

# URBAN SPACE AND WEB DATA

*Investigating Urban Leisure Activities using Web based  
Semantic Enrichment*

Demi van Weerdenburg

December 2017



**Utrecht University**

Cover image:

Skyline of Zwolle - Marnix Hägg (2015). Obtained from: <https://twitter.com/MarnixHagg>

Smart City Icons - CubicArt.com

# URBAN SPACE AND WEB DATA

## *Investigating Urban Leisure Activities using Web based Semantic Enrichment*

Demi van Weerdenburg  
4078004

This thesis is submitted in partial fulfilment for the degree of Master of Science  
in Urban Geography at Utrecht University

December 2017  
Utrecht, the Netherlands

*Supervisors*  
dr. Simon Scheider<sup>1</sup>  
dr. Bas Spierings<sup>1</sup>  
dr. Amit Birenboim<sup>1</sup>  
Hans Nouwens<sup>2</sup>

<sup>1</sup>Department of Human Geography and Planning, Utrecht University  
Utrecht, the Netherlands

<sup>2</sup> Stichting Geonovum  
Amersfoort, the Netherlands



**Utrecht University**



This research is conducted on behalf of platform Making Sense for Society\*, Geonovum during an internship from February until June 2017

\*A living lab for the Internet of Everything. This platform explores the use and potential of (sensor) data for the government, under which smart city concepts.



**Making Sense  
for Society** | Living Lab for  
the Internet  
of Everything





## Preface

When visiting a new city or country, you will likely search the Internet thoroughly for the best places to go, the most fun activities to do and where to eat the best dishes. Instead of just wandering around, our trips and leisure activities are increasingly more influenced by information provided by the Web. The importance of information and data is not only affecting you - currently, our cities and societies are becoming more digitalised than ever.

The past year was all about *smart cities* and data analysis for me. It caught my interest when I was involved in making a magazine about what smart cities actually are. It is an interesting concept which currently cities are using to label their city and ambitions. As part of the smart city; data and sensors are starting to play a bigger role in urban systems and policies. Hence why I found it really interesting to do my internship at Geonovum's Making Sense For Society platform to discuss and learn more about the governmental issues that arise within the smart city context. The subject of this thesis reflected on the use of web information that we daily use and create to orientate on what to do in the city.

First and foremost, I would like to express my gratitude to Simon Scheider for supervising me and helping me so much with the script that was used in my thesis, as this method was totally new to me. Furthermore, I would like to thank my other supervisors Bas Spierings and Amit Birenboim for your comments and valuable tips during the feedback sessions. Also, I am obliged to Ben Adams from the University of Canterbury for working on the multi-labelled version of the model.

I would also like to thank everyone at Geonovum and especially Hans Nouwens for an inspiring internship. I have learnt a lot about geo-information, standardization and sensors. I loved taking part of the ongoing discussion and especially enjoyed the enthusiasm you work with. I hope to take that with me in my future career.

Moreover, I am much obliged to everyone I have spoken along the way and who has given me valuable input or inspiration for my thesis. I found it interesting to hear how different parties are working with data and how everyone has its own idea about the city we are heading to. I also loved being advised to 'scrum' myself through the writing process. It took some time and sometimes it felt challenging, therefore I am thankful for my family and friends who were always there to support me. After almost a year of working on the subject, I still find the subject very interesting and I hope you enjoy reading this thesis as much as I have enjoyed working on it.

Demi van Weerdenburg  
Utrecht, december 2017





## Summary

In this thesis ‘*Urban Space and Web Data*’ was aimed at testing a method to allow enrichment of urban leisure-related activities based on combining place information with web texts. The value of web data to grasp and enrich place representations has been widely recognised (e.g. in Adams & Janowicz, 2014; Hobel, et. al., 2015 & Hu et. al., 2015). Nevertheless, most web data remains difficult to grasp when translating into a computational representation of place. Web data is acknowledged as a relevant source in understanding urban environments from a data-driven perspective. As leisure plays a bigger role in today’s urban economy, cities as Zwolle are aiming for being a vibrant and attractive city where citizens and visitors want to *be*. In doing so, it is their ambition to adopt more data-driven decision-making in order to know their guests better.

The following research question was designed: “*How does a web-based semantic model allow to extract and analyse urban leisure activities of people, combining place information with web texts, in the inner city of Zwolle?*”. In order to answer this question, four underlying sub-research questions were composed in the order of data-analysis.

The research focussed on testing a probabilistic topic model, Latent Dirichlet Allocation (LDA). Topic modelling is a tool to understand large amounts of texts and estimates a probability of potential activities into ‘*topics*’. There was aimed at both a single- and multi-labelled variant of LDA. The analysis was prepared by creating the ontology *Urban Leisure* in order to learn the model about related leisure activities. The used training data was manually collected from the Web, where unique web pages were selected from both sources created by institutions, such as the municipality and the tourist agency (VVV), and user-generated content (UGC). Eventually, 200 unique place names within Zwolle’s inner city were collected; in 20 activity classes, 105 sorts of referents and 62 different types of places. The major activity classes in the training data were *eating*, *drinking*, *shopping* and *watching*, which represented respectively 76% of the total data collection. This method considers to respect the ‘*voice of the consumers*’: it summarizes how people perceive the inner city of Zwolle (Guo et. al., 2016). The modelling and training involved a Python script, that enriches the data with different keys and evaluates the results with a ten-fold classification.

The results of the single-labelled LDA model have shown a considerable prediction quality between 60% and 68%. It succeeded a quite correct score of precision and recall, indicating that the model was considerably good at selecting the correct items. Nevertheless, the results of the multi-labelled LDA model have caused some difficulties and was considered to be not straightforward to compare with the single-label model. Regarding the resulting topics, both the single and the multi-labelled LDA models have summarized some topics decently while others seemed to be influenced by noise within the web texts. Therefore can be concluded that web-semantic research has proven to be adequate in summarizing how people perceive a certain place, but partially influenced by its too small sample size and limited obtained information in order for satisfactory results. When the results were visualized, the used method mainly focus on certain points of interest in the city while human conceptualization of ‘what to do where’ is based on a certain activity area (Hobel et. al., 2015).

The size and the topics of the training data have heavily influenced the results of the Latent Dirichlet Allocation topic models. Therefore is recommended to do further research in using web-based semantics to grasp human conceptualisation in computational representations with a bigger training dataset to filter out eventual noise within the data. There is also more research needed to overcome the multi-labelling issue regarding evaluation to get closer to a human perception of place as places are regarded as multi-functional areas (Hobel et. al., 2015). To make sense of the results in light of the data-driven policies, it is recommended to use web-semantic modelling as a qualitative data source to enhance city intelligence. Combining different datasets potentially enables the city to reflect on how people engage with their urban environment and therefore the city could use these insights making the city an attractive place.

***“You can have data without information, but you cannot have information without data”***

Daniel Keys Moran  
*Science fiction author & computer programmer*

# Contents

	<b>Contents</b> .....	<b>11</b>
	<b>List of figures and tables</b> .....	<b>13</b>
1	<b>Introduction</b> .....	<b>15</b>
2	<b>Theoretical framework</b> .....	<b>21</b>
2.1	Activity and place.....	21
2.1.1	Constructing place .....	21
2.1.2	Activities in time and space .....	22
2.1.3	Representing place .....	23
2.1.4	Earlier works on measuring activity and place .....	24
2.2	Leisure in urban context .....	25
2.2.1	Cities in the experience economy .....	25
2.2.2	Classifying leisure, users and areas.....	25
2.3	Web Data and the Data-driven city.....	27
2.3.1	The use of geographic information in leisure activities .....	27
2.3.2	Dealing with unstructured data .....	29
2.4	Conceptual framework .....	30
3	<b>Methodology</b> .....	<b>33</b>
3.1	Organizing unstructured data .....	33
3.1.1	Phases of data analysis.....	34
3.2	Ontology modelling .....	35
3.3	Preparing and enriching web data .....	37
3.4	Modelling LDA .....	41
3.4.1	Assumptions and supervision .....	41
3.4.2	Web-based enrichment: constructing training data .....	42
3.4.3	Cleaning text and model preparation .....	43
3.4.5	Classification: cross-validation .....	44
3.4.6	Substituting labelled LDA .....	44
3.5	Evaluation: validity, fitting and reliability .....	45
3.6	Visualising data .....	47
4	<b>Zwolle: a vibrant, leisure and data-driven city</b> .....	<b>49</b>
4.1	The vibrant city .....	49
4.2	The leisure city.....	50
4.3	The data-driven city.....	52

<b>5  </b>	<b>Results .....</b>	<b>55</b>
5.1	Collecting place affordances and data sources .....	55
5.1.1	Creating an ontology to describe urban leisure .....	55
5.1.2	Describing the training dataset .....	56
5.1.3	Reflection on collecting data .....	57
5.2	Training and evaluation of LDA with classifiers .....	58
5.2.1	Missing values and sample quality .....	58
5.2.2	Running the single-label LDA model .....	59
5.2.2.1	Resulting topics of model 1 .....	59
5.2.2.2	Resulting topics of model 2 .....	62
5.2.3	Evaluation of machine learning classifiers .....	64
5.2.3.1	Evaluation of model 1 .....	64
5.2.3.2	Evaluation of model 2 .....	67
5.3	Training and evaluating L-LDA .....	70
5.4	Visualising place affordance .....	73
<b>6  </b>	<b>Conclusion .....</b>	<b>77</b>
6.1	Answers to the research questions .....	77
6.1.1	Sub-research question I .....	77
6.1.2	Sub-research question II .....	78
6.1.3	Sub-research question III .....	78
6.1.4	Sub-research question IV .....	79
6.1.5	Main research question .....	80
6.2	Discussion .....	81
6.2.1	Volume, variety and velocity .....	81
6.2.2	Topic modelling and city intelligence .....	83
6.3	Implications and recommendations .....	85
	<b>References .....</b>	<b>88</b>
	<b>Appendix I - Encoding Sheet .....</b>	<b>96</b>
	<b>Appendix II - Machine Learning Classifiers .....</b>	<b>97</b>
	<b>Appendix III - Prediction quality of parameters and naïve classifiers for LDA .....</b>	<b>99</b>
	<b>Appendix IV - Classification Results LDA model 1 .....</b>	<b>100</b>
	<b>Appendix V - Classification Results LDA model 2 .....</b>	<b>101</b>
	<b>Appendix VI - Decision Tree LDA model 1 .....</b>	<b>102</b>
	<b>Appendix VII - Decision Tree LDA model 2 .....</b>	<b>103</b>
	<b>Appendix VIII - L-LDA Topics .....</b>	<b>104</b>

