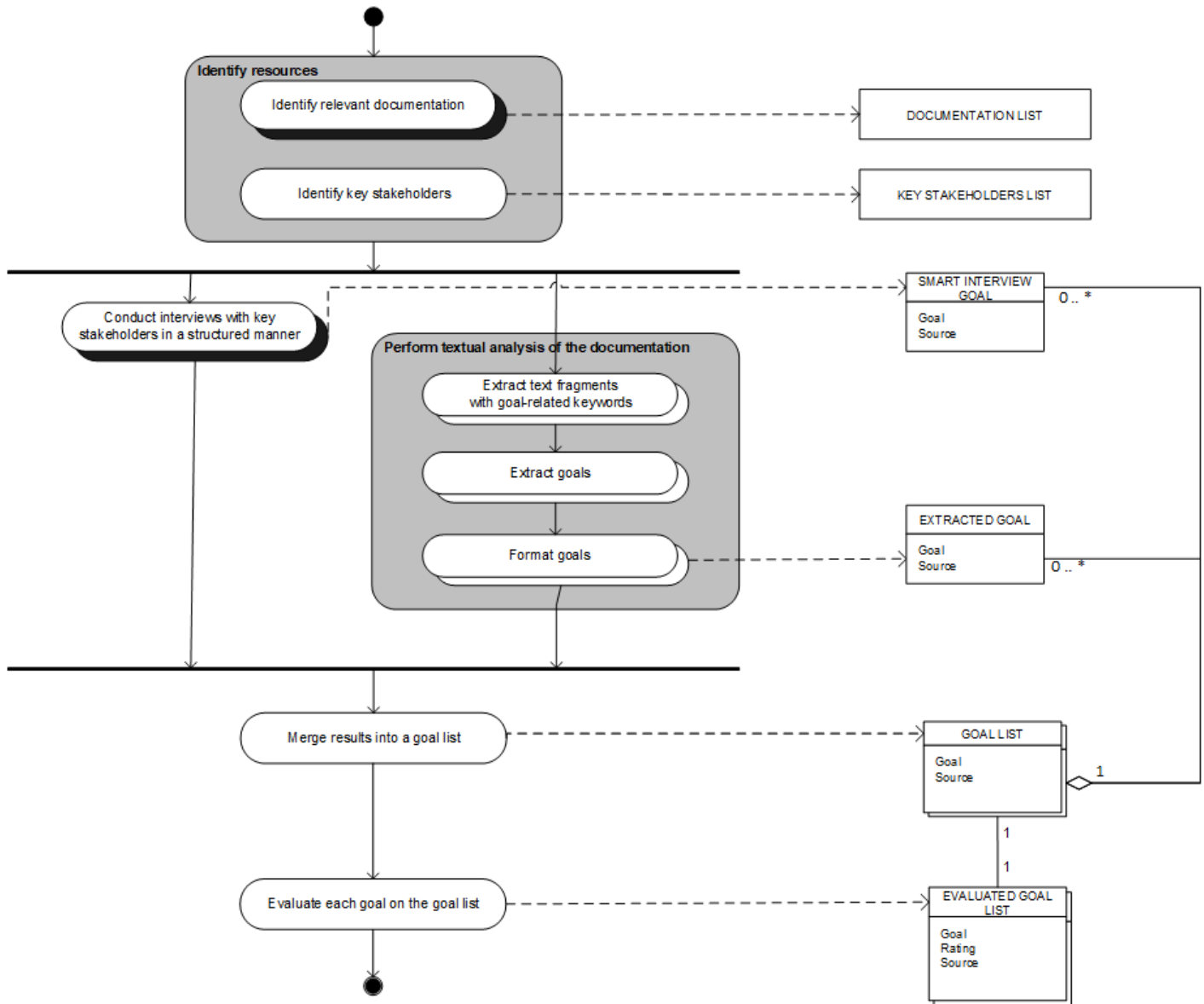


Figure 9: PDD of the proposed method

These at the same time being the steps, which they rank as the hardest to perform during the business understanding phase of the project. At the same time, the integration of results from different sources seems to be equally as problematic. To aid with this issue, we propose a method of determining business objectives, on which the further developed artefact is based. The method

is presented in the figure above, as a high-level process-deliverable diagram (van de Weerd, & Brinkkemper, 2009). It is further expanded to account for the details of the open activities in the next sub-sections of this document. The constituent elements of a process-deliverable diagram, and their explanations are attached in Appendix B.

The first step consists of two unordered activities (possible to be conducted in any sequence) relating to resource identification, namely identification of relevant documentation, and identification of key stakeholders. These resources are used in the two following steps after forking: the key stakeholders are interviewed in a structured manner to produce a list of SMART goals, while the documentation is subjected to textual analysis and information extraction. The text fragments containing relevant, goal-related keywords are drawn out, and through natural language processing, the goals are identified and extracted. After formatting they are placed in a list of "extracted goals".

5.2 Tool

The primary objective of the tool is natural language processing of textual data. Aside from that, it was designed to aid with the remaining parts of the method, namely saving the results of the goal identification through the interviews, subsequent merging of two sets of results, and their further evaluation.

The tool lets the participants pick a container where their textual data is stored. If the folder contains subfolders, the contents of all of them are taken into account when analyzing the data. When a container is selected, the tool identifies all files, which are composed of textual data. The first algorithm, "Goal Identifier", is applied to these files. It extracts the sentences, which contain the following keywords:

- Objective
- Aim
- Purpose
- Goal
- Achieve
- Require
- Lack
- Maintain
- Ensure
- Guarantee
- Motivate
- Improve
- Increase
- Decrease
- Reduce

- Enhance
- Enable
- Support
- Provide
- Make
- Need
- Must
- Ought
- Should
- Wish
- Want

This list was based on approaches analyzed in section 4 of this document; the presence of its constituents within a sentence suggests that a goal is introduced there. The algorithm extracts full sentences containing the keywords, by the use of regular expression, and saves them to another file. This is done to reduce the computation load, as parsing whole documents is costly and time-consuming. This way, only the potentially relevant sentences are subject to parsing.

The output is passed on to another component, "NLP Analysis". This component analyzes the sentences natural language processing. If the keyword is a verb, this verb is selected as the predicate. The algorithm then looks for the highest verb phrase tree with this keyword, and analyses its children and siblings, to find the complement (object and prepositional phrase), as well as the optional subject (noun phrase). In case of the keyword being a noun, the highest noun phrase tree is identified, and the sibling, which is a verb tree is selected to find the predicate. The complement is identified within this highest verb tree. The goals are saved to memory in the format specified in the previous chapter,

{(subject) predicate complement}

where only the predicate and the complement are obligatory.

In the meantime, the participants are able to input their ideas for goals. Depending on the structure of the interviews, if they are guided or not, this can be done by one person, interviewing the stakeholders, or an unlimited number of the stakeholders themselves. The second option should be picked only if the participants are well-acquainted with the process, and able to provide a SMART set of business goals. The results are then saved to the memory.

When this stage is completed, all sets of results (text analysis, as well as results from all the stakeholders) are merged into one and their order is randomized. This is done in preparation for the evaluation stage of the process. All the evaluators (could be the same number as interviewees from the previous step, or any other number, depending on the situation) are shown the results one by one and prompted to rate their usefulness. The scale ranges from 0 to 10, where 0 stands for irrelevant goal/not understandable/spam.

When everyone completes the evaluation, the points assigned to particular business goals are summed up. If a result does not receive more than 0 points, it is deemed irrelevant and discarded. The remaining ones are ranked and the resulting list is presented to the participants.

5.2.1 Technologies Used

The artefact was developed in Java programming language, using Eclipse JEE Neon as the integrated development environment. Aside from using common Java libraries, such as *swing*, *io*, or *awt*, other libraries were used to conduct regular expression pattern matching, and natural language processing of the results.

As we previously mentioned, not all of the sentences are parsed. To reduce the memory load, the keyword-containing sentences are first identified and saved, and only then passed to the parser. To extract these sentences, the tool makes use of pattern matching. The pattern is written in a form of a regular expression, which matches entire sentences containing the stem of a keyword. This is done through the use of *java.util.regex.Matcher* and *java.util.regex.Pattern* libraries.

When the sentences are extracted, another component is called. This component makes use of natural language processing libraries. After testing a couple of NLP application programming interfaces, Stanford CoreNLP Natural Language Processing Toolkit was chosen as the most fitting one.

Stanford CoreNLP is an open-source Java annotation pipeline framework, which provides most of the common core natural language processing steps (Manning, Surdeanu, Bauer, Finkel, Bethard, & McClosky, 2014), as presented on figure 10.

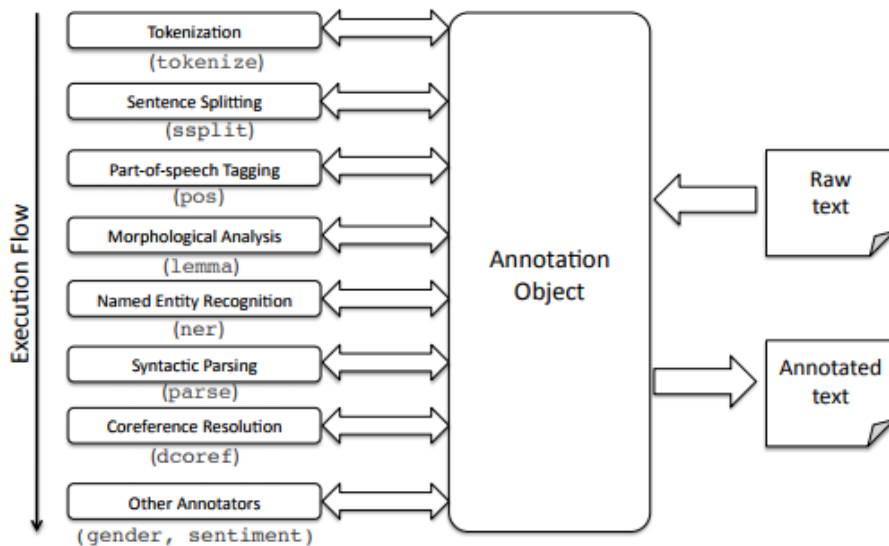


Figure 9: Stanford CoreNLP system architecture, taken from Manning et al., 2014

Picking the annotators in a pipeline, which are necessary for a project, is done through the basic Java properties in a Properties object. This object is also responsible for specifying the order in which the annotators are applied. The object is further passed to a pipeline, and can be applied to a list of sentences.

The artefact of this project uses a number of these annotators, namely:

- A Penn Tree Bank **tokenizer**, which tokenizes the text into a sequence of tokens. It has been developed with web text in mind, thus it deals quite well with noise, which can be found in electronic communication between people.
- A **sentence splitter**, which divides a sequence of tokens into sentences, thus preparing them for further analysis.
- A maximum entropy **part-of-speech tagger** (Toutanova, Klein, Manning, & Singer, 2003), which labels the tokens with their corresponding part-of-speech tags.
- A **parser**, which provides a full syntactic analysis of the input text. The dependency trees are developed in this stage of analysis (Klein and Manning, 2003; De Marneffe, MacCartney, & Manning, 2006)

After applying these annotators, the output text is subject to further analysis, in order to determine which sentences (or their parts) should be recorded as goals, and which can be discarded. The *edu.stanford.nlp.trees.** libraries are used to traverse the dependency trees of each sentence, and record the meaningful information.

The tags, originally developed for the Penn Treebank Project, which are applied to the parts of sentences are listed in the table below:

Part-of-speech Tags	
Tag	Description
CC	conjunction, coordinating
CD	cardinal number
DT	Determiner
EX	Existential "there"
FW	Foreign word
IN	Conjunction, subordinating, or preposition
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Verb, modal auxiliary
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Noun, proper singular
NNPS	Noun, proper plural
PDT	Predeterminer

POS	Possessive ending
PRP	Pronoun, personal
PRP\$	Pronoun, possessive
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Adverb, particle
SYM	Symbol
TO	Infinitival to
UH	Interjection
VB	Verb, base form
VBZ	Verb, 3 rd person singular, present
VBP	Verb, non-3 rd person singular, present
VBD	Verb, past tense
VCN	Verb, past participle
VBG	Verb, gerund or present participle
WDT	Wh-determiner (e.g. which)
WP	Wh-pronoun, personal (e.g. who)
WP\$	Wh-pronoun, possessive (e.g. whose)
WRB	Wh-adverb (e.g. when)
.	Punctuation mark, sentence closer
,	Punctuation mark, comma
:	Punctuation mark, colon
(Contextual separator, left paren
)	Contextual separator, right paren
Chunk Tags	
NP	Noun phrase
PP	Prepositional phrase
VP	Verb phrase
ADVP	Adverb phrase
ADJP	Adjective phrase
SBAR	Subordinating conjunction
PRT	Particle
INTJ	Interjection
PNP	Prepositional noun phrase

Table 5: Penn Treebank tags (taken from Part-of-Speech Tagging Guidelines For The Penn Treebank Project, Beatrice Santorini, 1990)

Tregex (Levy & Andrew, 2006), as one of edu.stanford.nlp.trees.* sub-libraries proves to be extremely useful, when extracting sentences, based on their tags. It allows for identification and extraction of trees matching specified patterns. For instance "NP < NN \$ VP" identifies a noun phrase, which is a parent of a singular form of a noun, while at the same time being a sister to a verb phrase. These patterns can get much more complicated than that, and extract much more specific content.