## List of figures and tables

Figure 1.1:	Data driven policy: from data to knowledge to execution.....	16
Figure 2.1:	The spatial triad of Lefebvre .....	22
Figure 2.2:	Accessible area within the space-time prism.....	23
Figure 2.3:	Circular action space - common for leisure activities .....	23
Figure 2.4:	The interrelated concept leisure.....	26
Figure 2.5:	Linkages between user, resources and region of activity.....	27
Figure 2.6:	Conceptual framework.....	31
Figure 3.1:	Flowchart of methodology.....	34
Figure 3.2:	Structure of Urban Leisure ontology.....	35
Figure 3.3:	Screenshot of the Grote Markt, Zwolle as shown on OSM.....	38
Figure 3.4:	Webpages describing the place affordance of the inner city I.....	39
Figure 3.5:	Webpages describing the place affordance of the inner city II.....	40
Figure 3.6:	The basic notions of LDA.....	41
Figure 3.7:	Extra information on OSM for restaurant ‘Bapas’ .....	43
Figure 3.8:	The principle of precision and recall.....	46
Figure 4.1:	Overview of Zwolle’s main attractions of the inner city .....	50
Figure 4.2:	Gastvrij Zwolle.....	51
Figure 5.1:	Visualisation of the ‘Food’ referent in Urban Leisure ontology.....	55
Figure 5.2:	Relative count of described activity classes within the dataset (N=326).....	56
Figure 5.3:	Visualisation of 18 topics derived from training ‘model 1’ .....	61
Figure 5.4:	Visualisation of 18 topics derived from training ‘model 2’ .....	63
Figure 5.5:	Topic 15, topic 0, topic 4 and topic 16 as discussed in the Decision Tree.....	67
Figure 5.6:	Topic 6 and topic 16 as discussed in the Decision Tree.....	70
Figure 5.7:	Labelled topic: Canoeing.....	71
Figure 5.8:	Labelled topic: Shopping.....	71
Figure 5.9:	Labelled topic: Eating.....	72
Figure 5.10:	Labelled topic: Watching.....	72
Figure 5.11:	Map of the place affordance of Zwolle’s inner city based on LDA topic modelling...	73
Figure 6.1:	Topic 5 and 8 (model 1).....	79
Figure 6.2:	Relevance of information resources in statistics.....	82
Table 3.1:	Open Street Map identifiers.....	37
Table 3.2:	Example of encoding of place affordance.....	41
Table 3.3:	Keys of enriched data file.....	42
Table 3.4:	Example of a document-term matrix.....	44
Table 5.1:	Top ten place and referent classes within dataset, absolute and relative.....	57
Table 5.2:	Different set parameters for each trained topic model.....	59
Table 5.3:	Places described by topic 8 in model 1.....	60
Table 5.4:	Cross-validated results and fit for Naïve Bayes classifier (model 1).....	64
Table 5.5:	Cross-validated results and fit for Logistic Regression classifier (model 1).....	65
Table 5.6:	Confusion matrix for Logistic Regression classifier (model 1).....	65
Table 5.7:	Cross-validated results and fit for Neural Net classifier (model 1).....	66
Table 5.8:	Confusion matrix for Neural Net classifier (model 1).....	66
Table 5.9:	Cross-validated results and fit for Decision Tree classifier (model 2).....	68
Table 5.10:	Cross-validated results and fit for Nearest Neighbours classifier (model 2).....	68
Table 5.11:	Confusion matrix for Nearest Neighbours classifier (model 2).....	68
Table 5.12:	Cross-validated results and fit for Neural Net classifier (model 2).....	69
Table 5.13:	Confusion matrix for Neural Net classifier (model 2).....	69



## 1 | Introduction

The 21<sup>st</sup> century is the century of the smart city. We are entering an era of *big data*: datasets of high volume, velocity and variety. Big data is often spatially and temporally referenced, which offers a new way of looking at geographical understanding (Kitchin, 2013a). Currently, urban authorities are testing the possibilities of data, data-driven policies and the *Internet of Things* to optimise urban processes. Authorities, businesses, citizens and people are generating enormous feeds of data in favour of urban policy and research. Governments are increasingly using big data to understand the city and “to better depict, model and predict urban processes” (Kitchin, 2013b, p. 2).

### Resilient and smarter cities

The inner city is often seen as the place where *it all happens* - the place where people live, work, shop and engage in leisure (Hall & Page, 2014). Traditionally, it is a multi-functional place and the prime centre of leisure activities in terms of quantity and variety. A changed consumption pattern of citizens and visitors and increased mobility has led to a renewed attention to the inner city in the past years (Aguiar & Hurst, 2007). Leisure-related activities have become one of the most important trip purposes into the city. Inner cities are focussing on attracting people to their city by branding their city as a *place to be* (Boersma & Raatgever, 2017). The city is being transformed into a place for fun and experience. This changing role of the inner city has implications for both its citizens and visitors. Nowhere the pressure of both societal and economic interest is as high as in the inner city. The report of the *Planbureau voor de Leefomgeving* (PBL; Netherlands Environmental Assessment Agency) on resilient cities states that different actors and interests in the city all have one thing in common: they aspire a fun, vital, accessible, livable and an attractive city (Evers et. al., 2015).

The change of inner cities to places about fun and experience is not entirely new: in 1998, Pine and Gilmore have introduced the concept of the *experience economy* to explain contemporary urban economies. People consume the city by looking for experiences (Oh et. al., 2007). The urban area has become the ‘*loci of consumption*’: an urban playground for visitors and citizens (Lorentzen, 2009).

In this light, new concrete challenges for the inner city are formulated. The city has to attract new functions other than retail. Cities should add *experience* into places where people want to *be*. Essential in this process is that cities need to be more resilience and cities needs to act smarter in doing so (Boersma & Raatgever, 2017). As most cities aspire to be fun, vital, accessible, livable and attractive, they facilitate enhancing and promoting inner cities as fun, vital and attractive places to be.

### The data-driven era

To be a vital and lively (inner) city, Vrolijk (2017) suggests that cities do not only have to know themselves better, they have to know their guests in order to improve decision-making. Already a lot of data is available of places in inner cities. They know where which business is situated. How often and on which moments people are in the city and how much money they are spending (ibid, 2017). Anthony Townsend (2013) describes the data-driven policy of cities as a historic shift in how we manage cities. Under the heading of *smart cities*, the use of data to tackle a wide range of urban problems has increasingly become popular (Meijer & Thaens, 2016, p.1; Townsend, 2013). Concepts such as ‘*big data*’, *sensors*, the ‘*Web 2.0*’ and the ‘*Internet of Things*’ are the newest forces that shape and sustain the urban environment and policy (Townsend, 2013, p. 3-4). Data creates a new layer around the city. The availability of data has acknowledged to have a potential impact on mobility and activity behaviour analysis but also in marketing, urban planning, health monitoring, travel- and transport geography (Rudinac et. al., 2017).

Currently, cities all over the world are experimenting with the use of big data in urban policy. Massive amounts of data about people's activities are used to determine the city. "Within the next twenty years, most of the data that we will use to understand cities will (...) be available in various forms" (Batty et. al., 2012, p. 488). Some will even state that without data, urban policies are on quicksand (Nouwens, 2016). Data-driven policies are focussing on collecting data and analysing information about the city, its surroundings and trends to gain knowledge and insights about the city. By executing these insights, data shapes the urban environment.

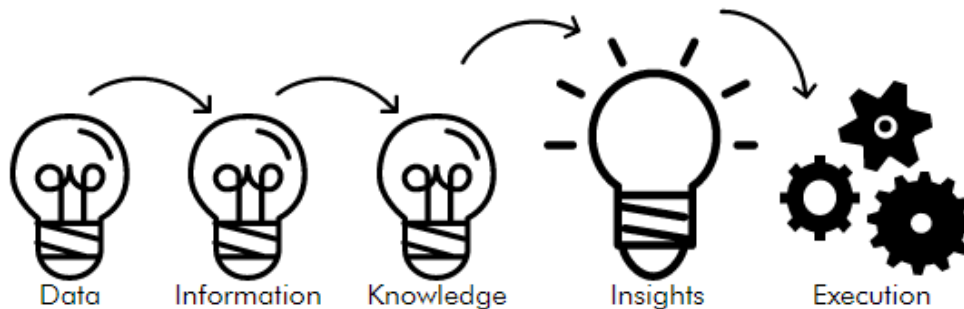


Figure 1.1 - Data driven policy: from data to knowledge to execution  
Based on Vergeer & Capelleveen (2017)

When looking at the available data on inner cities, a lot of quantitative data is available such as function descriptions of buildings, the number of flows in the city and the number of transactions. There is fewer, qualitative, data and information available about the use and perception of public space. As increasingly more people use the city for leisure purposes - and consume the city - the leisure industry becomes more relevant for urban policies (Boersma & Raatgever, 2017). Therefore it is relevant to research the potential of leisure in urban space. As it is the aspiration to be fun, vital, accessible, livable and attractive, the challenge within this data-rich era is trying to understand and analyse urban spaces with data sources in light of the behaviour, activities and experience of people.

#### The potential use of geographical Web information

Already a great amount of information describing geographic places is available on the Web in the form of websites, social media, user-generated content and other data sources. There is an explosion in interest of using the Web to create, assemble and distribute geographic information and knowledge about places, many referring to places with texts or pictures (Goodchild, 2007a; Townsend, 2013). Not only websites describe places; an increasing array of sensors are measuring the city and are connected to feed enormous amounts of data to the Web. Web information has become a basic notion for information retrieval: for both machine learning and for human interest. The potential value of geographic (big) data and information plays an important role in this new era: it forms the core in finding ways of explaining events and behaviour (van Oortmarsen et. al., 2014).

Place is tightly related to activity and the affordance of a place: the things which the environment enables to do (Hobel et. al., 2015). Hence why *place affordance* is a central concept in understanding the construction of places. Scheider and Janowicz (2014) state that place is deeply rooted in perception; perceptions which can be interpreted by Web information. Web users search and request for information about places and what can be done there. They search for information about retail, tourist attractions, accommodations, sports, entertainment, transport, public services and cultural heritage (Vaid et. al., 2005, p.1). As the Web becomes more open sourced and more based on *user-generated content*, information becomes more found on the perception of people itself. It is interesting to investigate how the potential use of different sorts of geographical Web information could enrich our assessment of human activity in space and time.



### Problem statement

Following the identification of a place based on the affordance, this study will concentrate on urban leisure-related activities in inner cities. The leisure industry is highly dependent on Web information retrieval and the available information determine people what to do (Költringer & Dickinger, 2015). The value of web data sources for enriching conceptual place representations has been recognised in several studies (e.g. in Adams & Janowicz, 2014; Hobel et. al., 2015 & Hu et. al., 2015). However, most web data remain unstructured, thus its meaning undefined, and is difficult to grasp. A method to translate this unstructured data into a computational representation of place is yet largely missing (Adams & Janowicz, 2014). Batty et. al. (2012, p. 488) further states: “*to interpret (...) data, we need to exploit and extend a variety of data mining techniques through which the visualisation of correlations and patterns in such data will be essential*”.

The main objective of this master thesis is to test a method that allows enrichment of urban leisure-related activities based on combining place information with web texts for urban areas of interest. The purpose of the research is twofold. First, to investigate how a topic model can be created to estimate the activity potential in public space based on web-based semantic modelling by using different kinds of web texts which describes urban areas. Eventually, this insights could provide assistance in the enrichment of public policy decisions regarding leisure space in indicating the spots within the city of high interest: places that attract people for an certain activity. It could optimise the understanding of the city as a vital and liveable place.

To investigate this research objective, the topic model will be tested on a particular inner city. The impact of (big) data has also gained interest in Zwolle. ‘Smart’ Zwolle wants to adopt data-driven innovations for issues and challenges within the city (iBestuur, 2016). Part of this ambition is the agenda for a vital, lively and attractive inner city that Zwolle has set up. In this agenda is the aim formulated for more research about how to measure behaviour and experience of people with data (Gemeente Zwolle, 2014a). Therefore, they focus on *city intelligence*: collecting data and analysing information about the city, its surroundings and trends to gain knowledge and insights about the city. Enriching current place information with web data provides an empirical basis for them to use in understanding the inner city as a place to be. The research question is thus as following:

*How does a web-based semantic model allow to extract and analyse urban leisure activities of people, combining place information with web texts, in the inner city of Zwolle?*

In order to answer the main question, four underlying sub-research questions are composed in the sequence within the research project. The first sub-research question has a more theoretical notion while the other sub-research questions focus on the methodological steps during modelling and analysis of the used topic model.

- I. *How can web-based text semantics be used for understanding and analysing activity potentials of people in certain urban spaces?*

The first sub-research question emphasizes the use of web-based semantics in geographical studies. The focus will be on the human perception of place and activity and how this can be conceptualised and represented. Hereby concepts as place affordance, leisure and web data need to be combined.

- II. *Which urban activity categories can be distinguished regarding urban leisure in the inner city?*

To answer this sub-research question is aimed to create an ontology. An ontology is a formal naming of different types and properties of urban leisure-related activities such as activity, referent and place type. This research question is relevant for inserting human language into a computational representation, which eventually teaches machine learning how to interpret bulks of texts.

- III. *Which web sources can be used to train a dataset for a web-text semantic model?*  
There are endless sources of data on the Web available, but which ones are useful to parse within the assumptions of the used model? This sub-research question therefore provides thoughts about data selection and the suitability of the eventually used training data.
- IV. *To which extent can urban leisure activities be estimated, extracted from knowledge from a web-text semantic model?*  
The final sub-research question aims to provide insights in the accuracy of the used model and focusses again on the relation between machine learning and geographical, human interpretation. How accurate are the results and what is relevant information for further use or within urban policies?

### Methodological background

The study will do a first step in exploring a method to analyse potential activity-based references of places. In essence, the used web-based semantic model is one of the most frequently used variant of topic modelling. This is a tool to understand a large amount of text by machine learning. It mines through web data and distributes a probability of topics, which in this case predicts a probability of potential activities. Topic modelling is widely used for investigating shared opinions and assisting in better decision making (Huang & Li, 2016).

The used models in this research are based on the *Latent Dirichlet Allocation* topic model (LDA), introduced by Blei in 2002. This is an algorithm that models documents in order of a mixture of topics, where each topic is characterised by a distribution over words (Blei et. al., 2002, p. 966). Therefore this model assumes that each word in a web text is generated from one underlying topic (Ramage et. al. 2009, p. 248 - 249). To conclude, the method aims to combine place information with web texts about urban areas to give enrichment of leisure-related activities in cities. The potential use could be summarized as determining the relevance of places for people and the *place affordance*, based on the activity potential as derived from the model.

### Relevance

The academic relevance of this thesis is built upon the gap between the growing production of web-based geographical information and the undiscovered potential of this data to enrich our knowledge about the relevance of places. The production of general available rich datasets has been recognised to allow new ways of understanding, analysing or generalising urban areas (Hobel et. al., p. 19) but also have raised the question how to conceptualise human representations in Geographic Information Systems (GIS). Different studies have explored methods to extract and analyse patterns of urban activity with web data. Hu et. al. (2015) have analysed Flickr photos to indicate hotspots in the city while Hollenstein & Purves (2010) did similar research by comparing the number of photos uploaded and the boundaries of the city centre. Photos were also used to discover spatiotemporal patterns of tourists in Amsterdam (van der Drift, 2016). Also in practice, sensors and other counting devices are monitoring the city in real-time. Web texts remains quite underexposed, while web texts are considered as a valuable, qualitative data source in understanding behaviour and activity in our cities. This is however not easily captured: *“a tremendous amount of place knowledge remains obscured from formal computational representation because it exists in formats designed for humans, not machine, consumption”* (Adams & Janowicz, 2014, p.2).

Within the scope of *place affordance*, the complexity of place is that people know what can be done where. The identification of a place is dependent on the perception of people, often characterised by lifestyle patterns of the urban crowd (Wakamiya et. al., 2011). Most research on understanding the relevance of place is done by surveys of human participants, which provide enough insights but are considered to be labour-intensive, time-consuming, rapidly outdated and not able to scale because of the fact place representations are complex (Hu et. al., 2015)

The identification of place is well-discussed within the field of human geography, while it poses methodological questions for large-scale computer representations about places for geographic information systems (Adams & Janowicz, 2014). GIS-based literature states that humans think and communicate in concepts such as *vague places*. These are cognitive, perceptual ideas about places but complicated to grasp into computational representation (Montello et. al., 2003). Most content about places resides in natural language, hence why current computational representations are mostly limited to facts such as population count, geographic locations and relations to other entities (Adams & Janowicz, 2014). Previous work in text extraction associates activities with place names but places could have multiple names or could have multiple types of activities that can be done, depending on the source or purpose of the information (Alazzawi et. al., 2012; Scheider & Janowicz, 2014). The relevance of place and how people use place is however fluid. Places are often determined by the potential of activities that can be carried out there because people know where a certain activity can be done (Scheider & Janowicz, 2014, p. 101). Hence why in this research an explorative research is conducted in how we can use web texts in enriching and representing our knowledge about places.

The societal relevance of this research enhances the ambition of data-driven urban policy in cities. The explosion of the availability of data creates new interest in the potential of data but also raises questions for cities. What are the new challenges of using data? Which data is useful? What do they miss? Dutch inner cities are driven by demand but since digital innovations, it is possible to act actively in real-time: in hours, minutes or seconds (Batty, 2013). Potentially this means the city has gotten a digital reflection which could lead to improvement of currently used systems but also leads to developing new services, business models and *evidence-based* policies. However, in practice, it seems that not all data is optimal to use for improving policies (Raatgever, 2017).

It is relevant to research the potential use how to enrich the knowledge of place affordances with web semantic modelling. As previous stated, the pressure on inner cities is high as the role of it has changed because of the rise of the new experience economy. It is both the ambition and the challenge to be a fun, vital, accessible, livable and attractive city. Edwards et. al. (2010) states that to cope with this kind of challenges, planners and policy-makers need detailed information about the whereabouts in the city. It is relevant to explore, next to quantitative and sensor data, qualitative data sources that describe the city. Place enrichment of leisure-related activities could be of value to develop further policies, for example by mapping hot- and cold spots of activity in the city. This information gives them the opportunity to reflect on how people engage with their urban environment in respect of their place affordance.

### Reading guide

To conclude, the objective of this research is aimed at exploring a way to model place affordance of a place by testing multiple data sources from the Web and test whether it is possible to model place by its human concepts. This thesis continues with chapter two, in which the most relevant concepts of place affordance, urban leisure and web data are examined. It ends with a visual representation that follows a synthesis of literature within the expectations of this research.

Chapter three elaborates on the research's methodology. The chapter justifies and explains made choices during the research. This will be described through the steps of data-analysis. Chapter four describes Zwolle which is used as a case study to apply the Latent Dirichlet Allocation topic model on. It follows a description of the inner city of Zwolle based on their ambitions to be vibrant and a leisure city. The municipality of Zwolle is aiming doing this data-driven, therefore is described how this ambition fits within the relevance of this research.