To develop the tool, a sample set of textual corporate data was needed. Due to privacy concerns and legal restrictions, sets like that are not easy to obtain. For this reason, the so-called Enron Corpus (Klimt & Yang, 2004) was used. Enron Corporation was an American energy company, based in Houston, which bankrupted on December 2nd, 2001 in the wake of accounting fraud scandal. The data set of 619,446 e-mails sent and received by 158 Enron employees was collected in 2002 during the investigation into company's collapse, as commissioned by Federal Energy Regulatory Commission. It was later purchased and released to researchers, being the first publicly available mass collection of corporate e-mails.

The corpus is often used by researchers in natural language processing, as well as other fields. For the purposes of this research, in this case the development of the artefact, parts of the data set were used to intermediately analyze the correctness of the algorithms, as well as assess the precision and recall of the tool.

5.2.2 Explanation of the key components

The components of the tool, which provide its key functionality for goal identification, are sentence extraction, natural language processing analysis and tree pattern matcher. This subsection of this document will briefly introduce these components, and explain the code behind them, at the same time matching them to the open activities of previously introduced process-deliverable diagram.

Sentence Extraction Component

Using *java.util.regex.Pattern*, we can define a regular expression, to which the analyzed text will be compared. As the idea behind the extraction was to select entire sentences containing the keywords, the following regular expression provides the pattern for that:

```
[^.!?]*stem_of_the_keyword[^.!?]*[.!?]
```

The first bracket, with a circumflex inside, indicates a negated set, an asterisk at the end of it means that it can repeat 0 or more times. In this case, `[^.!?]*` means, that the matching starts with an unlimited number of signs, which are not a dot, not an exclamation point, and not a question mark. After this unlimited number of signs, the pattern introduces the stems of our keywords, followed by another set of 0 or more signs, which do not signify the end of the sentence. The pattern concludes with `[.!?]`, which is an expression matching one of: a dot, an exclamation point, or a question mark (which mean that the sentence ends there). Essentially, matching against this pattern will result in strings, which are limited by a dot, a question mark, or an exclamation point both at the beginning and at the end, and which contain the stem of the keyword.

The pattern used within the code looks as follows:

```
( "[^.!?]* (objective|aim|purpose|goal|achiev|requir|lack|maintain|ensur|guarantee|motiv|improv|increas|decreas|reduc|enhanc|enabl|support|provid|make|need|must|ought|should) [^.!?]* [.!?]", Pattern.CASE_INSENSITIVE)
```

The pattern is then applied to the text within the *BufferedReader*, by using *java.util.regex.Matcher*. If a match is found, it is recorded in a set, and when all the matches are retrieved, the set is saved to a text file, which is later analyzed by the NLP component of the tool.

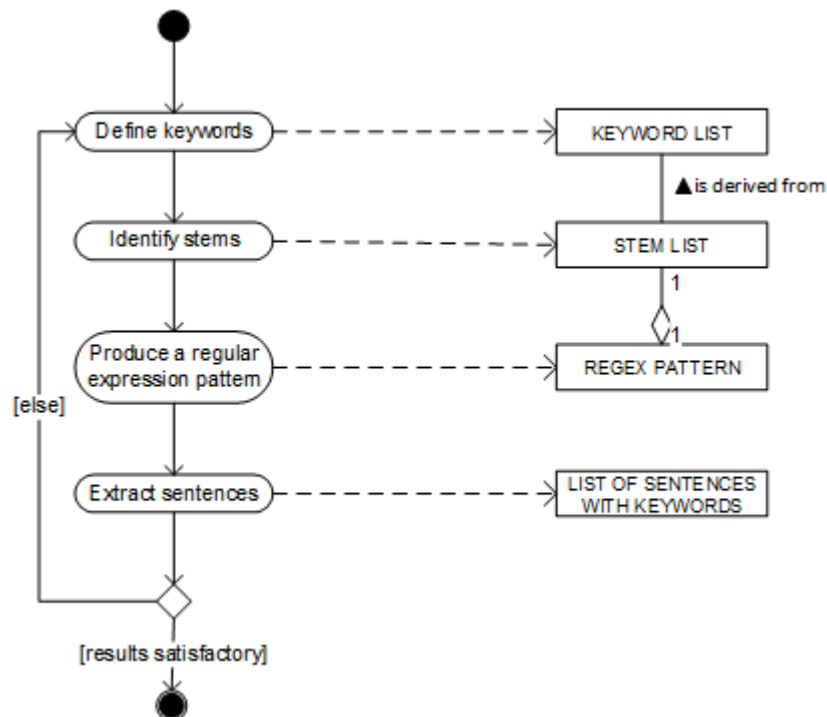


Figure 10: PDD of extracting sentence fragments with goal-related keywords

Natural Language Processing Component

The NLP component of the tool receives the text file passed to it by the sentence extraction component and reads it thanks to the *java.io.FileReader* library. A Stanford NLP pipeline with annotator properties: *tokenize, ssplit, pos, parse* is created and a *CoreMap* of sentences saves the annotations. To facilitate the further extraction of matches, each of the sentences is converted to a Penn Treebank tree, and the matchers are applied. There are three different matchers, set to find goals of three types of structures, as defined earlier.

The first matcher, based on the pattern:

```
VP << (VBZ|VB|VBP|VBG|VBG|VBD < [verb keywords]) ?$ PP ?$ NP
```

Looks for a verb phrase, which dominates a keyword-verb in any form, at the same time extracting the adjacent prepositional and noun phrases (if they exist).

The second matcher, based on the pattern:

```
NP [ << (NN < [noun keywords]) | << (NNS < [noun keywords plural]) ] ?$ PP ?$ VP !> NP
```

Looks for a noun phrase, which dominates a keyword-noun in any form, while at the same time extracting the adjacent noun and prepositional phrases. It must not be directly dominated by another noun phrase (as the first relationship is a domination, not necessarily a direct one, the result is the highest possible noun phrase with the keyword).

The third matcher is aided by the following pattern:

```
VP < (MD < [modal keywords]) < (VP !<< have)
```

The results of applying it are composed of a verb phrase, which consists of an auxiliary modal keyword, and another verb, which is not "have". The word "have" was eliminated, due to a high number of false matches, such as "I should have been there". In these cases the modal verb does not extract a goal, but a wish that something in the past went differently. This matcher was however made optional in the final version (and set to *off* by default), due to the fact that applying it resulted in certain obfuscation of the results. While the recall of the tool with the third matcher in place was slightly improved, when compared to the tool without it, the precision rates dropped significantly. More on that subject can be found within the *Evaluation* section of this document.

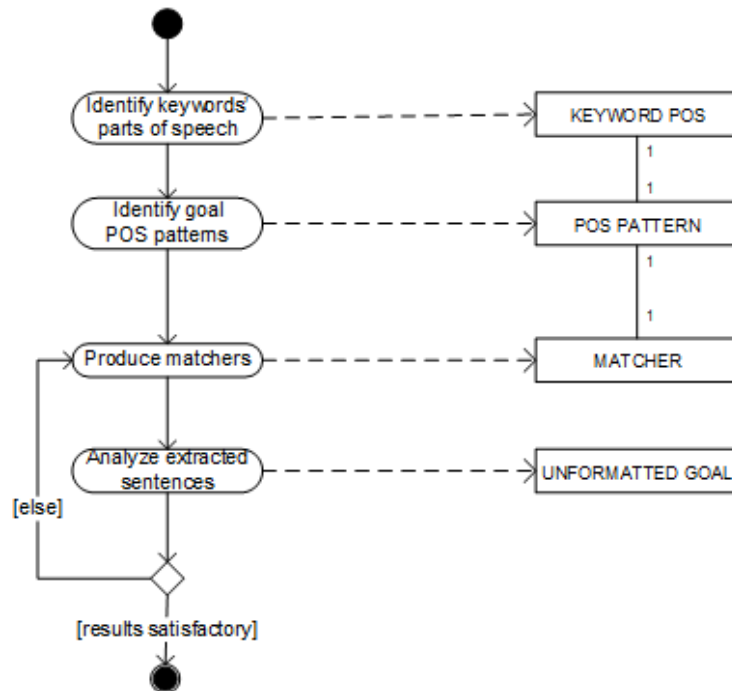


Figure 11: PDD of extracting goals

Goal Formatting Component

When the results of pattern matching are discovered, 3 types of goal-formatters are applied, in order to convert the goal to the formula specified in section 5.2, namely: [(subject) predicate complement] with only the predicate and the complement necessary.

For the first matcher, this is done through:

- Saving the keyword as the predicate,
- Re-mapping verb phrase to the highest verb phrase in the sentence, by iteratively replacing the VP with its parent if it is a verb phrase. This is done up to the point when the parent is no longer a verb phrase, thus the highest VP was identified.
- Extracting two tree lists, one of the siblings of the highest VP, and one of its children.
- Identifying the noun phrase within the siblings' tree list and saving it as the subject.
- Identifying the noun phrase/prepositional phrase within within the highest VP's children tree list and saving it as the complement. If no complement gets identified, the goal is deemed incompatible and discarded.

For the second matcher, this is done through:

- Extracting a tree list of the siblings of identified highest noun phrase,
- Identifying the highest verb phrase, and saving the verb as predicate,
- Identifying the noun phrase/prepositional phrase within within the highest VP's children tree list and saving it as the complement. If no complement gets identified, the goal is deemed incompatible and discarded.
- Extracting the tree list of the highest verb phrase,
- Exploring the tree list in pursuit of the subject.

For the third matcher this was done through:

- Extracting two tree lists, one of the siblings of the highest VP, and one of its children.
- Saving the verb adjacent to the modal verb as a predicate,
- Identifying the noun phrase within the siblings' tree list and saving it as the subject.
- Identifying the noun phrase/prepositional phrase within the highest VP's children tree list and saving it as the complement. If no complement gets identified, the goal is deemed incompatible and discarded.

After the application of this sequence of components, the tool returns a list of identified goals, in the previously specified format.

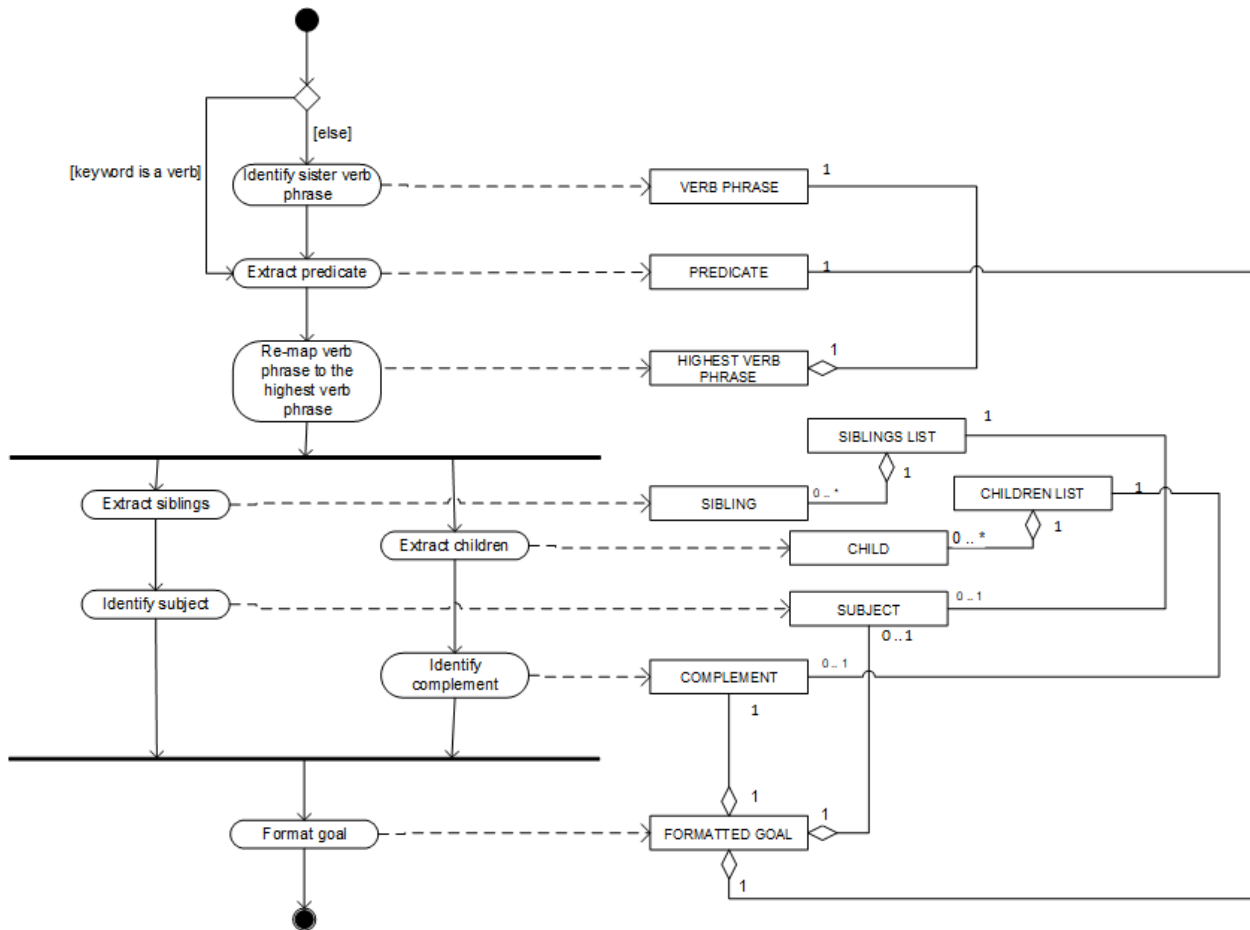


Figure 12: PDD of formatting goals

Filtering of the results

At this stage, a goal list is created, however one more step is necessary to make it useful, which is automatic filtering of as many as possible unwanted false positive results. This was done by analyzing a sample of 447 e-mails from the Enron corpus, while looking for common ground between the false positive matches, which was not shared with true positive matches. This analysis yielded the following results:

- False positives often concern personal information exchange, which is not related to the organization, however gets labelled as a match due to the appearance of keywords within it. A high number of this type of false positives has the participants of the information exchange addressing each other personally, e.g. by using the word "you". This is not as prominent within the true positives group, where the e-mail authors rarely address the recipients personally within the same sentence (or – sentence fragment) that contains the business goal.

- False positives often concern the organization of: lunches, dinners, meetings, conferences, sleeping accommodation etc. The language used often contains keywords, such as "provide", "ensure", or the modal ones used in the third matcher.
- False positives often contain first and/or last names of people, e.g. "Provide John with an update". This is not encountered that often within the true positives, however there are instances where an excerpt from a sentence can be classified as a business goal, even while explicitly stating a person's name.

Having these observations, the following filters were applied:

- Any match, which personally addresses the recipient of the e-mail is filtered out, thus not being included in the results. This had an extremely positive effect on the precision of the tool, while having an insignificantly negative effect on the recall.
- Any match, which includes the words "lunch", "dinner", "meeting", "conference", "accommodation" is filtered out as well. This filtering had a slightly positive effect on precision, while not affecting the recall at all.

Since filtering first and last names put of the results would require a next instance of a higher level natural language processing algorithm, which would in turn need much more computational power and time, while not improving the combined measure of precision and recall significantly, this idea was abandoned.

Interview and Evaluation Component

While the tool is conducting the processing of text, the user has an option of adding his/her own goals, or conducting interviews and saving the goals of multiple participants. These goals are saved with the information about their sources, and later mixed up with the text processing results. When the NLP components finish executing, and the interview goals are added to the list, the evaluators can step in and rate them.

The rating is done to eliminate the low-quality goals (or false matches) identified by the NLP components, as well as to rank the output in terms of usability. This concludes the usage of the tool within the business understanding phase of CRISP-DM.

5.2.3 Sample Results

Application of the NLP components of the tool to a folder with Monika Causholli's sent and received e-mails from the Enron Corpus produces the following output (matches from the third matcher marked in dark red):

{600m broadband investments improve disclosure }
{achieve = the mandated reductions }
{achieve account goals and an activity plan }

{achieve long-term success }
{achieve more energy-efficient usag = e of equipment }
{achieve our business objectives }
{achieve the 2 = 5 percent electrical consumption reduction target }
{avoid a downgrade }
{avoid power outage }
{clarify for everyone the process objectives }
{CorpPurpose reduce fragmentation and increase performance }
{decrease the amounts of its annual raises and bonuses }
{decrease their newsprint consumption }
{enable the Company to maintain efficiency in the balance of inventory , order flow and machine capacity in response to weaker economic conditions in the key U }
{enable the units to boost their paper production }
{enhance any and all sections }
{enhance safety and product quality }
{ensure balance between supply and customer demand , Domtar Inc. }
{ensure that sufficient storage is available for the SAP Production systems }
{ensure they are functional before I call them to encourage them to become Registered Users }
{improve 10/17/2001 05:59:07 , PRNewswire STAMFORD , Conn. }
{improve after September }
{improve by middle of 2002 }
{improve coated woodfree prices }
{improve disclosure of information surrounding the off-balance sheet liabilities }
{improve in a short period of time }
{improve its indent levels in Europe }
{improve leadership skills such as coaching and effective communication that have a direct impact on performance and managing people }
{improve operating synergies }
{improve product quality }
{improve removal of particles such as ink , wax and stickies }
{improve the group 's transparency for investors }
{improve the inventory situationsomewhat }
{increase 5 % or so while The Wall Street Journal 's net effective general ad rate will be up 3 }
{increase awarene = ss }
{increase by a few cents }
{increase by a further 10,000 ad tonnes in 2002 }
{increase by another 220,000 mt in the year 2003/4 }
{increase capacity }
{increase demand for Toronto dailies }
{increase indent prices for coated woodfree paper }
{increase its capacity at Belo Oriente facility , Minas Gerais state , from the current 800,000 to 1 million tonnes per year }
{increase mill uptime , efficiency and profitability by addressing filtration for raw water intake ; nozzle protection ; additives and coatings ; and , resource recovery needs for water , thermal and resource recovery materials }
{increase speed of presentation development since users can email their selected information to the development team }
{increase the capacity of Pacifico mill -LRB- CMPC -RRB- that will come online on Q1 2003 }
{increase the prices in November in line with what they have announced }

{increase the use of Canadian wood-frame construction technology and Canadian wood products in China }
{increase their stocks even if they wish to }
{increase total coverage to 81 % by adding only one more company }
{maintain efficiency in the balance of inventory , order flow and machine capacity in response to weaker economic conditions in the key U }
{maintain our business and our meritocracy }
{maintain the company 's competitive edge }
{make efforts to increase the prices in November in line with what they have announced }
{make for lost ad dollars }
{make sure each is configured properly by running our requirements check }
{make sure the hotfix works on a production box }
{make sure these correlations are really that high }
{My goal is to move to a trader 's position , mostly financial trader to work with derivatives }
{Oji 's 1 million tonne/yr expansion goal includes building a large PM that has an uncoated and coated woodfree capacity of 500,000 tonnes/yr }
{producers increase June output }
{provide a platform for stability }
{provide a similar analysis for mixed tropical hardwoods versus NBSK }
{provide Accredited Continuing <DIV> </DIV> Education in Nursing -LRB- CNE -RRB- with courses now available online for <DIV> </DIV> nurses , RN and other healthcare or medical professionals }
{provide all employees with background information on Anthrax and up-to-date guidance for handling any possible Anthrax exposures }
{provide an additional resource for employees who do not currently feel comfortable going to either their supervisor or their Human Resources }
{provide an opportunity for Q&A }
{provide an overview of recovered paper markets in Europe and the U. com -RSB- }
{provide any help in the future }
{provide better performance , server redundancy , and backups }
{provide continuity of earnings for permanent employees }
{provide data by port }
{provide feedback by the end of this week }
{provide greater efficiency }
{provide guarantees for the supply of raw materials }
{provide information for discussion purposes }
{provide its own outlet }
{provide manufacturing solutions -LRB- < A href = }
{provide more details in the coming weeks regarding this significant but necessary change to our Email environment }
{provide more segmented information about our business units }
{provide oversight and support for up <DIV> </DIV> to seven senior corporate auditors }
{provide some immediate relief to other Toronto dailies , such as the Toronto Star , }
{provide some information as to the future direction of contract prices }
{provide structured solutions }
{provide sufficient impetus for pulp prices to creep up again in the fourth quarter , according to one source }
{provide the answer to our ad page dilemma }
{provide with a general overview of the currency 's value and the prevailing economic environment in Brazil , as well as information relevant to the pulp market }
{provide with an insurance card }

{provide with more detailed information as it became available }
{reduce administrative , operational , financial and tax costs }
{reduce corporate overhead expenses by 16 percent , or approximately \$ 30 million annually }
{reduce demand in the second half of the year , which could involve continued capacity curtailments in the paper and pulp industry during the autumn }
{reduce downtime costs }
{reduce eliminate its purchase of slush pulp from Abitibi 's Fort Frances -LRB- ON -RRB- mill }
{reduce its sales of certain grades of these papers that = have been negatively impacted by ICMS , a form of value-added business tax }
{reduce operating costs }
{reduce operating expenses through this transition period }
{reduce our earnings estimates }
{reduce production by 17,500 tonnes }
{reduce production output by 10,000 tonnes }
{reduce Spain 's raw materials deficit , both paper recovery and the harvesting of fast-growing species on wasteland need to be increased }
{reduce the production of the Company 's least profitable products }
{reduce their trading with Enron }
{require several weeks to complete }
{support Email addresses that do not follow the standard format of firstname }
{support the families of those most affected by the tragic events of September 11th }
{support the price increases in pulp }
{try maintain prices }
{work improve this further }
{= 09Visitors to Enron facilities be escorted by an Enron employee or b = adged contractor at all times }
{= 09Visitors to the Enron Center produce a valid photo ID when signing = in at the lobby reception desk and must completely fill out the visitor ca = rd. }
{00The above tables be viewed using Courier or other fixed fonts }
{1 % over September be accompanied by a chorus of `` Happy Days are Here Again " }
{75The above tables be viewed using Courier or other fixed fonts }
{absent new , negative surprises provide a platform for stability }
{all information , documents and communications related to the merger between Enron and Dynegy be coordinated through and approved by Mark Muller , Lance Schuler , Robert Eickenroht , Mark Haedicke , Rob Walls or Greg Whalley of Enron }
{all users respond by dialing 1-800-97-ENRON or 1-800-973-6766 }
{backups complete without problems }
{be badly hurt }
{be noted , however , thatOctober 2000 counted 5 Sundays -LRB- as opposed to 4 in October 2001 -RRB- }
{be shared = outside Enron Industrial Markets }
{be used for the letters of recommendations and transcripts }
{com and begin making the necessary arrangements to start using this Internet address format if they are not using it already }
{com 's Spot Prices Center be confused with contract pricing reported in Pulp & Paper Week and PPI This Week }
{CorrelationsIt be all the German blood coursing through my veins because I 'm always suspicious of such a positive result }
{discuss matters related to the lawsuits with anyone other than the appropriate persons at Enron and its counsel }
{do that plus the rates I should use }
{Each plan be concise enough to fit on one page , and contain three sections - key account goals , financial and physical products targeted to achieve account goals and an activity plan }

{employees demonstrated the greatest contribution and behaviors , which individuals should be given greater responsibility and leadership , }
{empty the Deleted Items folder after deleting items or the space will not be freed }
{Enron restore confidence and improve disclosure of information surrounding the off-balance sheet liabilities }
{existing application systems or business cards that reference a non-supported Internet Email address need to be changed to reference the only supported firstname }
{fill out the visitor ca = rd }
{find an analysis of a publicly-traded company which should serve as an example for what I 'm looking for }
{form the principal basis for any investment or trading de = cision }
{forward their resume to Tektronix }
{general economic recovery bring relief to the industry }
{G-P be able to internally fund capex and dividends -LRB- < A href = " }
{individuals be given greater responsibility and leadership }
{intended solely for the personal use o = f the recipient , should not form the principal basis for any investment }
{Investors obtainadvicebased on their own individual circumstances before making an investmentdecision }
{it represent the very bottom of pricing in the current cycle }
{know that Enron will defend these lawsuits vigorously }
{know that this document preservation requirement is a requirement of Federal law }
{lumber producers be able to positively influence the outcome of > these > final duties by not selling low-grade products into the US }
{minimize the native AOL browser and }
{occur by year-end 2001 }
{One talk down somebody who 's talking up the market , " which is , of course , very different from agreeing with the viewpoint }
{pstHi Monika receive an e-mail in a day or 2 confirming access to the site }
{serve as an example for what I 'm looking for }
{signing = in at the lobby reception desk must completely fill out the visitor ca = rd. }
{The = views set forth herein are intended solely for the personal use of the reci = pient should not form the principal basis for any investment or trading de = cision are not intended as advice or as a recommendation or solicitatio = n of any
{the Anthrax organism be rubbed into abraded skin , swallowed , or inhaled as a fine , aerosolized mist }
{The end of FYI Toronto provide some immediate relief to other Toronto dailies , such as the Toronto Star , and possibly increase demand for Toronto dailies }
{the following measures be implemented Restore confidence in Enron }
{the forest product 's book arrive early next week , it 'll most likely come to my desk }
{The reduction be measured and achieved for = consecutive thirty-day periods commencing May 31 , 2001 }
{The restructuring reduce administrative , operational , financial and tax costs }
{their badges present a valid = picture }
{their respective activities be undertaken at arm 's length until the transaction has closed }
{then such disclosure be approved by one of the foregoing individuals }
{This be completed within the next two weeks }
{This be the person that is doing most of the work on pulp }
{This clothing bag be given to the emergency responders for proper handling }
{This form be submitted to Payroll by October 15 , 2001 }
{this information is highly proprietary should not be shared = outside Enron Industrial Markets }
{this is a rumor at this point should be viewed as such = }
{Understanding which employees have demonstrated the greatest contribution and behaviors , which individuals should be given greater responsibility and leadership , identifying our top and bottom talent }
{was very positive should provide a platform for stability absent new , negative surprises }