In chapter five, the results of the Latent Dirichlet Allocation topic models will be shown and evaluated. Afterwards, the last chapter is the conclusion of this thesis. It provides answers to the four sub-research questions and the main research question. This will be followed up by a discussion of what the results implies and how this can be applied in city intelligence. Hereby will be reflected on the application and insights of topic modelling and recommendations will be given for further use in urban policy and geographical research.



## 2 | Theoretical framework

As people use and experience places every day - the relevance of a place is determined by the daily interaction between people and space. People use space to move from A to B, people spend time in spaces and furthermore; they make decisions in and about places (Jordan et. al., 1998). Consequently, digital data is increasingly influencing the way people perceive place as they search for and retrieve information about possible activities that can be done at a certain place. This chapter follows an examination of three concepts: place affordance, urban leisure and web data. Place affordance and place representations in GIS will be discussed with respect to leisure-related activities within the inner city. Urban areas are often characterised by patterns of the crowd - indicating various places for working, drinking, eating, living, shopping or sightseeing (Wakamiya et. al., 2011, p. 101). Characterizing urban areas is seen as important for decision-making. As increasingly more people will use cities for leisure purposes and the leisure industry becomes more relevant for urban policies as well - it is relevant to research the activity potential of leisure in urban space. This theoretical framework further stress the emergence of new (big) data resources and how (big) data could be used for understanding and analysing activity potentials in urban space.

The main discussion is about how the perception of the crowd can be captured and analysed in the digital sphere. Computational place representations based on Geographic Information Systems (GIS) are often limited to statistical facts while the identification of a place is fluid (Adams & Janowicz, 2014). Relevant types of place exist because humans know where a certain activity can be done: the relevance of a place is determined by the potential activity (Scheider & Janowicz, 2014, p.101). Nonetheless, most content about what can be done where resides in unstructured data on different websites and in written words, difficult to grasp for machine learning.

### 2.1 Activity and place

Various studies have explained what makes a place important and how people assign a meaning to it. A fundamental concept is the principle of *space* and *place*. As Tuan (1977, p. 35) puts it: “*place is a space infused with meaning*”. Space is understood as a neutral container of coordinates, distances, topology and directions and treated as a static point on Earth (Gao et. al., 2014, p. 173). In GIScience, often the world is indeed represented by coordinates in representing space. However, space becomes a *place* when it is conceptualized by individuals or a group and becomes meaningful because of belonging activities and experiences. This ‘*patial*’ perspective focusses less on coordinates but more on place names, linguistic descriptions and semantic relationships between places (ibid). The principle of space and place forms the main difficulty in representing and understanding human perceptions and activities in a digital environment.

#### 2.1.1 Constructing place

Lefebvre (1991) has illustrated how space is constantly produced and contested by different users and planners. Not only the user itself constructs space, but increasingly the entrepreneurial city is becoming interested in constructing spaces as well in order to attract more people (Shim & Santos, 2016). In his spatial triad, Lefebvre explained how space is constructed by *spatial practice*, *representations of space* and *representational space*. Figure 2.1 shows the dynamics between how space is represented (lived) and perceived (practice) by its users and how it interacts with professionals, reflected in master plans and designs, compromises a produced space. Space is always socially produced and always in a process influenced by practices, perceptions and experiences. Understanding this interaction between users and planners is relevant in understanding how people interact with their environment and how governmental branches can act upon.

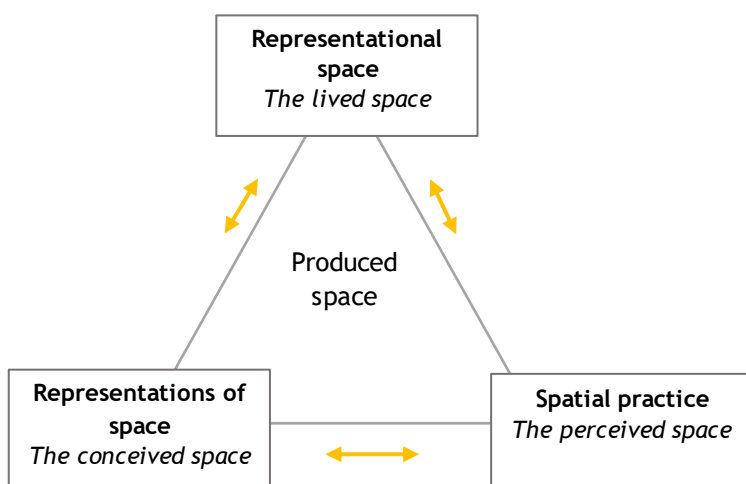


Figure 2.1 - The spatial triad of Lefebvre

Based on: Lefebvre (1991)

Place is therefore always in a *state of becoming*. It is more than a static and definable point on Earth: it is related to cognition and attachment of individuals. In GIScience is problematized how this concept and its duality can be represented in a geographic information system (GIS). A GIS presents a simplified view of the world by translating space into a series of basic spatial entities such as lines, points and areas. These entities are attached to coordinates in a two-dimensional world (Heywood et. al., 2011). Therefore raises the question how to conceptualize and digitalise perceived space into such a two-dimensional system (Jordan et. al., 1998).

### 2.1.2 Activities in time and space

Place is tightly related to activity and behaviour because people *do* things in places (Hobel et. al., 2015; Scheider & Janowicz, 2014). This makes human activity a central concept in constructing place: “many relevant types of places (...) exist only because we know that some activity can potentially be carried out there” (Scheider & Janowicz, 2014, p. 101). Linking back to the spatial triad of Lefebvre, the spatial practice of a place plays an important role in producing space.

The relevance of place is determined by perception and experience but also by the activities that take place there. Studies of human activity patterns in space have emerged in the 1960s through the use of activity-travel diaries and time-budget studies, but have gained renewed attention due to a significant wave of IT and data related research in human geography. With the increasing availability of geo-referenced data and innovations in GIS, it is more likely to operationalize and implement time-geographic constructs in the analysis (Kwan, 2002). Information technology could eventually lead to new dimensions of understanding of urban areas, patterns of human activities and enhances the understanding of everyday life in relation to the local geographical and social context (ibid, p. 476).

Activities are not restricted to a single point in space: they have an extent in time and space (Hobel et. al., 2015). One of the first studies to human activity patterns is the classical work of Alexander Szalai. In *The Use Of Time* (1972) examines activities as the outcome of pastime because “time can only be spent, not ‘earned’” (ibid, p.1). Conducting activities is basically using time. A time budget is a systematic record of the use of time during a given period and mostly describes the temporal characters of the activity as duration, frequency, timing and the sequential order. This particular book still has an influence on more recent research on activities and time-use (Kwan, 2002).

In addition to time-budget and the temporal character of activity, the spatial character and the distribution of activities are evenly important. People tend to behave differently depending on where they are. The field of time geography provides an understanding of spatial and temporal behaviour patterns. The origins of time geography lie in the work of Torsten Hägerstrand (1970), and analyses human movement and activities in space by combing time as the third factor. Hägerstrand assumes that activities are constrained by space and time. Space and time constraints activities such as sleeping at home and having a meeting at the work office. Other activities are determined by

opportunities in time and space. Opportunities in time geography are visually designed in a *potential action space*, or *activity area* (see figure 2.2). The potential action space is the area in which potentially activities can take place in a certain matter of time, represented by a space-time prism.

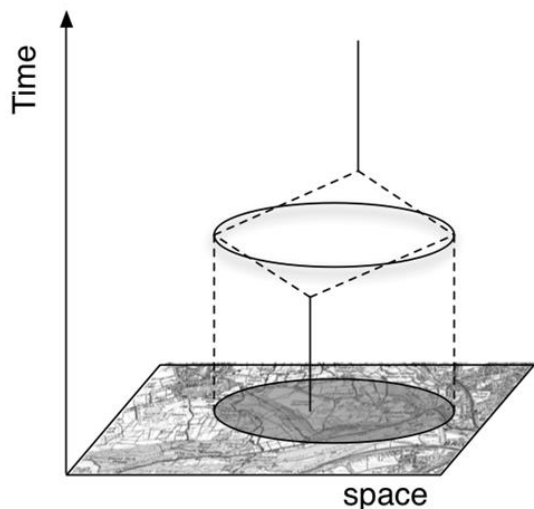


Figure 2.2 - Accessible area within the space-time prism  
Dijst (2009, p. 270)

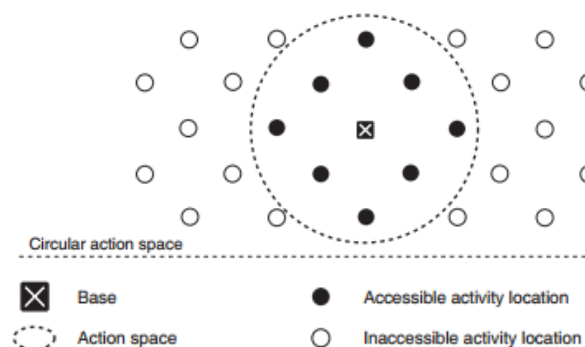


Figure 2.3 - Circular action space - common for leisure activities  
Malleon & Birkin (2014)

The work of time geography shows how space and time are related to the potential activity that can be carried out in space. The activity is dependent on a certain area in where the activity can be conducted. This spatiotemporal accessibility determines activity potential: what is one able to do within a certain space within a certain time? This can be described as *time affordance*. Dijst (2009) differentiates three forms of space-time prisms (or frames) namely circular, linear or elliptic. Leisure-related activities can be mostly understood from a circular action space. The activity mostly starts and ends at the same base such as a train station, parking spot or accommodation (see figure 2.3). The activity is thus the result of a relation between space and time: “*human action is not necessary situated, it occurs in a context*” (Hägerstrand, 1970, p. 35 - 36).

### 2.1.3 Representing place

It is recognized that place is a concept deeply rooted in perception and that activities are related to a certain action place in which the activity can occur. This has led to the main question in geographical information science how to conceptualise human representations of place.