{We adapt our employee programs to fit the immediate needs of our company during this time of transition }
{we address to move Enron forward }
{We continue to look for ways to reduce operating expenses through this transition period }
{we define a sample that would be most representative of the total consumer inventory picture }
{We need to restart your computer in order to complete this action }
{we provide an additional resource for employees who do not currently feel comfortable going to either their supervisor or their Human Resource }

Table 6: Sample results of the application of the tool

And while some of the identified goals might be missing context, it is possible to identify the file, which contained the text in order to find the context manually. There are some false positives as well, such as *{this is a rumor at this point should be viewed as such = }*, but this is unavoidable on this level of natural language processing. It is possible to observe however, that the third (modal) matcher produces a higher number of false positives (marked in dark red within the table) than the other two. The next section of this document will deal with this subject and assess the precision and sensitivity of the tool, both with and without the modal matcher.

6 Evaluation

To evaluate the artefact's effectiveness, we picked another set of e-mails from the Enron corpus, which was unused during the development. We chose the */maildir/lay-k/sent* folder, which contained 266 e-mails sent by Kenneth Lay (who was the CEO of Enron at this time) and his assistants, as well as quotations of the e-mails he, or his assistants, were responding to. The e-mails were carefully read, and each of the business goals written down within them was marked as one, and saved in an Excel spreadsheet. This spreadsheet was later compared to the output of the tool, both with and without the inclusion of the modal matcher. This was done to determine the precision and recall of the tool, compile the f-measures of both versions, and compare them to each other.

Measures

In Information Retrieval with binary classification (either *relevant* or *not relevant*), the effectiveness of an algorithm can be measured by precision, recall, and the F-measure (also known as F₁ score, or F-score), which combines these two scores into one measurement (Goutte & Gaussier, 2005). Precision and recall can be presented graphically, as on the following image:

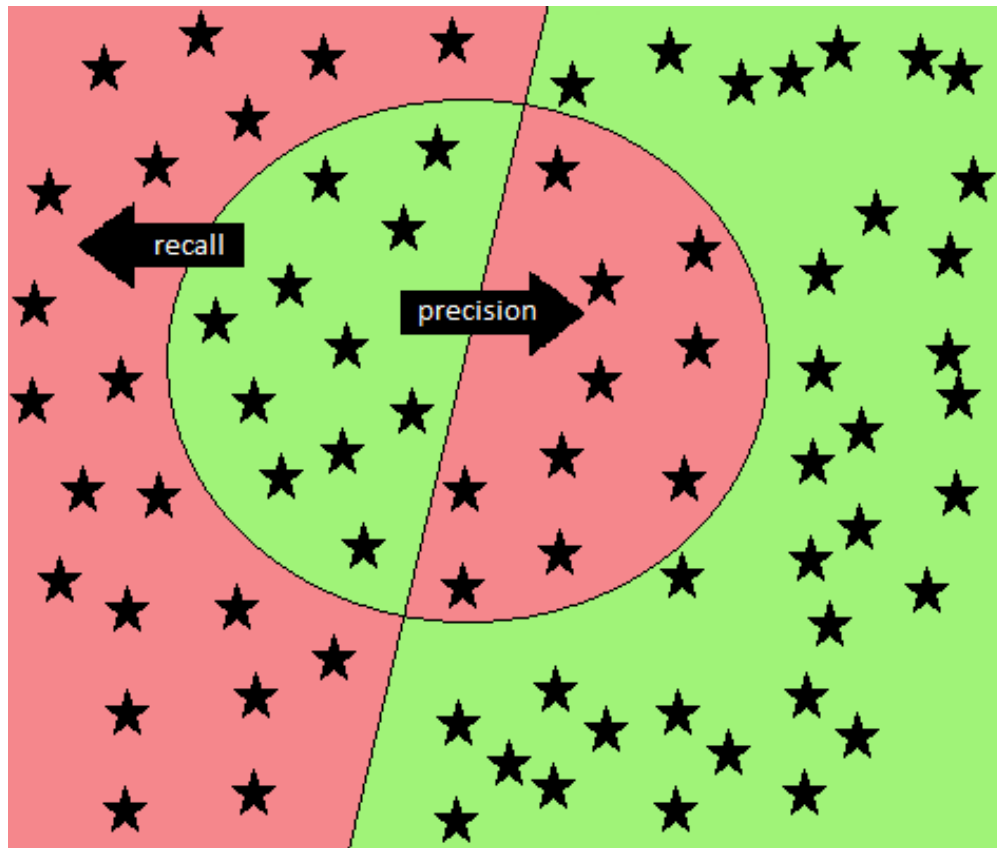


Figure 13: Precision and recall illustrated

The left-hand side represents the positives (the stars on that side being the relevant data, which should be retrieved), while the right-hand side the negatives, which should be discarded. The circle stands for the results of the application of our retrieval algorithm. The red background depicts the errors: false negatives on the left-hand side, and false positives within the circle. Precision is defined as the fraction of relevant data which was retrieved among all the data which was retrieved (green side of the circle as a part of the whole circle):

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

Recall is the fraction of relevant data which was retrieved among all the data which should have been retrieved (green side of the circle as a part of the whole left-hand side):

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

The F-measure combines these two values into one measurement – it is a harmonic mean of precision and recall.

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Data Retrieved

The table below is composed of all the business goals found within Kenneth Lay’s sent folder, and the data retrieved by the tool:

BUSINESS GOAL	MATCH (version with the 3rd matcher)	MATCH (version without the 3rd matcher)
These objectives could be accomplished through amendments to SB27X.	{= 20These objectives could be accomplished through amendments to SB27X }	{= 20These objectives could be accomplished through amendments to SB27X }
achieve a competitive retail market when == 20the utility is removed completely from the procurement function	{achieve a competitive retail market when == 20the utility is removed completely from the procurement function }	{achieve a competitive retail market when == 20the utility is removed completely from the procurement function }
it is time to take the next steps and achieve even greater progress in our public education system	{achieve even greater progress in our public education system }	{achieve even greater progress in our public education system }
achieve such a plateau	{achieve such a plateau }	{achieve such a plateau }
Decrease demand	{decrease demand = }	{decrease demand = }
decrease in member outputs	{decrease in member outputs }	{decrease in member outputs }
achieve even greater profitability	{enable our customers to achieve even greater profitability from their email marketing programs }	{enable our customers to achieve even greater profitability from their email marketing programs }

enhance the future financing needs of public companies through a consistent, fair stock valuation	{enhance the future financing needs of public companies through a consistent, fair stock valuation }	{enhance the future financing needs of public companies through a consistent, fair stock valuation }
enhance their ability to increase revenues and market share by differentiating themselves from their competition	{enhance their ability to increase revenues and market share by differentiating themselves from their competition }	{enhance their ability to increase revenues and market share by differentiating themselves from their competition }
help Enron maintain its status as the top money raiser of all MS150 rides in the NATION	{Enron maintain its status as the top money raiser of all MS150 rides in the NATION }	{Enron maintain its status as the top money raiser of all MS150 rides in the NATION }
ensure accuracy	{ensure accuracy }	{ensure accuracy }
ensure that local, parochial=20interests cannot block otherwise beneficial distributed generation projects	{ensure that local , parochial = 20interests can not block otherwise beneficial distributed generation projects = }	{ensure that local , parochial = 20interests can not block otherwise beneficial distributed generation projects = }
ensure that the executive orders (D-22-01 thru = 20D-26-01) issued by the Governor to expedite plant siting and maximize plant = 20output apply equally to smaller scale, = 01 & distributed generation = 018 facil = ities	{ensure that the executive orders -LRB- D-22-01 thru == 20D-26-01 -RRB- issued by the Governor to expedite plant siting and maximize plant == 20output apply equally to smaller scale , = 01 & distributed generation = 018 facil = ities }	{ensure that the executive orders - LRB- D-22-01 thru == 20D-26-01 - RRB- issued by the Governor to expedite plant siting and maximize plant == 20output apply equally to smaller scale , = 01 & distributed generation = 018 facil = ities }
ensure that this = 20requirement applies to all generation facilities	{ensure that this == 20requirement applies to all generation facilities , including distributed = 20generation }	{ensure that this == 20requirement applies to all generation facilities , including distributed = 20generation }
ensure that it is virus free and no responsibility is accepted by Century Development or any of its affiliates.	{ensure that it is virus free and no responsibility is accepted by Century Development or any of its affiliates }	{ensure that it is virus free and no responsibility is accepted by Century Development or any of its affiliates }
improve product development, strategic sourcing = g, = 20supply planning, manufacturing, and procurement processes across the = 20extended supply chain.	{improve product development, strategic sourcing = g, = 20supply planning, manufacturing, and procurement processes across the = 20extended supply chain }	{improve product development, strategic sourcing = g, = 20supply planning, manufacturing, and procurement processes across the = 20extended supply chain }
improve the energy consumption efficiency of their products and use of energy at their factories.	{improve the energy consumption efficiency of their products }	{improve the energy consumption efficiency of their products }
improve the program	{improve the program }	{improve the program }
improve the quality of his empire's output by better integrating its disparate parts	{improve the quality of his empire 's output by better integrating its disparate parts }	{improve the quality of his empire 's output by better integrating its disparate parts }
should improve this East and North African oil	{improve this East and North African oil }	{improve this East and North African oil }
improve this service	{improve this service }	{improve this service }
improve this by further addressing the difficulties that project developers face == 20in securing air emission reduction credits to meet the air permit = 20requirements included in the CEC 's certification requirements	{improve this by further addressing the difficulties that project developers face == 20in securing air emission reduction credits to meet the air permit = 20requirements included in the CEC 's certification requirements }	{improve this by further addressing the difficulties that project developers face == 20in securing air emission reduction credits to meet the air permit = 20requirements included in the CEC 's certification requirements }
improve upcoming issues	{improve upcoming issues }	{improve upcoming issues }
increase its non-news/current affairs production in centres other than Sydney or Melbourne by 40 percent.	{increase its non-news/current affairs production in centres other than Sydney or Melbourne by 40 percent }	{increase its non-news/current affairs production in centres other than Sydney or Melbourne by 40 percent }
increase our exposure and ability to accomplish our goals	{increase our exposure and ability to accomplish our goals }	{increase our exposure and ability to accomplish our goals }