Today's most common strategy is representing place by *points of interest* (POI). POI is associated with an individual location such as a restaurant, museum or a place worth seeing (Hobel et. a., 2015). It is basically attaching an attribute to a point. This is widely adapted in navigation software and in tourism information. According to Kang et. al. (2006), people who are planning to visit a city are provided large amounts of information but cannot easily process this information in a short time, hence it is easy to process information into points of interest. For representing the place and the potential activities that can occur in the context of a place, such way of mapping does not always match the perceptions of the users. Points are absolute locations on the Earth's surface while activities are not restricted to one single point in space. Hobel et. al. (2015, p.20) identifies three problems with POI:

- “People typically conceive place as a region;
- Different persons tend to associate different spatial footprints to the same place;
- Conceptualization of a place relies on the activities that are possible to carry out at that spatial location - referred as *place affordance*.”

Hobel et. al. (2015) argue that activities are not bound to a single point in space but that people relate an activity within the region which enables them to pursue the activity. In relation to time geography, the activity occurs in a context. Within that context, multiple footprints can occur. This can be related to time. Therefore Hobel et. al. (2015) argue that the spatial accessibility of places should be represented by its *place affordance* - the things which an environment enables one to do and how people perceive their environment (Jordan et. al., 1998). The affordance of a place creates potential activities for users. The concept, derived from ecological psychology, is used in GIScience towards a more intuitive mapping and brings people's experiences with the real world into computational representations. It can be applied in multiple domains such as tourism, city planning and transport analysis. In its potential, it relates interesting activities per domain to eventually reveal regions based on a certain interest.

An example from the research of Hobel et. al. (2015) shows that shopping not only involves actually buying goods but can involve sitting down at a café or going to the nearest ATM to withdraw money. This is, for example, a conceptually homogenous area which affords people being able to pursue activities such as shopping. Human perceptions of places are in GIScience closely referred to *vague places*. Montello et. al. (2003) states that people think and communicate in terms of vague concepts, which do not have precise references or sharp boundaries and therefore are fuzzy and probabilistic. People use language to explain spatial relations (e.g. *near*, *around*) or use vague boundaries as '*downtown*' based on their cognitive or perceptual ideas of regions. Therefore Hobel et. al. (2015) argue that places should be presented as an *area of interest* (AOI) and in a polygon instead of a point which enables new functions and makes information search more accurate according to human perception.

#### 2.1.4 Earlier works on measuring activity and place

Existing work on analysing human activity and place especially have focussed on activity patterns, the spatiotemporal aspects of activity and how they perceive their regions (Huang & Li, 2016). Investigating how people use their time explains patterns of leisure (Crosbie, 2006; Evans & Jones, 2011). Self-administered activity diaries are seen as a major method to capture time use, while other methods such as trajectories and walk-along interviews focus more on mobility and the embodiment of place (van Duppen & Spierings, 2013; Evans & Jones, 2011).

Using activity diaries is retrieved from time allocation studies as Szalai (1972). Basically, activity-based diaries focus on how people allocate their time during the day and is particularly focussed on the conducted activity during the day. By using activity diaries, it is possible to create space-time paths. Nevertheless, it is questioned whether this method is reliable in researching activity patterns because the relation with place remains underexposed.

Human movement trajectories and walk-along interviews focus on people's understanding and embodiment of place. A small number of geographers is using walk-along as interview technique to investigate attitudes and knowledge about a certain place. Engaging with place is seen as a major advantage of the method, as participants are more familiar with the place and more spatially focussed (van Duppen & Spierings, 2013; Evans & Jones, 2011).

The considered methods focus mainly on exploring, understanding and or modelling activity patterns of an individual in a spatiotemporal dimension (Huang & Li, 2016). Certainly, more standard methods such as surveying and interviewing could be used for measuring urban activity. Surveys stress larger research samples than interviews but face difficulties such as time constraints. Nonetheless, different methodologies, both quantitative and qualitative are eventually complementary. This thesis will focus on machine learning as a method to analyse place affordance by using unstructured data on the Web. The difference with existing methods is that machine learning can address a varied amount of data to reveal the perceptions of the crowd. This method addresses on the 3 V's of big data: *volume*, *variety* and *velocity* of data (further discussed in paragraph 2.3). A special interest in big data lies within the domain of leisure - whereas different studies focus on the spatial footprint of people within the city (Hu et. al., 2015).



## 2.2 Leisure in urban context

The city has always been a place of consumption, production and leisure - but nowadays the *leisure city* has become a focal point of a competitive branding strategy. Leisure has become part of an ambitious re-branding and redevelopment of consumption spaces in inner cities to make the city competitive on a regional or even (inter)national scale (Thibault & Lavigne, 2014; Spierings; 2009). The matter of this paragraph stresses how leisure has changed the city and produced the city.

The definition of leisure, as given in the Dictionary of Human Geography is “*either the freedom from doing some things or the freedom to do other things*” (Gregory et. al., 2009, p. 416). Both indicate a state of pleasure, amusement and enjoyment alongside choice. Leisure activities emerged from reductions in the working week and an increase in the paid holiday time.

However, the form of leisure activities themselves has changed. Originally, in the time of the book of Szalai, leisure is seen as ‘*free time*’: the portion of individual time left when obligations as work and household were fulfilled (Robinson et. al., 1972). Contemporary studies acknowledge that considering leisure just as spending time would be empty of its content. Leisure is more than time and is described as the cross-over of activity, time and experience. It is seen as a meaningful activity (Roberts, 2006): leisure is an important part of an *economy of experience* in which people are seeking for *moments* (Gregory et. al., 2009; Lorentzen, 2009). Understanding activity in relation to place could enrich the understanding of urban life and experience. Studying leisure activities in relation to place eventually answers the question ‘*which leisure activities can you perform there?*’

### 2.2.1 Cities in the experience economy

Cities have a long history of producing and consuming leisure. Since the evolution of cities into places where people live, work, shop and engage in leisure - cities became places with various leisure events and consumer spaces in order to be competitive (Spierings, 2009). Urban areas compete in an urban hierarchy characterised by an intense competition to attract and ensure investments, jobs and visitors. Therefore urban authorities had to create economically and symbolically important new consumer spaces. In this era, the city is defined by and through consumption (Jayne, 2006, p. 3). Leisure-related activities have gotten a new significance in the post-modern city as the result of several (societal) processes such as the new division of labour, increased mobility, the transformation into the 24-hour city and accompanying urban regeneration. Changes in leisure customs have led to an increasing complexity in leisure patterns (Hall & Page, 2006).

A keyword in current urban policy and economy is ‘*experience*’. Pine and Gilmore (1999) introduced the concept *experience economy* whereas people now see the memory itself as a commodity (*the experience*). In essence, what people primarily seek and consume at places is engaging in experiences accompanied by the goods or service that is offered at the destination. Leisure activities are a form of human experience and have become a major aspect of economic development and government responsibility (McLean & Hurd, 2014; Thibault & Lavigne, 2014). Destinations are positioned as experiences (Oh et. al., 2007). The consolidation of leisure has led to a debate whether such activities should be a municipal enterprise. City governments are key actors that focus on enhancing and promoting these memories in their city. Leisure has reshaped the city, by constructing and increasingly focussing on creating the inner city as a place where people wants to be (Crouch, 2006; Hall & Page, 2006).

### 2.2.2 Classifying leisure, users and areas

The city has been reshaped into a place where people engage in all kinds of leisure activities such as shopping, strolling along bars and cafés at day- and night-time, visiting touristic attractions, doing sightseeing or visiting festivals and events. The (inner) city area provides leisure functions for various visitors regardless the prime motivation of visiting (Hall & Page, 2006). In literature, terms like leisure, tourism and recreation are widely used as complementary, related and overlapping concepts. Hall & Page acknowledge that recreation and tourism are both part of the wider concept of leisure.

Figure 2.4 shows how the terms are related. In general, leisure is distinguished from work-time as free time. Two exceptions are given, namely business travel and serious leisure. Within leisure, you find both tourism and recreational activities. This indicates that leisure, tourism and recreation are interrelated and overlapping concepts. Based on this figure, the *leisure city* is both used for tourism and recreational activities.

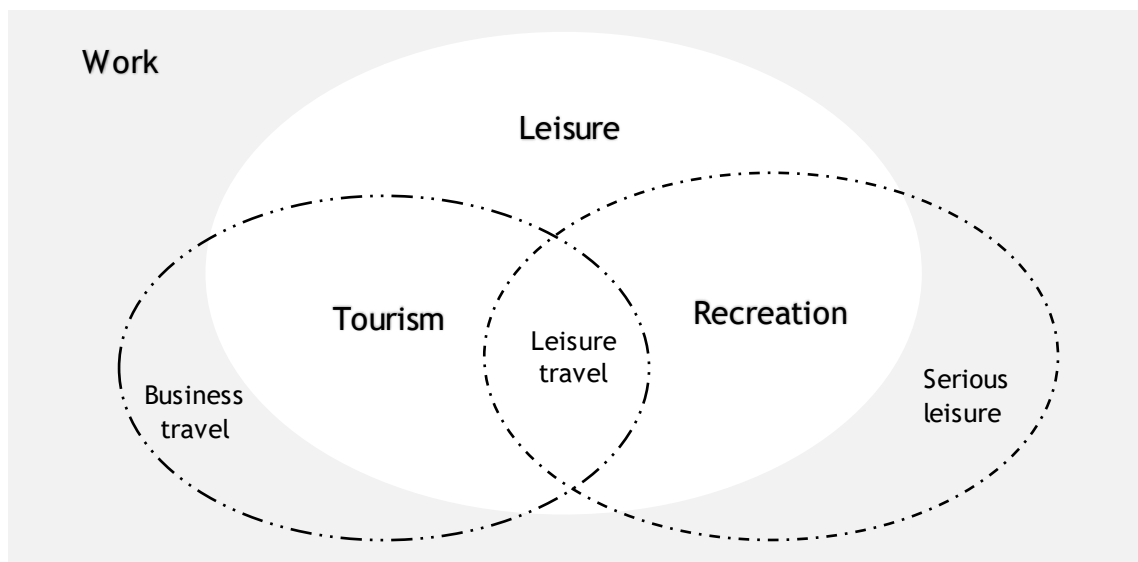


Figure 2.4 - The interrelated concept leisure  
Hall & Page (2006)

Regarding recreation and tourism in the city, Ashworth & Page (2011) have identified that the distinction between the citizen and visitor (or tourist) have been blurred as they use the same facilities, resources and environment: “*tourist make use of almost all urban features, they make an exclusive use of almost none*” (Ashworth & Page, 2011, p.3).