increase profitability and market share	{increase profitability and market share through the use of rich media over the Internet }	{increase profitability and market share through the use of rich media over the Internet }
dramatically increase response rates	{increase response rates }	{increase response rates }
Increase supply	{increase supply }	{increase supply }
increase the amount permitted to paid into nuclear decommissioning reserve funds primarily for Commonwealth Edison	{increase the amount permitted to paid into nuclear decommissioning reserve funds primarily }	{increase the amount permitted to paid into nuclear decommissioning reserve funds primarily }
maintain flexibility	{maintain flexibility }	{maintain flexibility }
maintain its freedoms in the faceof politically motivated assaults	{maintain its freedoms in the faceof politically motivated assaults }	{maintain its freedoms in the faceof politically motivated assaults }
make a commitment to help TFAeliminate the digital divide across Houston and beyond	{make a commitment to help TFAeliminate the digital divide across Houston and beyond }	{make a commitment to help TFAeliminate the digital divide across Houston and beyond }
make and authorize apolitical contribution in the name of or on behalf of another unless theperson discloses in writing to the recipient the name and address of theperson actually making the contribution	{make authorize apolitical contribution in the name of or on behalf of another unless theperson discloses in writing to the recipient the name and address of theperson actually making the contribution }	{make authorize apolitical contribution in the name of or on behalf of another unless theperson discloses in writing to the recipient the name and address of theperson actually making the contribution }
reduce pollution and greenhouse gas emissions with littleeffort	{investing in energy-efficiency improvements in their homes andsimultaneously reduce pollution and greenhouse gas emissions with littleeffort }	{investing in energy-efficiency improvements in their homes andsimultaneously reduce pollution and greenhouse gas emissions with littleeffort }
Our aim is not to ban information but to restrict subscriber andadvertising revenue to an errant channel.	{Our aim is not to ban information but to restrict subscriber andadvertising revenue to an errant channel }	{Our aim is not to ban information but to restrict subscriber andadvertising revenue to an errant channel }
provide a briefing to the media > and interested members of the general public	{provide a briefing to the media > and interested members of the general public }	{provide a briefing to the media > and interested members of the general public }
provide a comprehensiveunderstanding of all aspects of the power industry in the countries of theMiddle East.	{provide a comprehensiveunderstanding of all aspects of the power industry in the countries of theMiddle East }	{provide a comprehensiveunderstanding of all aspects of the power industry in the countries of theMiddle East }
provide a limited PUHCA exemption for companies with a high level of financial stability as determined by independent market analysts	{provide a limited PUHCA exemption for companies with a high level of financial stability as determined }	{provide a limited PUHCA exemption for companies with a high level of financial stability as determined }
provide a number of logistical inputs	{provide a number of logistical inputs }	{provide a number of logistical inputs }
provide a quick and informative recap of world media news of thepast week to our friends, patrons and clients.	{provide a quick and informative recap of world media news of thepast week to our friends , patrons and clients }	{provide a quick and informative recap of world media news of thepast week to our friends , patrons and clients }
provide a succinct , but thorough assessment of the investment that will inform a discussion on how to best add value and manage the investment going forward	{provide a succinct , but thorough assessment of the investment that will inform a discussion on how to best add value }	{provide a succinct , but thorough assessment of the investment that will inform a discussion on how to best add value }
provide an economic overview of India as well as US-India economic engagement	{provide an economic overview > of India as well as US-India economic engagement }	{provide an economic overview > of India as well as US-India economic engagement }
provide an overview of Indo-US relations	{provide an overview of > Indo-US relations }	{provide an overview of > Indo-US relations }
provide a unanimous consent for your signature to formalize	{provide aunanimous consent for }	{provide aunanimous consent for }

provide Enron special recognition on both the Cougar Vision Scoreboard and over the PA in both the first and second half of the contest	{provide Enron special recognition on both the Cougar Vision Scoreboard and over the PA in both the first and second half of the contest }	{provide Enron special recognition on both the Cougar Vision Scoreboard and over the PA in both the first and second half of the contest }
provide quicker == 20stimulus	{provide quicker == 20stimulus }	{provide quicker == 20stimulus }
provide rebates=20directly to customers to fund the installation of advanced metering and=20control systems	{provide rebates=20directly to customers to fund the installation of advanced metering and=20control systems }	{provide rebates=20directly to customers to fund the installation of advanced metering and=20control systems }
provide small and medium size businesses with an affordable, yet robust solution to conduct targeted email campaigns	{provide small and medium size businesses with an affordable , yet robust solution to conduct targeted email campaigns }	{provide small and medium size businesses with an affordable , yet robust solution to conduct targeted email campaigns }
provide the recent context for President Clinton's visit to India	{provide the recent context for > President Clinton 's visit to India }	{provide the recent context for > President Clinton 's visit to India }
provide us a solid foundation with which to work	{provide us a solid foundation with which to work }	{provide us a solid foundation with which to work }
provide video streaming media appliances, software applications, technology consulting and design for the on-demand service	{provide video streaming media appliances , software applications , technology consulting and design for the on-demand service }	{provide video streaming media appliances , software applications , technology consulting and design for the on-demand service }
provide with a complete analysis * All major oil and gas field/projects	{provide with a complete analysis * All major oil and gas field/projects }	{provide with a complete analysis * All major oil and gas field/projects }
provide workforce training opportunities to half amillion persons over the next three years	{provide workforce training opportunities to half amillion persons over the next three years }	{provide workforce training opportunities to half amillion persons over the next three years }
reduce their demand for a sustained period	{reduce demand = more = 20economically by running an auction to determine the payments businesses wou = ld = 20be willing to receive to reduce their demand for a sustained period -LRB- e. nsfAll responses to Ken Lay should be sent to kenneth }	{reduce demand = more = 20economically by running an auction to determine the payments businesses wou = ld = 20be willing to receive to reduce their demand for a sustained period -LRB- e. nsfAll responses to Ken Lay should be sent to kenneth }
reduce expenses	{reduce expenses }	{reduce expenses }
reduce nitrogen oxide emissions as part of theTexas state plan for controlling smog.	{reduce nitrogen oxide emissions as part of theTexas state }	{reduce nitrogen oxide emissions as part of theTexas state }
reduce running costs	{reduce runningcosts }	{reduce runningcosts }
we would be able to reduce the> risk> factor associated with the recent release of their current software> and> hardware architectures.	{reduce the > risk > factor associated with the recent release of their current software > and > hardware architectures }	{reduce the > risk > factor associated with the recent release of their current software > and > hardware architectures }
strengthen and enhance those ties	{strengthen enhance those ties }	{strengthen enhance those ties }
support extending the offer	{support extending the offer }	{support extending the offer }
support load curtailment implementation	{support load curtailment implementation }	{support load curtailment implementation }
support our transmission open access provisions in turn for our PUHCA exemption support	{support our transmission open access provisions in turn for our PUHCA exemption support }	{support our transmission open access provisions in turn for our PUHCA exemption support }
would like to support this action	{support this action }	{support this action }
support this new Sports Arena Deal.	{support this new Sports Arena Deal }	{support this new Sports Arena Deal }
the team's goal is to raise \$350,000	{the team 's goal is to raise \$ 350,000 }	{the team 's goal is to raise \$ 350,000 }
This year's goal is to raise \$100,000	{This year 's goal is to raise \$ 100,000 }	{This year 's goal is to raise \$ 100,000 }

aim to attract the leading market players by providing a first-class program with top-rated=speakers	{ aim to attract the leading market players by providing a first-class program with top-rated=speakers }	{ aim to attract the leading market players by providing a first-class program with top-rated=speakers }
determine if we can resolve current projects before taking on additional ones.	{determine if we can resolve current projects before taking on additional ones }	{determine if we can resolve current projects before taking on additional ones }
develop advertising and other ancillary revenue streams	{develop advertising and other ancillary revenue streams }	{develop advertising and other ancillary revenue streams }
develop, execute and measure a successful campaign	{develop execute measure a successful campaign }	{develop execute measure a successful campaign }
promote energy and environmental goals	{promote energy and environmental goals }	{promote energy and environmental goals }
promote energy services	{promote energy services }	{promote energy services }
promote the image campaign , changes at the university , the tier 1 status , etc.	{promote the image campaign , changes at the university , the tier 1 status , etc. }	{promote the image campaign , changes at the university , the tier 1 status , etc. }
promote the reinvestment of profits, and a reduction in banks=01, reserve>=20requirements, which would free up money for loans.	{promote the reinvestment of profits }	{promote the reinvestment of profits }
Policy on GHG reduction must reflect the circumstances of each country and be based on a combination of regulatory measures, economic measures and voluntary measures.	{2 -RRB- Policy on GHG reduction reflect the circumstances of each country and be based on a combination of regulatory measures , economic measures and voluntary measures }	
Enron is looking at ways to allocate money to the firm	{advised that Enron is looking at ways to allocate money to the firm , and that a decision should be reached sometime in May }	
Any non-essential travel to Melbourne during this time frame should be rescheduled	{Any non-essential travel to Melbourne during this time frame be rescheduled }	
Establish a truly competitive retail electricity market	{any solution to California 's = 20crisis focus on four issues : Increase supplyDecrease demandEstablish a truly competitive retail electricity marketReturn California = 01 , s Investor-owned utilities to solvencyIncrease supply -- Legislative vehicle : SB28X -LRB- Sher -RRB- To site and construct a power plant in Texas takes approximately 2 years }	
should be the environmental agenda for business in the 21st century	{be the environmental agenda for business in the 21st century }	
the goal of having all customers served by a=20non-utility provider within 36 months.	{California begin immediately to phase the utility out of the procurement = 20function entirely , with the goal of having all customers served by a = 20non-utility provider within 36 months }	
design those rate=20increases with the dual goal of returning the utilities to solvency without>=20=01&shocking=018 the economy or household budgets	{design those rate = 20increases with the dual goal of returning the utilities to solvency without = = 20 = 01 & shocking = 018 }	
increase revenues and market share by differentiating themselves from their competition	{increase revenues and market share by differentiating themselves from their competition }	
issue a `` concept release '' to solicit broad public comments on a `` supplementary framework for reporting intangibles	{issue a `` concept release '' to solicit broad public comments on a `` supplementary framework for reporting intangibles }	
keep in mind to manage the investment going forward	{keep in mind manage the investment going forward }	
give a new boost to the country's economic and social progress	{Most key sectors benefit from these reforms , which will thus give a new boost to the country 's economic and social progress }	
provide more airtime of the Enron logo and further placement of "The New Power	{provide more airtime of the Enron logo }	

Company" in the retail consumer/viewer's venue.		
should take this as a vote of noconfidence in our online operations	{take this as a vote of noconfidence in our online operations }	
The current two-tier stoc = k = 20system should be reviewed.	{The current two-tier stoc = k = 20system be reviewed }	
the Committiee has previously requested that we optimize our existing allocation to minority money managers	{The proposal be well received as the Committiee has previously requested that we optimize our existing allocation to minority money managers }	
the system must possess t=he=20flexibility to respond to the reduced availability of power supply.	{the system possess t = he = 20flexibility to respond to the reduced availability of power supply }	
These policies should succee = din the context of liberalizing energy markets and increasing interest in th = edevelopment of a broadly based energy service industry	{These policies succee = din the context of liberalizing energy markets and increasing interest in th = edevelopment of a broadly based energy service industry }	
evaluate the impact of what is happening on our project at Dabohl with reference to enticing other foreign investment	{this forum evaluate the impact of what is happening on our project at Dabohl with reference to enticing other foreign investment }	
advise on strategy with key members of Congress and the Governor		
assess the progress of the bold and extensive measures proposed in the recent budget		
attain a level of success that will set us aside from any other group		
build strong working relationships		
conduct sophisticated online direct marketing campaigns easily and affordably		
create a task force to do the work		
deal with issues that impede, or facilitate, market development as well as enterprise strategy and metamorphosis		
detail strategy for The Center for Houston's Future capital campaign		
Energy conservation is to make up the cost reduction of products and recycling will bring about the effect on manufacturing cost saving in the long run.		
establish contract terms with th=e=20goal of entering into a power purchase agreement as soon as possible.		
exclude water and sewage connection fees from gross income as contributions to capital		
gain from a changed economic relationship		
gain market access for energy services		
help them see the relationship between power and energy related objectives and environmental issues including GHG emissions		
hold a series of competitive solicitations over the 36-month period =in=20which competing service providers would bid for the right to serve segments=20of utility load.		

identify and implement the major environmental themes for the plenary session		
insure stockholder protection along with consumer protection		
keep the tight financial and risk controls in place		
license these technology assets to third parties both domestically and internationally		
net their contribution obligation against other contributions that they solicited		
optimize our existing allocation to minority money managers		
Policies will have to be devised that take account of developments in scientific knowledge and opinion and that also maintains flexibility.		
profit from the current difficult market		
pursue new funding opportunities		
reduce the risk of black outs this summer		
retaining an exceptional workforce in our world class city		
review where we are, the political environment, and continued Center involvement in this program area.		
spin this so that businesses create the bridge for the divide and not government		
see where there are opportunities for EXIM to work more closely with the company.		
track and analyze results and generate real-time reports		
validate to the Swiss insurance company that is putting up financing.		
We must not allow our city to lose its competitive advantage on or off the court		
	{make the cost }	{make the cost }
	{Mr Lay provide a number of logistical inputs , including a biography and photo , as well as information on any remarks he might choose to make }	{Mr Lay provide a number of logistical inputs , including a biography and photo , as well as information on any remarks he might choose to make }
	{products and services require the approval of }	{products and services require the approval of }
	{provide additional detail }	{provide additional detail }
	{provide an update on some of the new and successful research underway for Alzheimer 's disease }	{provide an update on some of the new and successful research underway for Alzheimer 's disease }
	{provide for Mr. about the opportunity }	{provide for Mr. about the opportunity should }

	{provide his perspective on > India and the United States how they perceive and misperceive each other , > and how some of the misperceptions might be addressed }	{provide his perspective on > India and the United States how they perceive and misperceive each other , > and how some of the misperceptions might be addressed }
	{provide keynote address }	{provide keynote address }
	{provide local personal security agents if }	{provide local personal security agents if }
	{provide Mr. The impact on the business }	{provide Mr. The impact on the business }
	{provide us with the following information }	{provide us with the following information }
	{provide valueand solution to }	{provide valueand solution to }
	{provide with our honest > evaluation in }	{provide with our honest > evaluation in }
	{reduce = 018 ratin = gs }	{reduce = 018 ratin = gs }
	{reduce wordiness }	{reduce wordiness }
	{support Malaysiakini }	{support Malaysiakini }
	{This special purpose PAC is subject to thegeneral Texas laws that govern campaign fundraising }	{This special purpose PAC is subject to thegeneral Texas laws that govern campaign fundraising }
	{us install develop test }	{us install develop test }
	{develop a broader , more multi-faceted relationship between the world 's > }	{develop a broader , more multi-faceted relationship between the world 's > }
	{promote the notion of a person `s personal or religious beliefs }	{promote the notion of a person `s personal or religious beliefs }
	{me promote tolerance and acceptance of all people }	{me promote tolerance and acceptance of all people }
	{promote own position }	{promote own position }
	{ = 01 & on-site = 018 -RRB- generation that is 50 MWs = or = 20greater receive certification from the California Energy Commission an = d = 20therefore face all of the impediments to development that large-scale = 20generation faces }	
	{> > companies do to become radical innovators }	
	{2 page A4 -RRB- in French and EnglishPhotograph -LRB- colour or black/white -RRB- The biography be sent as a word document by return emailalong with a photograph in jpeg or tiff format or an original }	
	{a session propose topics }	
	{arenot suitable and 35mm slides or graphic presentations -LRB- PowerPoint -RRB- beused }	
	{Australia 's ABC decentralise to lessen Sydney bias }	
	{be listed here or on ourwebsite }	
	{Bids be returned by 2pm Wednesday , April 11 , 2001 }	
	{California shift control over interconnection away from th = e = 20utility and place that control with the California ISO }	

	{Cavallo take a different tack than L }	
	{Chair the Commerce Committee should the Democrats regain control of the House }	
	{Chile , which led Latin American nations in adopting == 20free-market reforms two decades ago , bet big on attracting high-te = ch = 20industries }	
	{companies do to become radical innovators }	
	{Congressman Dingell -LRB- D-MI -RRB- and Congressman Markey -LRB- D-MA -RRB- actively oppose this limited exemption , especially given the fact that Congressman Dingell -LRB- who would Chair the Commerce Committee should the Democrats regain control of the House -RRB- has recently said that whether or not he is Commerce Committee Chairman next Congress , that comprehensive electricity restructuring legislation including PUHCA repeal is `` at least three years away from happening }	
	{contact either Jaime Alatorre or Max Yzaguirre }	
	{doc -RRB- Sarah , e-mail me exactly where they should come once they get through traffic }	
	{everyone be congratulated that worked on the planning }	
	{feel free however }	
	{first issue is on its way should arrive shortly }	
	{give `` heads-up `` in case }	
	{go ahead equip it }	
	{It be understood , that there has not been a final decision > }	
	{It be understood , that there has not been a final decision > and > management has not approved a migration plan }	
	{It influence not > > only top management , but also every employee who , indeed , is the CEO of > > their own business life }	
	{It influence not only top management > > but also every employee who , indeed , is the CEO of their own business > > life }	
	{MD saysThe Australian Broadcasting Corporation decentralise both staff and activities to prevent it from becoming a Sydney mouthpiece }	
	{must see the irony here not to mention the double standard in how Enron 's Visions & Values are applied to assistants all over Enron }	
	{need further assistance }	
	{nsfAll responses to this e-mail be sent to kenneth }	
	{nsfFURTHER CLARIFICATION TO THE MESSAGE BELOW the members of the Office of the Chairman approve the requests }	
	{proposed sessions and inquiries be submitted to : David Williams , Executive Director , USAEE/IAEE28790 Chagrin Blvd. }	
	{PUHCA apply should the company lose its high rating }	
	{read , copy , use or disclose this communication }	
	{send his resume to Celeste Roberts at -LRB-713-RRB- 853-0555 }	
	{That language sound familiar }	
	{The administration be embarrassed }	

	{The biography be sent as a word document by return email along with a photograph in jpeg or tiff format or an original }	
	{The contributions be paid to the referendum campaign by the partnership , on behalf of each partner , upon receipt of instructions from each such partner that its share of the amount prepaid should be contributed to the campaign instead of being distributed }	
	{the internet had its place within media should not be sidelined just because of the recent dot com gloom }	
	{The sites sign up at a ' NewsStand ' section to secure their free delivery of each day 's top stories }	
	{they agree to place an MSNBC }	
	{They arrive in Davos about 9:00 a. DeLay has requested support from and interaction with the BCCA and the Partnership on that day }	
	{they come once they get through traffic }	
	{use , copy , disclose or take any action based on this message or any information herein }	
	{We allow our city to lose its competitive advantage on or of = 20 the court }	
	{We pack the punches that we ought to " }	
	{we say a polite " no " }	

Table 7: Results of the application of the tool to Enron CEO's "sent" folder

The green background of the cells represents the correct matches (true positives), while the pink incorrect ones (false positives), or lack of matches for existing business goals (false negatives). In case of incomplete matches (at first sight inconclusive if they are business goals due to incomplete extraction from the sentence), the user of the tool has a possibility of determining if they are noteworthy by being redirected to the complete textual environment of the match. This table considers that step to be completed by dividing these matches to true or false, based on their textual environment.

Results

This creates the ground for calculations of precision, recall, and F-measure. In case of the first version of the tool (including the modal matcher), these are:

Precision:

$$P_1 = \frac{92}{92 + 69} = 0.571$$

Recall:

$$R_1 = \frac{92}{92 + 32} = 0.742$$

F-measure:

$$F_1 = 2 \cdot \frac{0.571 \cdot 0.742}{0.571 + 0.742} = 0.645$$

These results can be compared to their equivalents for the second version of the tool, which does not use the modal matcher. In this case, the results are the following:

Precision:

$$P_2 = \frac{74}{74 + 22} = 0.77$$

Recall:

$$R_2 = \frac{74}{74 + 50} = 0.597$$

F-measure:

$$F_2 = 2 \cdot \frac{0.77 \cdot 0.597}{0.77 + 0.597} = 0.672$$

This confirms the slight advantage of the second version of the tool, when it comes to its usability. High precision seems to be more important than high recall, when it comes to analyzing tens of thousands lines of text, especially considering the fact that not all business goals can be translated to data mining goals. The modal matcher was however left in the code, and can be activated, if needed.

Another aspect of evaluation of the tool is its applicability within the business understanding phase of CRISP-DM. This concerns the usefulness of the extracted goals, when it comes to their translatability to data mining goals. This cannot be however objectively done without knowing the data mining potential of the organization's resources. Assuming that the organization collects and/or has access to all the data, which could be mined to support the identified business goals, the applicability of the results is the following:

Extracted business goal	Potential for a data mining goal?	Extracted business goal	Potential for a data mining goal?
{= 20These objectives could be accomplished through amendments to SB27X }	NO	{provide a succinct , but thorough assessment of the investment that will inform a discussion on how to best add value }	YES
{achieve a competitive retail market when = 20the utility is removed completely from the procurement function }	YES	{provide an economic overview > of India as well as US-India economic engagement }	YES
{achieve even greater progress in our public education system }	NO	{provide an overview of > Indo-US relations }	YES
{achieve such a plateau }	YES	{provide a unanimous consent for }	NO

	(refers to drawing the most successful alumni of University of Houston to the company)		
{decrease demand = }	YES	{provide Enron special recognition on both the Cougar Vision Scoreboard and over the PA in both the first and second half of the contest }	NO
{decrease in member outputs }	YES (refers to production cuts and subsequent increase in oil prices)	{provide quicker == 20stimulus }	YES (refers to planning rate cuts)
{enable our customers to achieve even greater profitability from their email marketing programs }	YES	{provide rebates=20directly to customers to fund the installation of advanced metering and=20control systems }	NO
{enhance the future financing needs of public companies through a consistent, fair stock valuation }	YES	{provide small and medium size businesses with an affordable , yet robust solution to conduct targeted email campaigns }	YES
{enhance their ability to increase revenues and market share by differentiating themselves from their competition }	YES	{provide the recent context for > President Clinton 's visit to India }	NO
{Enron maintain its status as the top money raiser of all MS150 rides in the NATION }	YES	{provide us a solid foundation with which to work }	YES (refers to planning increases in participation)
{ensure accuracy }	YES	{provide video streaming media appliances , software applications , technology consulting and design for the on-demand service }	YES
{ensure that local , parochial = 20interests can not block otherwise beneficial distributed generation projects = }	NO	{provide with a complete analysison * All major oil and gas field/projects }	YES
{ensure that the executive orders - LRB- D-22-01 thru == 20D-26-01 - RRB- issued by the Governor to expedite plant siting and maximize plant == 20output apply equally to smaller scale , = 01 & distributed generation = 018 facil = ities }	YES	{provide workforce training opportunities to half amillion persons over the next three years }	YES
{ensure that this == 20requirement applies to all generation facilities , including distributed = 20generation }	YES	{reduce demand = more = 20economically by running an auction to determine the payments businesses wou = ld = 20be willing to receive to reduce their demand for a sustained period -LRB- e. nsfAll responses to Ken Lay should be sent to kenneth }	YES
{ensure that it is virus free and no responsibility is acceptedby Century Development or any of its affiliates }	NO	{reduce expenses }	YES
{improve product development, strategic sourcin=g.=20supply planning, manufacturing, and procurement processes across the=20extended supply chain }	YES	{reduce nitrogen oxide emissions as part of theTexas state }	YES
{improve the energyconsumption efficiency of their products }	YES	{reduce runningcosts }	YES
{improve the program }	NO	{reduce the > risk > factor associated with the recent release of their current	YES

	(refers to obtaining someone's opinion)	software > and > hardware architectures }	
{improve the quality of hisempire 's output by better integrating its disparate parts }	YES	{strengthen enhance those ties }	NO
{improve this East and NorthAfrican oil }	YES	{support extending the offer }	NO
{improve thisservice }	NO (refers to obtaining someone's opinion)	{support load curtailment implementation }	YES
{improve tho by further addressing the difficulties that project developers face == 20in securing air emission reduction credits to meet the air permit = 20requirements included in the CEC 's certification requirements }	YES	{support our transmission open access provisions in turn for our PUHCA exemption support }	NO
{improve upcoming issues }	NO (refers to obtaining someone's opinion)	{support this action }	NO
{increase itsnon-news/current affairs production in centres other than Sydney orMelbourne by 40 percent }	YES	{support this new Sports Arena Deal }	NO
{increase our exposure and ability to accomplish our goals }	YES	{the team 's goal is to raise \$ 350,000 }	NO
{increase profitability and market share through the use of rich media over the Internet }	YES	{This year 's goal is to raise \$ 100,000 }	NO
{increase response rates }	YES	{investing in energy-efficiency improvements in their homes andsimultaneously reduce pollution and greenhouse gas emissions with littleeffort }	YES
{increase supply }	YES	{Our aim is not to ban information but to restrict subscriber andadvertising revenue to an errant channel }	YES
{increase the amount permitted to paid into nuclear decommissioning reserve funds primarily }	NO	{provide a briefing to the media > and interested members of the general public }	NO
{maintain flexibility }	NO (refers to not making information public)	{provide a comprehensiveunderstanding of all aspects of the power industry in the countries of theMiddle East }	YES
{maintain its freedoms in the faceof politically motivated assaults }	NO	{provide a limited PUHCA exemption for companies with a high level of financial stability as determined }	NO
{make a commitment to help TFAeliminate the digital divide across Houston and beyond }	NO	{provide a number of logistical inputs }	YES
{make authorize apolitical contribution in the name of or on behalf of another unless theperson discloses in writing to the recipient the name and address of theperson actually making the contribution }	NO	{provide a quick and informative recap of world media news of thepast week to our friends , patrons and clients }	YES

{ aim to attract the leading market players by providing a first-class program with top-rated=speakers }	NO	{promote energy and environmental goals }	YES
{determine if we can resolve current projects before taking on additional ones }	YES	{promote energy services }	YES
{develop advertising and other ancillary revenue streams }	YES	{promote the image campaign , changes at the university , the tier 1 status , etc. }	YES
{develop execute measure a successful campaign }	YES	{promote the reinvestment of profits }	NO

Table 8: Potential usability of the results

The results look promising, however a case study at an organization would be a much more fruitful way of assessing the potential of the results, as the stakeholders could help with that assessment. What is more, the data set (assuming being granted access to the organization’s e-mail repository) would be more relevant than the 17-year-old Enron Corpus.