It is typical that the leisure city is consumed in a sort of bundles. Ashworth and Page (2011) acknowledge that people are attracted by urban features and the *urban way of life* but still the city is used in clusters such as *culture, history, business, sports, gastronomy, night-life* and *shopping*. Activities take place at a particular location because of certain characteristics such as the history, traditions, built environment, facilities or the atmosphere of the place (Lorentzen, 2009). Thus leisure activities are inherently spatial: leisure activities are produced and consumed in space (Crouch, 2016; Johnson & Glover, 2013). Often when urban areas are represented, the city is represented in a significant theme which indicates a certain place affordance and produces a certain experience for its users.

Figure 2.5 shows a model which was developed by Burtenshaw, and applied by Hall and Page that argues the multifunctionality of urban areas. It shows different activity profiles based on the type of activity (resources), city type (regions of activity) and user. It is clear that different users make use the same resources: both residents, local visitors and business delegates could visit a cinema or theatre. Furthermore, it shows how the ‘*Tourist City*’ overlaps with the ‘*Nightlife City*’ and the ‘*Leisure Shopping City*’. Urban places are therefore multifunctional areas with multiple types of activity and multiple areas, which complicates identifying areas of leisure. In addition to the application of Hall and Page, there has been a line added from ‘*tourist*’ to ‘*shops and consumptive activities*’ because leisure shopping is nowadays acknowledged as one of the most important activities in tourism (Kemperman et. al., 2009). The assumption based on this graph is that there is almost none differentiation between different users: the multifunctionality of place overlaps different consumption patterns and leisure activities as all urban features are used for different and overlapping leisure activities (Ashworth & Page, 2001, p.10).

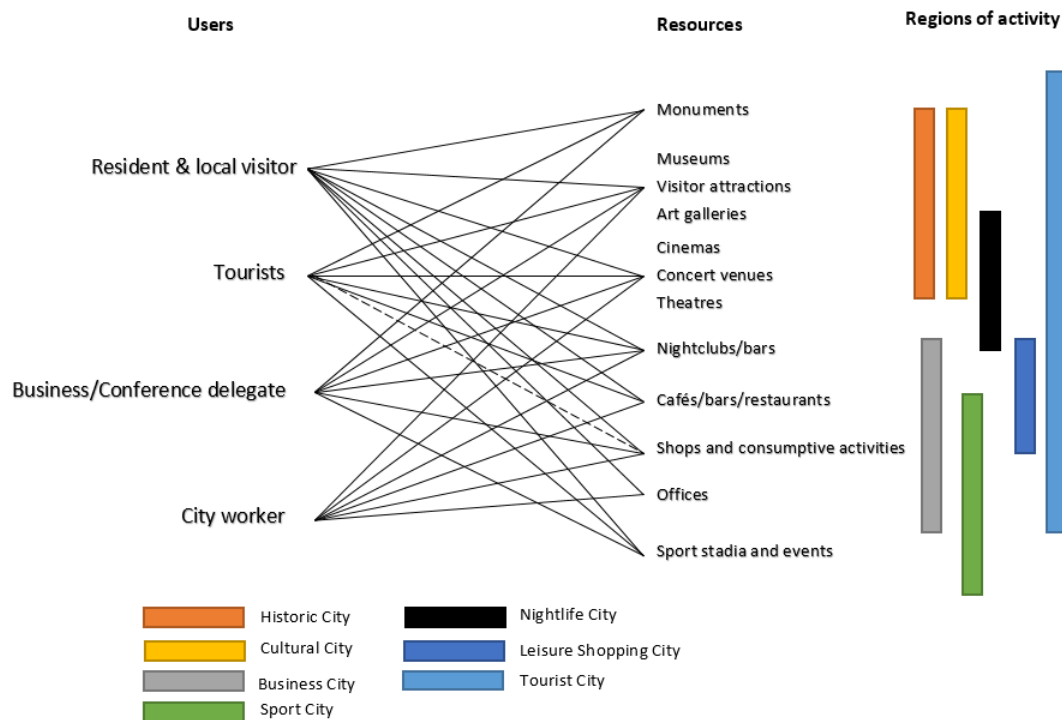


Figure 2.5 - Linkages between user, resources and region of activity

Adopted version from Burtenshaw et. al. (1991); Hall & Page (2006)

To conclude, urban leisure is in this thesis defined as a (meaningful) activity done in the free-time, after obligations such as work and the household, outside the house in a state of enjoyment, making use of the features of the urban environment. This includes both tourism and recreation done by either visitors or citizens as they make use of the same features and an exclusive use of none.

## 2.3 Web Data and the Data-driven city

The city offers a wide range of activities, in different types, for different users. Within this increasing complexity, it becomes inconvenient for users to have a complete overview of the activity potential within the city. Defining urban areas and giving information about the characteristics is crucial for making geographic decisions because everything we do takes place in space (Wakamiya et. al., 2011). Humans retrieve, organize and use information for their daily behaviour: information seeking determines what people are going to do. Consequently, seeking for and retrieving geographical information has become a commodity itself (Fabrikant & Buttenfield, 2001).

### 2.3.1 The use of geographic information in leisure activities

The way how users of urban space search and use information have changed. Different researchers acknowledge how, especially in leisure, web information became the most important source for planning activities (e.g. in Akehurst, 2009; Költringer & Dickinger, 2015). The internet became the most important source for answering the question *'what can be done there'*? Furthermore, web information became in this field one of the most valuable communication channels (Akehurst, 2009). An increasing number of web pages consist of information in texts, pictures and multimedia about places. The emergence of Web 2.0 and user-generated content have caused a worldwide sharing of place representations (Költringer & Dickinger, 2015). Improvements in the storage capacity, distribution and processing of digital data have caused an explosion in generating and publishing data on the web, also describing places. It provides a repository of perceptions, additional information and reports of places which potentially influences people what do to while being at a place.

The emergence of web information has become reliant on leisure because of the increasing demand of consumers who request information. The participatory internet influences how people gaze upon a certain place (Magasic, 2016). The image or perception a certain area has is influenced by what is shared on the Web. People are inspired by activities they read about online, creating a vicious circle of sharing and retrieving information. These developments have led to a new strategic management for leisure-related actors as the existing information causes a gaze upon what people actually are planning to do (Költringer & Dickinger, 2015; Magasic, 2016). The benefits of web information usage are unprecedented: from private to the public sector, from scientists to citizens. Kitchin (2014, p. 129) even suggests that data-intensive research has reached into a new paradigm in science with a focus on exploratory research, statistical exploration and data mining. Geographic data and information play an important role in this new era - it forms the core of explaining events and behaviour (van Oortmarssen et. al., 2014).

The availability of web data is constantly growing: roughly 90 percent of the web data that currently exists is not older than two years (ibid). This triggers a growing interest in the use of web information in urban spaces (Jones & Purves, 2008). Van Oortmarssen et. al. (2014) states that geographic information has been changed from just '*a spot in the atlas*' into a culmination point of data. Users of the Web search and request information about, for example, services related to retail, tourist attractions, accommodation, sport, entertainment, transport, public services and cultural heritage (Vaid, 2005, p.1). The retrieval and relevance of geographic information have been changing over the decades and can be distinguished in three sorts of data from different decades: *institutional web data*; *volunteered geographic information* (VGI) or *user-generated content* (UGC) and the emergence of newer "big" data sources.

### 2.3.2 Datafication: changing cities, changing Web

As the way we manage cities is changing so is the Web. In perspective, web data is changing because of new involvements of the users. In the early days, the Web was primarily one-directional. The content was created by a few with a majority of users simply consuming information. Information was very top-down regulated (Cormode & Krishnamurthy, 2008). The way how users could retrieve information is comparable to going to the library. This Web is a retronym compared to what today is described as Web 2.0.

New developments in the early 2000s allowed users to interact and provide information to websites (Goodchild, 2007a). Except for retrieving information, the Web allows users to participate. The new development consists interactive features, is user-friendly and mobile. This interactive content is defined as *user-generated content* (UGC): users are able to generate and collect data for instance with a smartphone. A special type of UGC is volunteered geographic information (VGI) (Goodchild, 2007b). This type has a geospatial component and has become the richest source of geographic data - in gazetteers <sup>[1]</sup>, in social media, blogs, news forums and more (Gao et. al., 2017), especially in leisure where it is used to share information with potential new users and visitors. Anyone with an Internet connection can select an area on the Earth's surface and provide it with own descriptions and links to other sources. Some descriptions are extensive and include hyperlinks, others only describe its features within the city. Two examples described by Goodchild (2007b) are geotagged photos on the photo-sharing website Flickr and the open platform Open Street Map (OSM). Many places which are referred by users are not specified by official gazetteers. This implicates that UGC adds to a rich and generous set of information with the potential to benefit scientific research and decision making because of semantically rich new information about a certain place (Goodchild, 2007a). User-generated content is thus considered as a platform for rethinking cities from the bottom up (Townsend, 2013, p. xiv).