The artefact has multiple possible directions for evolution. The goal identification component can be readjusted to account for other ways of detection of business goals, if these are developed. The modularity of the tool also allows for the addition of extra components, or the replacement of the currently used ones, if needed.

7 Conclusions

This section of the document presents the conclusions of the research, as well as limitations of the created artefact, and possible directions of future research on this subject.

7.1 Research Questions

This research was conducted to investigate if the stated research questions can be answered. Starting with the sub-questions:

RSQ1: Are there any theoretical models designed to help with data mining business goals definition?

The literature review gives a clear answer to that question. Guidelines on how to deploy the early stages of data mining projects do exist, however they severely lack in detail, when it comes to the definition of business goals. What is more, researchers agree that this phase should be clearer and more feasible to implement than it is now, especially considering the fact that it is crucial to the success of the entire project.

RSQ2: What do enterprises do in practice to define their data mining business goals?

The results of the survey are inconclusive in this matter. Only a quarter of the respondents admits to using non-standard tools and services to aid with the early stages of their data mining projects, however there is a wide variety of choices between them and no typical answer. It is worth noting, that CRISP-DM, while being the most common data mining process used in the field, is still deployed only by less than 12% of the respondents. The overwhelming majority chooses to use either their own processes, or the ones created by the organizations they work for. This confirms the idea that a more specific set of directions is needed within the industry standard process, in order for it to be more widely acknowledged.

RSQ3: Which textual resources could be helpful to bootstrap the business understanding phase of data mining efforts in practice, and in what way?

A number of reviewed scientific papers confirms that e-mail exchanges between the members of the organizations contain a lot of valuable knowledge about the businesses and their needs, and that capturing this knowledge exchange can be a valuable asset in overall knowledge management strategy of the enterprise. As this type of repository is relatively easy to mine, it can be used for future information extraction, in this case extraction of the business goals. These business goals can be later evaluated in the context of data mining, and – if applicable – translated to data mining goals. This leads us to the main research question, which was stated as:

RQ: *To what extent can enterprises semi-automatically uncover the definition of their data mining business goals using text analytics?*

Text analytics can be deployed to find the definitions of business goals at an organization relatively precisely when it comes to using an internal e-mail repository as a subject to analysis. Depending on the effort one might want to put into the second-level manual analysis and rejecting the false positives, the recall results of the analytics can also be enhanced, but – as in all cases – it is a trade-off against precision. Not all of these goals however are translatable to data mining goals. This can be one of the subjects of further research (section 7.3) on this topic.

7.2 Limitations

There are several limitations, which were noted while developing this project, the first being the definition of a business goal. As there is no common framework of identification of business goals through their syntax, I used the only suggestions found in the literature to develop a textual formula. This formula is not by any means perfect, and – as visible in the evaluation section – it is inevitable for the encoded business goals to slip through the algorithm and not be recorded as matches. Aside from that, the spelling mistakes, or formatting errors are not accounted for. Due to that, for instance the business goal written down as “reduc=e=20the risk of black outs this summer” was not recorded as one in the evaluation section of this document.

Another limitation to this research is the fact of it being partially developed (the filtering component) and fully tested on the basis of a 17-year-old e-mail data set. While being an incredible repository of thousands of real-life corporate e-mails, it is unfortunately quite outdated in terms of content. While correspondence from a company, which needed funding to revolutionize e-mail by making it auto-play voice messages was an entertaining read, other e-mails contained a lot of personal messages to and from outside of the organization; i.e. friends, family, newsletters, etc.; or multiple “test-messages” to see if the e-mail reaches the respondent; some inboxes were shared between the employees and their assistants, who replied to their e-mails. Using a newer data set would be more beneficial to the research, especially considering how many more goals related to data mining it could contain, however there was no better fit available.

7.3 Future Research

Multiple directions for future research can be taken, starting with a clearer definition of a business goal for its extraction, or even a formula for a business goal, which could be translated to a data mining goal. As mentioned before, academic literature does not provide a unified formula for identifying business goals, or even a guideline for writing them down. Undoubtedly, if such a definition existed, it would be translatable to the natural language processing terms, which would make the business goals easier to extract with higher precision and recall results. Considering how

beneficial a framework for goal extraction through text analysis could be for organizations, not only for data mining, our first recommendation for future research is the development of a clearer formula for encoding business goals. When it comes specifically to data mining, and CRISP-DM, the business goals identified through the tool are not always translatable to data mining goals. This could be another expansion of this research – identification of data mining goals within the extracted business goals, thus bootstrapping the business understanding phase along with data understanding.

Another recommendation would be a multiple case-study: deployment of the tool at multiple organizations, using their internal e-mail repositories as data sets. This way, the obtained results can be analyzed, and the natural language processing component, along with the filtering component can be readjusted accordingly to these results. This would measure the generalizability of this research, and make it possible to enhance it, wherever the room for improvement can be found.

References

- Aggarwal, C. C., & Zhai, C. (Eds.). (2012). *Mining text data*. New York, NY: Springer Science & Business Media.
- Angles, R., & Gutierrez, C. (2008). Survey of graph database models. *ACM Computing Surveys (CSUR)*, 40(1), 1.
- Ananiadou, S., & McNaught, J. (2006). *Text mining for biology and biomedicine* (pp. 1-12). London: Artech House.
- Berry, M. J., & Linoff, G. (2000). *Data mining techniques: for marketing, sales, and customer support*. New York, NY: John Wiley & Sons, Inc..
- Becher, J. D., Berkhin, P., & Freeman, E. (2000). Automating exploratory data analysis for efficient data mining. *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 424-429).
- Bergmair, R. (2004). Natural Language Steganography and an "AI-complete" Security Primitive. *21st Chaos Communication Congress, Berlin (December 2004)*.
- Blockeel, H., & Sebag, M. (2003). Scalability and efficiency in multi-relational data mining. *ACM SIGKDD Explorations Newsletter*, 5(1), 17-30.
- Brachman, R. J., & Anand, T. (1996, February). The process of knowledge discovery in databases. *Advances in knowledge discovery and data mining* (pp. 37-57). American Association for Artificial Intelligence..
- Britos, P., Dieste, O., & García-Martínez, R. (2008). Requirements elicitation in data mining for business intelligence projects. In D. Avison, G.M. Kasper, B. Pernici, I. Ramos, & D. Roode (Eds.). *Advances in Information Systems Research, Education and Practice* (pp. 139-150). New York, NY: Springer US.
- Campbell, C. S., Maglio, P. P., Cozzi, A., & Dom, B. (2003, November). Expertise identification using email communications. In *Proceedings of the twelfth international conference on Information and knowledge management* (pp. 528-531). ACM.
- Casagrande, E., Woldeamlak, S., Woon, W. L., Zeineldin, H. H., & Svetinovic, D. (2014). NLP-KAOS for systems goal elicitation: Smart metering system case study. *IEEE Transactions on Software Engineering*, 40(10), 941-956.
- Cerf, L., Besson, J., Nguyen, K. N. T., & Boulicaut, J. F. (2013). Closed and noise-tolerant patterns in n-ary relations. *Data Mining and Knowledge Discovery*, 26(3), 574-619.

- Chan, A. P., Scott, D., & Lam, E. W. (2002). Framework of success criteria for design/build projects. *Journal of management in engineering*, 18(3), 120-128.
- Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, 37(1), 51-89.
- De Marneffe, M. C., MacCartney, B., & Manning, C. D. (2006, May). Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC* (Vol. 6, No. 2006, pp. 449-454).
- Domingos, P. (2003). Prospects and challenges for multi-relational data mining. *ACM SIGKDD explorations newsletter*, 5(1), 80-83.
- Džeroski, S. (2003). Multi-relational data mining: an introduction. *ACM SIGKDD Explorations Newsletter*, 5(1), 1-16.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37.
- Feldman, S. (1999). NLP Meets the Jabberwocky: Natural Language Processing in Information Retrieval. *ONLINE-WESTON THEN WILTON-*, 23, 62-73.
- Fisher, D., Brush, A. J., Gleave, E., & Smith, M. A. (2006). Revisiting Whittaker & Sidner's email overload ten years later. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work* (pp. 309-312). ACM.
- Goutte, C., & Gaussier, E. (2005, March). A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *ECIR* (Vol. 5, pp. 345-359).
- Grobelnik, M., Mladenic, D., & Fortuna, B. (2009). Semantic technology for capturing communication inside an organization. *IEEE internet computing*, 13(4).
- IBM. (2011). *IBM SPSS Modeler CRISP-DM Guide*.
- Hearst, M. A. (1999). Untangling text data mining. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 3-10).
- Klein, D., & Manning, C. D. (2003, July). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1* (pp. 423-430). Association for Computational Linguistics.
- Klimt, B., & Yang, Y. (2004). The enron corpus: A new dataset for email classification research. *Machine learning: ECML 2004*, 217-226.
- Kohavi, R., & Provost, F. (2001). *Applications of data mining to electronic commerce* (pp. 5-10). New York, NY: Springer US.

Kurgan, L. A., & Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review*, 21(1), 1-24.

Levy, R., & Andrew, G. (2006, May). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of the fifth international conference on Language Resources and Evaluation* (pp. 2231-2234).

Lezcano, R., Alfonso, L., Guzmán, L., Alberto, J., Gómez, A., & Alonso, S. (2015). Extraction of goals and their classification in the KAOS model using natural language processing. *Ingeniare. Revista chilena de ingeniería*, 23(1), 59-66.

Liddy, E. D. (2001). Natural language processing.

Lucero, A. (2015). Using affinity diagrams to evaluate interactive prototypes. In *Human-Computer Interaction* (pp. 231-248). Springer, Cham.

Madhusudhan, C. H., & Rao, K. M. (2015). Trends in Multi-Relational Data Mining Methods. *International Journal of Computer Science and Network Security (IJCSNS)*, 15(8), 101.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55-60.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1, No. 1). Cambridge: Cambridge university press.

Marbán, O., Segovia, J., Menasalvas, E., & Fernández-Baizán, C. (2008). Toward data mining engineering: A software engineering approach. *Information systems*, 34(1), 87-107.

Merali, Y., & Davies, J. (2001, October). Knowledge capture and utilization in virtual communities. In *Proceedings of the 1st international conference on Knowledge capture* (pp. 92-99). ACM.

Mooney, R. J., Melville, P., Tang, L. R., Shavlik, J., Castro Dutro, I. D., Page, D., & Costa, V. S. (2002). *Relational data mining with inductive logic programming for link discovery*. Austin, TX: Texas University at Austin.

Nahm, U. Y., & Mooney, R. J. (2002). Text mining with information extraction. *Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases*, 60-67.

Piatecki, G. (2014, October). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. Retrieved April 17, 2016, from <http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>

Prat, N., Comyn-Wattiau, I., & Akoka, J. (2014). Artifact Evaluation in Information Systems Design-Science Research - a Holistic View. In *PACIS* (p. 23).

- Santorini, B. (1990). Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision).
- Scupin, R. (1997). The KJ method: A technique for analyzing data derived from Japanese ethnology. *Human organization*, 56(2), 233-237.
- Shahin, A., & Mahbod, M. A. (2007). Prioritization of key performance indicators: An integration of analytical hierarchy process and goal setting. *International Journal of Productivity and Performance Management*, 56(3), 226-240.
- Sharma, S., & Osei-Bryson, K. M. (2009). Framework for formal implementation of the business understanding phase of data mining projects. *Expert Systems with Applications*, 36(2), 4114-4124.
- Sharma, S., Osei-Bryson, K. M., & Kasper, G. M. (2012). Evaluation of an integrated Knowledge Discovery and Data Mining process model. *Expert Systems with Applications*, 39(13), 11335-11348.
- Spool, J. M. (2004). The KJ-technique: A group process for establishing priorities. *User interface engineering*.
- Soderland, S. (1999). Learning information extraction rules for semi-structured and free text. *Machine learning*, 34(1-3), 233-272.
- Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003, May). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (pp. 173-180). Association for Computational Linguistics.
- van de Weerd, I., & Brinkkemper, S. (2009). Meta-modeling for situational analysis and design methods. In *Handbook of research on modern systems analysis and design technologies and applications* (pp. 35-54). IGI Global.
- van Lamsweerde, A. (2009). *Requirements engineering: From system goals to UML models to software* (Vol. 10). Chichester, UK: John Wiley & Sons.
- Vleugel, A., Spruit, M., & van Daal, A. (2010). Historical data analysis through data mining from an outsourcing perspective: the Three-phases model. In R. T. Herschel (Ed.). (2012). *Organizational Applications of Business Intelligence Management: Emerging Trends* (236-260). Hershey, PA: IGI Global.
- Wang, H., & Wang, S. (2008). A knowledge management approach to data mining process for business intelligence. *Industrial Management & Data Systems*, 108(5), 622-634.

Appendix A

Full set of answers to the questionnaire

What is the size of the organization you work for?