---

<sup>1</sup> a geographical index or dictionary

### Volume, variety and velocity

The era of datafication implicates that web data have become an important input for innovation in cities. There is a drive for data in order to measure different components of the city (Marshall, 2012). Big data is a concept that describes the phenomenon of big datasets (both structured as unstructured) on different aspects of both environment and society which are created by millions of people and their smartphones, wearables, GPS and other digital sensors. Big data is not only considered as 'big' because of its volume - also because its variety and velocity. Big data is often described as following (Kitchin, 2013, p. 262):

- Huge in volume: consisting of terabytes or petabytes of data
- High in velocity: being created in or near real-time
- Diverse in variety: being structured and unstructured in nature

The importance of big data in cities is emerging. The sources of big data already include both directed and automated data in forms of digital devices, sensors, scans, interactions, volunteered data, social media, crowdsourcing and citizen science (Kitchin, 2014). The rise of big data has been spurred by the collection of data on activities in which humans are involved in, under which leisure activities (Batty, 2013). Batty argues that big data will be the new understanding of how cities function. Automatic, routinely (real-time) use of data will make the city plannable and predictable by "*minutes, hours and days rather than years, decades or generations*" (Batty, 2013, p. 276). This has caused the rise of the '*Internet of Things*' (IoT): different sensor objects (things) are able to interact with other objects via the internet through data.

It is likely that often sensors and sensed data will be associated with using big data but there are other datasets considered as big data as well (which needs new algorithms and measurements). For instance, generated by human responses on the Web. The challenge that all this developments and increase in data resources brings is how to make use of this large datasets. Kitchin (2013) states that within the field of human geography it offers the potential for insights in our social and spatial world. However it also poses challenges and risks - traditional methods (as discussed in paragraph 2.1.4) are designed for small and scarce data and the human geography discipline is unprepared in methods and theories about the use of big data in our cities and surroundings. The 'hype' surrounding big data suggests that it is superior to traditional studies - which is risky. The interest of policymakers is too quickly shifting to big data. However as "*big data may (...) be exhaustive, but as with all data, they are both a representation and a sample*" (Kitchin, 2013, p. 265).

#### 2.3.2 Dealing with unstructured data

Big data generally captures expressions such as what has been typed, swiped, scanned or sensed (Kitchin, 2013, p. 265). Thus big data describes action and behaviour. The difficulty of these voluminous datasets is that they are weaker in capturing complex concepts such as emotions, values, beliefs and opinions and the way people interact and make sense of the world (ibid). The array of web-based geographical information concerning leisure-related activities is predominantly based on text. It describes the city in every day's language, considered as a *vernacular language*. *Vernacular geography* considers the spatial knowledge which resides in the minds of people thus "*the body of knowledge that people have about their surrounding world*" (Hollenstein & Purves, 2010). Such notions are becoming of interest in information search and travel planning as it resonates the *voice of the consumer* (Guo et. al., 2017).

Therefore the problem rises that place related queries are treated the same as other search terms. The resulting query is lacking geographical relevance (Vaid, 2005). An often used solution is geotagging: the process of adding geographical information to any unstructured digital content such as text, photos and other not pre-defined sources. By doing so, users produce their own location information (Goodchild, 2007b). Nonetheless, not all content is geotagged and not all content is geotagged consistently.

The complexity and ambiguity of place have posed a 'riddle' for geographic information systems (Adams & Janowicz, 2014, p.2). This has caused explorations in methods to extract and analyse patterns of activity and characterizing urban areas with semantic research. Hereby is acknowledged that user-generated content can be used to understand the perceptions of users about a certain place. The prevalence of geographic information on the Web has increased explorative methods trying to grasp and understand human languages and perceptions, in machine learning indicated as *natural language*. For instance, in the study of Hu et. al. (2015), Flickr photos were analysed to indicate areas of interest within the city, by assuming the most photos of certain places is an interesting place in the area. Hollenstein & Purves (2010) did a similar research by comparing a number of photos uploaded and the boundaries of the city centre. They all question how to identify *vague* concepts and the multifunctionality of place in geographic information systems. This example merely focus on grasping and understanding human concepts with pictures.

In conceptualizing *vague places*, vernacular language plays an important role. Current developments on the Web try to let computers learn our natural language (*Natural Language Processing*). Computers cannot reason on bits of information available throughout the web. As an example, humans know that a rose is both a flower and a plant; computers need an ontology to understand (Bekel, 2008; Slaghuis, 2009). Hereby semantics can be defined as '*learning the meaning of*'. An important issue is that the meaning of entities and relationships between entities are available throughout the Web instead of on one single web page or application. Adams and Janowicz (2014) therefore suggest that more semantic techniques have to be developed for better evaluation of the state of semantic modelling with unstructured data.

## 2.4 Conceptual framework

Deriving from the theory is the following conceptual framework which expresses the expectations within this thesis (figure 2.6). The conceptual framework helps to focus the research and in answering the main research question: '*How does a web-based semantic model allow to extract and analyse urban leisure activities of people, combining place information with web texts, in the inner city of Zwolle?*'.

The chapter followed an examination of the concepts of place affordance, urban leisure and web data and how they influence place representations in geographic information systems. The concept of place is tightly related to activity because people do things in places, which is in this case leisure related (Scheider & Janowicz, 2014). But not only the activity itself constructs place - relevant places exists when people know what can be done there referring to a potential action space. Therefore, place is determined by both perception based on experiences and the activities that take place - which means that place is influenced by its *place affordance* (Jordan et. al., 1998). This is visualised in the left part of the conceptual framework. Because humans know what to do at a certain place, they have a representation of the place and knows about the place affordance and activity potential of that particular place.

The problem central in this thesis is that a tremendous amount of place knowledge is available on the Web in various data forms - but remains hidden for computational representation because it is designed for humans instead of machine consumption (Adams & Janowicz, 2014, p.2). People know what can be done at a certain place - or retrieve this information from the Web. Web information has become the most important source for planning activities, especially since the emergence of user-generated content (Költringer & Dickinger, 2015). It determines what people do in space; indicating a vicious circle of constantly sharing and retrieving information about places and its affordance (Fabrikant & Buttenfield, 2001). Roughly 60 percent of all data online contains geographical information about a place (van Oortmarssen et. al., 2014). This information is mainly unstructured data because it is undefined in use for machine learning and often in natural language. In this thesis will be focused on web data in various types: data from institutional sources (top-down), user-generated content (UGC) and the upcoming field of other big data sources. All these different data resources will be tested in use for probabilistic topic modelling.

The focus of the methodology is pointed out with the grey frame, indicating its explorative character whether a model can be created to estimate the activity potential of a public space and describe how information on the Web is describing what can be done in the inner city of Zwolle.

Eventually, in light of the experience economy, the results could optimize the understanding of place representations of people and human behaviour in place. Space is an active element in producing the city. In this data-fied era, space becomes increasingly mediated and enhanced by code, algorithms, sensors and data (Kitchin & Dodge, 2011). Batty (2013) has argued that this will cause a new understanding of how cities function. Policymakers could use the insights to reflect and improve their policy and enhance the experiences in their city. This evidence-based policy potentially influence how city planners and architects *conceive* the city.