101-250 employees (medium), 51-100 employees (medium), 250+ employees (large), 250+ employees (large), 250+ employees (large), 51-100 employees (medium), 1-10 employees (micro), 11-50 employees (small), 250+ employees (large), 11-50 employees (small), 250+ employees (large), 250+ employees (large), 11-50 employees (small), 51-100 employees (medium), 1-10 employees (micro), 250+ employees (large), 250+ employees (large), 11-50 employees (small), 250+ employees (large), 11-50 employees (small), 250+ employees (large), 250+ employees (large), 1-10 employees (micro), 250+ employees (large), 250+ employees (large), 250+ employees (large), 250+ employees (large), 250+ employees (large), 250+ employees (large), 250+ employees (large), 250+ employees (large), 101-250 employees (medium), 250+ employees (large), 11-50 employees (small), 250+ employees (large), 11-50 employees (small), 250+ employees (large), 250+ employees (large), 51-100 employees (medium), 250+ employees (large), 11-50 employees (small), 250+ employees (large), 250+ employees (large), 11-50 employees (small), 51-100 employees (medium), 51-100 employees (medium)

Does your organization collect data to support decision making?

Yes	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes		

Does your organization have a specialized Data Mining division?

Yes	No		Yes	No	No	No		Yes	Yes	Yes	No
Yes	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Yes	No
No	Yes	No	Yes	No	Yes	No	Yes	No	No	Yes	No
Yes	No		No	No	Yes	No	No	Yes	No		

Which methodology do you use the most often for data mining?

Your organizations', CRISP-DM, , Your organizations', Your own, Your organizations', Other, , Your own, Your organizations', Other, Your own, Your own, Your own, Your own, Other, Your own, Your own, Other, Your own, Other, Your own, CRISP-DM, CRISP-DM, Your own, Other, CRISP-DM, Your organizations', Your organizations', Your own, CRISP-DM, Your organizations',

Your own, Your organizations', Other, Your own, Your own, Your organizations', , Your own, Your own, Your organizations', Your organizations', Your organizations', Your own, Your own

Which of these activities do you perform at the early stages of the project?

{Assess background to the project, Conduct interviews with stakeholders}, {Assess background to the project, Conduct interviews with stakeholders, Define the success criteria}, {}, {Assess background to the project, Review corporate documentation}, {Assess background to the project, Conduct interviews with stakeholders, Define the success criteria}, {Assess background to the project, Review corporate documentation, Conduct interviews with stakeholders, Assess risks, Define the success criteria, Write a project plan}, {Conduct interviews with stakeholders, Define the success criteria, Write a project plan}, {}, {Conduct interviews with stakeholders, Write a project plan}, {Assess background to the project}, {Write a project plan}, {Assess background to the project}, {Assess background to the project, Define the success criteria, Write a project plan}, {Write a project plan}, {Conduct interviews with stakeholders, Define the success criteria, Write a project plan}, {Assess background to the project, Conduct interviews with stakeholders, Assess risks}, {Assess background to the project, Write a project plan}, {Define the success criteria}, {Assess background to the project, Review corporate documentation, Conduct interviews with stakeholders, Assess risks, Define the success criteria, Write a project plan}, {Assess background to the project, Conduct interviews with stakeholders, Define the success criteria, Write a project plan}, {Assess background to the project, Review corporate documentation, Conduct interviews with stakeholders, Assess risks, Write a project plan}, {Assess background to the project}, {Assess background to the project, Review corporate documentation}, {Assess background to the project, Conduct interviews with stakeholders, Define the success criteria, Write a project plan}, {Assess background to the project, Assess risks, Define the success criteria, Write a project plan}, {Review corporate documentation, Conduct interviews with stakeholders, Write a project plan}, {Assess background to the project, Review corporate documentation, Conduct interviews with stakeholders, Assess risks, Define the success criteria, Write a project plan}, {Assess background to the project, Review corporate documentation, Conduct interviews with stakeholders, Assess risks, Define the success criteria, Write a project plan}, {Assess background to the project, Review corporate documentation, Write a project plan}, {Assess background to the project, Conduct interviews with stakeholders, Assess risks, Define the success criteria, Write a project plan}, {Assess background to the project, Conduct interviews with stakeholders, Define the business problem}, {Assess background to the project, Review corporate documentation, Conduct interviews with stakeholders, Assess risks, Define the success criteria}, {Assess background to the project}, {Assess background to the project, Review corporate documentation, Conduct interviews with stakeholders, Assess risks, Define the success criteria, Write a project plan}, {Assess background to the project, Conduct interviews with stakeholders, Assess risks, Define the success criteria}, {Assess background to the project, Review corporate documentation, Conduct interviews with stakeholders, Assess risks, Define the success criteria, Write a project plan, Profile available data}, {Write a project plan}, {Assess background to the project, Review corporate

documentation, Conduct interviews with stakeholders, Assess risks, Define the success criteria, Write a project plan}, {}, {Assess background to the project, Review corporate documentation, Conduct interviews with stakeholders, Assess risks, Define the success criteria, Write a project plan, Data gathering rules and regulations}, {Assess background to the project}, {Assess background to the project, Assess risks, Define the success criteria, Write a project plan}, {Conduct interviews with stakeholders}, {Conduct interviews with stakeholders, Define the success criteria}, {Assess background to the project, do we have the data and resources we need? Assess background to the project, Write a project plan}

How do you rate the feasibility of the activities from the previous question under the chosen methodology?

3	2		3	3	4	3		3	3	4	3
3	1	3	3	3	3	5	2	3	3	3	4
2	3	1	4	2	3	4	2	3	1	3	4
3	3		1	3	4	1	3	3	3		

Which of these activities prove to be most difficult, or time-consuming?

Interviews, Asses risks , Gathering and clean data, Defining the success criteria, Asses background, Data Cleaning, , Data Cleansing/ ETL, Can't compare the difficulty level of any step. Since, it depends upon the nature of each project. In some project one activity could be difficult or time consuming but in other project any other activity could be difficult or time consuming, Missing value treatment , Getting background information, Assess background of project, Nothing, , Conducting interviews, Interviews,Assess background to the project, Data processing, Define the success criteria, Writing a plan, Inexistent corporate doc, Assess background to the project, , Documentation, Stakeholder goals, Data integration across disparate platforms, Stakeholder interview, Risk assessment, Assess risks, Review company documentation, Project Background, , Define business problem, Interviews, Docs review, Writing a project plan , Assessing project back ground, Data prep, No idea, interviews, Conduct interviews with stakeholders, Assess risks, Define success criteria or acceptance criteria, Yes, Interviewing stakeholders, preparing data, Y

Do you use any tools and / or services to help define the business goals of the project?

No	No		No	No	No	Yes		Yes	No	No	No
No	Yes	No	No	No	No	No	No	No	No	No	No
Yes	No	Yes	No	No	No	No	Yes	No	No	No	No
No	Yes		Yes	No	Yes	Yes	No	No	Yes		

If you do, please mention which ones.

SAS, SDLC, Project Plan in Excel, KJ method, Excel, Kj analyses, SMART goals, Corporate KPIs, Excel sheets and flowcharts, In house tool, Affinity diagrams

Do you think some steps of the initial phase (i.e. Business Understanding in case of CRISP-DM) can be (semi-)automatized? If yes, which ones?

I don't know what that is, Data preparation and modeling, No, Not sure, Initial phase is the hardest to automate, even using AI, -, no, Business understanding with the help of data can be think for automatized. Through this the current situation of the business can be understand very well, Work flow, Not our organization, No, N, No, Not yet, No, It varies for each project, basically, as many as possible, Look at the current model baseline and understand the key factors and personas influencing the over all business environment in which the model is going to get implemented down the time. Maybe a bit of business case generation can also be automated, No, Not really, No, Understanding the motivation, Not sure, No, No, No, None, No, No not really, maybe putting findings into a DB and using for shared fields, No, No, Yes, risk assessment, No, Possibly, No, No, Don't know, Can't tell, Give more relevant articles to the subject, Discussion tracking, No, Don't know, No, No

At which point do you know that you know enough to define the business goals of the project?

I'm afraid we're not that structured as we're still a startup, It is iterative, so not really one step, Globally 10% in, Once KPIs have been established, Feasible project objects that are explainable to all members of the company, w, After the feasibility study, After understanding the working of the company as well as the difficulties an organization is facing in achieving it's desired outcomes, business goal of the project can be defined, Pre coluser, It varies, After multiple seesion of client meeting, Y, When I understand the business goals, Acceptance of the proposed analysis by the requester and my manager, When you have a fully understanding of what the business wants to set as business goals and how you can translate that with the information and company's system, The goal is never achieved. With each success, the customer's and management's appetite grows, After couple of detailed workshops with key business and technical stakeholders of the project, Usually after a few meetings with the client, After interviews, Early on, Initial phase, Stakeholder is happy with the presentation, After brainstorming, I am quite familiar with domain to begin capturing goals fairly early in discovery. , Middle I understand p&l implications of decisipn, Analysis, Data preparation, Use case development, Project has been defined in terms of decision making, When I understand how all of the dots connect, Initial discussion, I've done enough that i can assess business goals of the project from day 1, Assement of the problem, When I can produce a POC that resonates with most stakeholders, -, When the correct question has been formulated and there is sufficient ly detailed data to support finding the answer, After preliminary analysis of data, understand project, assess risks , define success/goals, Well. It's depends upon customers priority and timelines, Not yet, At defining success criteria, Depends on the project, Dont know

How would you envision the steps which should be made to define the data mining business goals at your organization?

I think we should start from the needs of users in terms of what they want to learn from data. Then, we should see if we have the data necessary to answer these questions or if we need to collect the data. Then, we should evaluate the costs for that, and the techniques that we should employ. Then, if all of this is reasonable, then we should proceed.

tech m

Understanding data, ETL of data, model, evaluación of models, respond business objectives and integration of the models

Do you think any of these steps could be performed (semi-)automatically? If yes, please mention which ones.

Yes, analyzing the business goal could be assisted by some software to help collect data about the needs. Also, some more or less automatic data mining techniques can be used for data analysis. For example, unsupervised learning algorithm do not require to provide as much information as supervised learning methods.

Yes

Yes, data preparation

Do you know of any tools and / or services, which could help with defining the data mining business goals? If yes, please mention them by name.

No. I am from academia. So I don't know exactly the tools used in the industry.

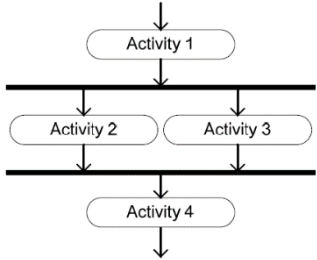

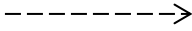

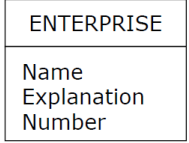
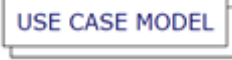


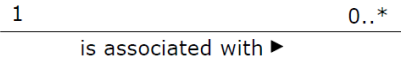

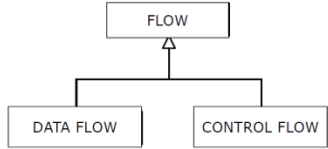
Sas

Dataiku

Appendix B

Process-Deliverable Diagram

Symbol	Name	Explanation
	Initial state	The state before any activity has been started
	Final state	The state after all activities have been finished
	Flow arrow	The flow
	Standard activity	An activity that can be performed
	Open complex activity variant A	A (higher level) activity of which the sub-activities are represented elsewhere in the model(s)
	Open complex activity variant B	A (higher level) activity of which the sub-activities are represented directly inside the activity
	Closed complex activity	A (higher level) activity of which the sub-activities are not represented in the model(s)
	Sequential activities	Activities that are performed in sequence
	Unordered activities	Activities that can be performed in any order

	<p>Concurrent activities</p>	<p>Activities that are performed concurrently</p>
	<p>Conditional activities</p>	<p>Activities that are performed based on a certain condition</p>
	<p>Link arrow</p>	<p>An arrow that links activities to its deliverables</p>
	<p>Concept</p>	<p>A deliverable that is instantiated from performing an activity</p>
	<p>Concept with properties</p>	<p>A concept that has properties</p>
	<p>Complex open concept</p>	<p>A (higher level) concept which has sub-concepts represented elsewhere in the model(s)</p>
	<p>Complex closed concept</p>	<p>A (higher level) concept of which the sub-concepts are not represented in the model(s)</p>
	<p>Association</p>	<p>A relation between two or more concepts, or one concept and itself</p>
	<p>Multiplicity</p>	<p>Indication of the multiplicity of instances of a certain concept</p>
	<p>Aggregation</p>	<p>Relationship between a concept containing other concepts</p>
	<p>Generalization</p>	<p>Express a relation between a general concept and a specific concept (e.g. All data flows are flows, but not all flows are data flows)</p>