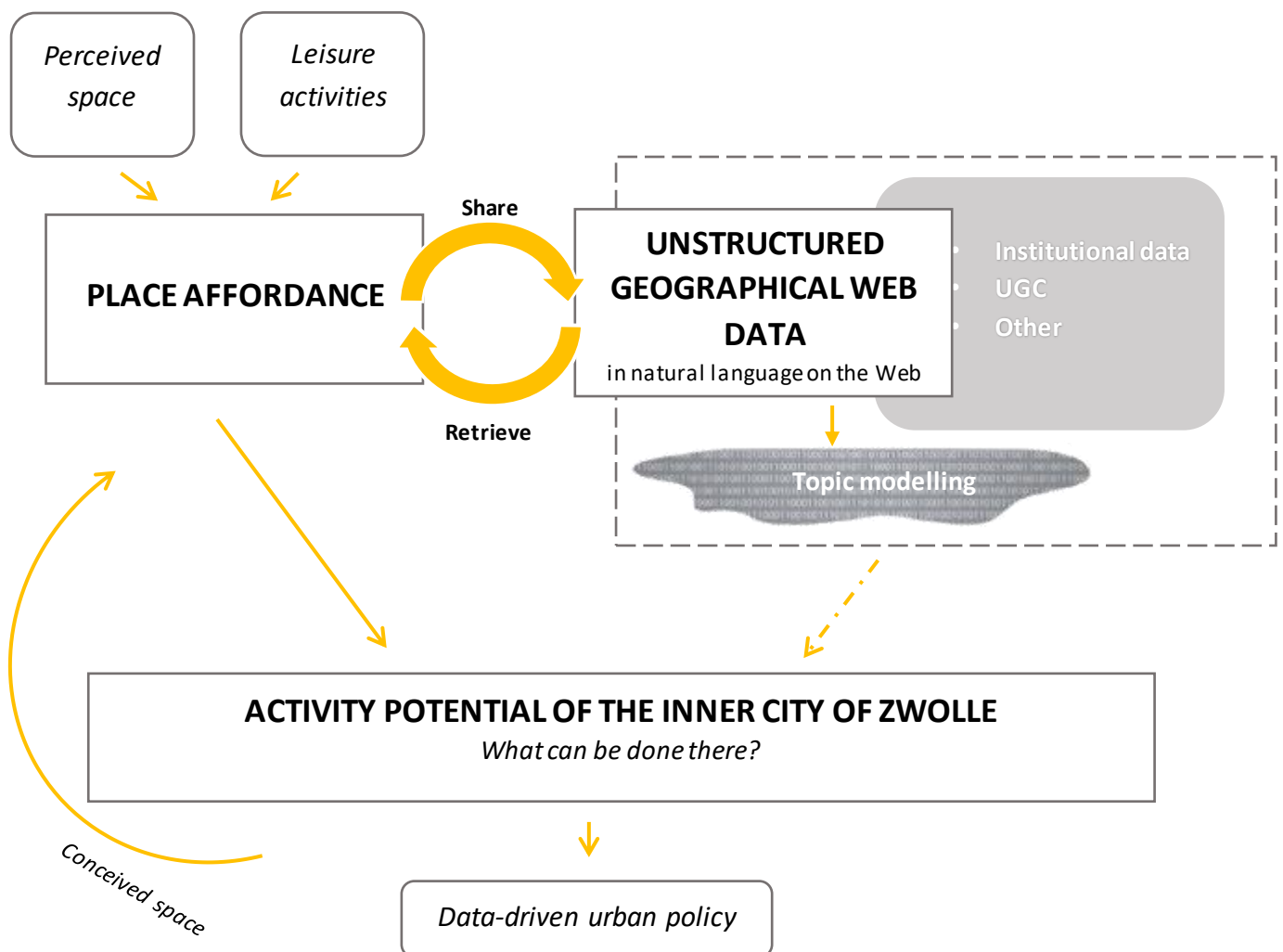


Figure 2.6 - Conceptual framework





## 3 | Methodology

The prevalence of geographic information on the Web and the need for more intuitive descriptions of place has increased the demand for machine understanding of place descriptions (Jones et. al., 2008). The intention of this thesis is to explore, use and discuss the potential of web-based semantic research on measuring activities in urban areas. Therefore a frequent used probabilistic topic model is used. Topic modelling is a tool for understanding a large amount of text by machine learning. It mines through web data and distributes a probability of topics describing place affordance (Blei et. al., 2003). The potential use of web data is increasingly changing the way how different scientific fields are processing and analysing information. The volume, velocity and variety of (web) data have caused a shift to data-driven city intelligence and the emergence of the interdisciplinary field of data science and analysis (Kitchin, 2013a).

### 3.1 Organizing unstructured data

Data science is an interdisciplinary field in which statistics, data analysis and related methods such as data mining, web scraping and predictive analytics are used. The essence of data science is that with analysis it is possible to extract *information* from the large amounts of data because data is not useful itself. The challenge of using (big) data analysis is that often unstructured forms of data are used. Commonly unstructured data is based on text, where the meaning and relevance remains hidden between the lines. *Knowledge* remains hidden from formal computational representation because the knowledge fits human designed formats instead of machine designed formats. Over the years, different methods have been developed to mine through the data and fetch the information people are looking for (Bansal, 2016).

The first sub-research question focus on the potential use of web-based text semantics: ‘*how can web-based text semantics be used for understanding and analysing activity potentials of people in certain urban areas of interest?*’. Semantics has become a key tool for extracting knowledge from text because there is too much (unstructured) information on the Web for humans to process (Gangemi, 2013). Therefore machine learning researchers have developed *probabilistic topic modelling*. Topic models are statistical algorithms which aim to discover and annotate large samples of documents with thematic information (Blei, 2012). Originally, topic modelling is used for understanding topics in large bulks of texts without any annotating or labelling (Blei, 2012, p. 78). It takes a collection of documents, mines through data and tries to find a structure or a *topic* within this collection. A topic is defined as “*a repeating pattern of co-occurring terms in a corpus (...) as ‘health’, ‘doctor’, ‘patient’, ‘hospital’ for a topic ‘healthcare’*” (Bansal, 2016, p.1). This is a comprehensive way to explore and structure a large set of unstructured documents: it organizes the documents based on the words that occur in them and on a similar vocabulary (ibid).

The technique helps to organize, search and understand the unstructured data - as semantic studies search the meaning of language and to process everyday used language (Giri, 2011). Mining data proves insights into the descriptive question ‘*what happened*’ by analysing what people have written on the Web. Topic modelling is used in different disciplines to describe shared opinions and topics and assists in better decision-making (Huang & Li, 2016). It also has been applied to understanding activity (patterns). Hasan & Ukkisuri (2014) have used topic modelling to discover daily routines based on wearable sensor data. Another example is by Farrahi & Gatica-Perez (2011) who have classified urban activities by extracting social media sources.

Nair (2016) describes three benefits of topic modelling: it helps to discover hidden patterns of topics across the documents, it helps to annotate documents according to these topics and it helps to use these annotations to organize, search and summarize texts. Latent Dirichlet Allocation (LDA) is one of the most popular topic models introduced by Blei et. al. (2003). By using an algorithm, it identifies patterns, structures, clusters and could summarize and predict data. LDA topic modelling is based on three related concepts: *documents* which are produced based on a pre-set list of *topics*. These topics

generate *words* based on a probability distribution within the document. LDA backtracks the topics that have created the document. This causes a problem of credit attribution: within a single text exists multiple themes but not all topics are equally relevant. In this thesis is aimed at both learning a single- and multi-label version of LDA (Ramage et. al., 2009). These models are supervised in a way that the training data ‘teaches’ the algorithm just as a teacher is supervising the learning process of its class by annotating classes.

### 3.1.1 Phases of data analysis

The enormous pile of unstructured data which is available on the Web that describes certain activities and places is not immediately useful for analysis. To analyse data and to get information, data analysis often exist of a few essential phases. Data analysis is often considered as a sequence of *importing*, *cleaning* and *preparing*, sometimes *manipulating*, *modelling*, and *visualizing* data (Adèr, 2008).

The flowchart below (figure 3.1) describes different phases of this research. The methodology behind every phase will be further elaborated in the upcoming paragraphs. The first phase involves building the ontology that describes different categories of possible leisure-related activities, urban places and including referents (paragraph 3.2). The next phase is preparing the training dataset. This dataset is manually collected from different websites that describe different places and their affordance in the inner city of Zwolle. To further prepare the date for the modelling phase, the training dataset is linked to the ontology and Open Street Map tags. Furthermore, the training dataset is extended by enriching the set with tags and reviews from Google. Eventually, before training the model, a few more preparations like a document-term matrix have to be prepared. These steps and assumptions are discussed in paragraph 3.3 and parts of paragraph 3.4. Paragraph 3.4 further stresses the actual modelling phase. As this thesis is an explorative research to analyse potential activity-based references of places, the evaluation of the model is an important phase that has been added to the flowchart (paragraph 3.5). Here is discussed how the statistical evaluation based on precision and recall works. The final paragraph 3.6 stresses the last step of the data analysis: visualising the results in a map.

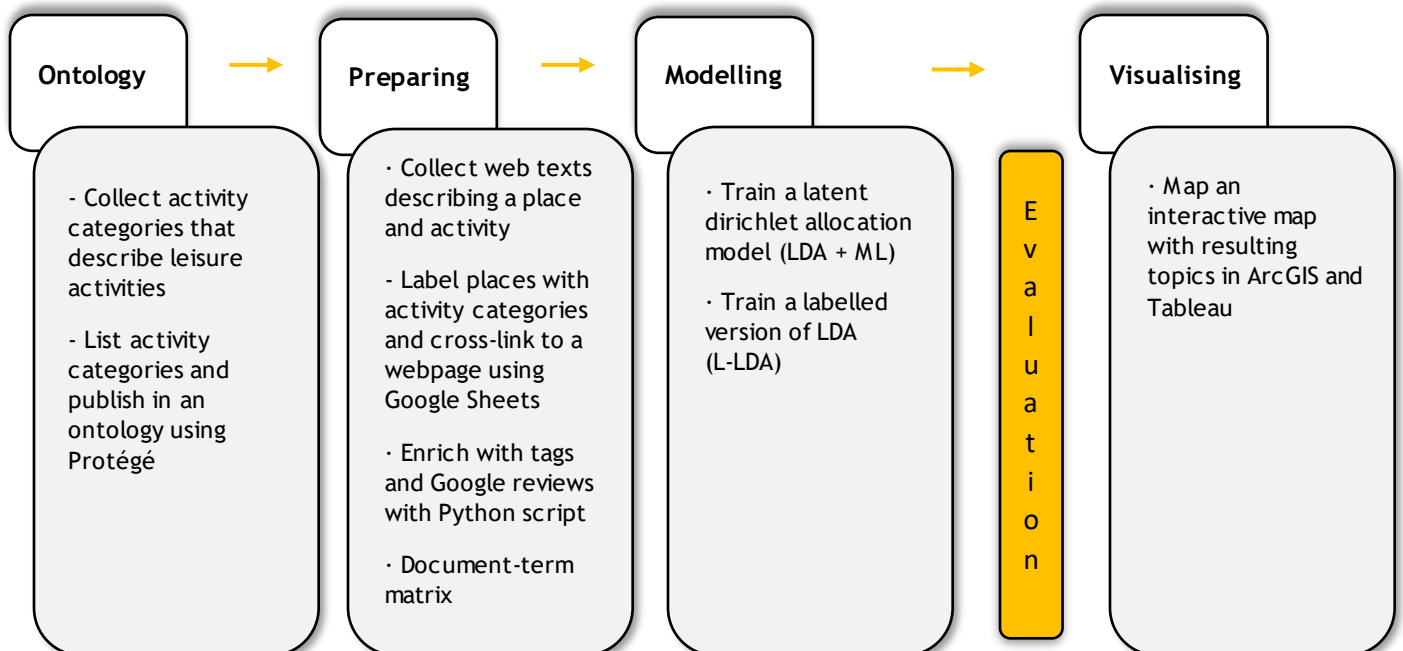


Figure 3.1 - Flowchart of methodology

### 3.2 Ontology modelling

In the process of going towards a semantic understanding of web information, creating an *ontology* is one of the first essential elements in semantic modelling. The ontology was used to categorize the used dataset - but is furthermore an approach to structure and classify the concept of leisure activities in urban space and its relationships. In this context, leisure is constructed as *a (meaningful) activity done in the free-time, after obligations such as work and the household, outside the house in a state of enjoyment, making use of the features of the urban environment*. This activity can be done by both citizens in their free-time (recreation) or by visitors coming into the city (tourism) - as they make use of the same features, and an exclusive use of none (Ashworth & Page, 2011, p.3). This definition is based on previous studies focusing on leisure, as discussed in paragraph 2.2. Hereby both the traditional time budget studies as from Szalai (1972) and the change of leisure into a commodity are acknowledged (Gregory et. al., 2009; Lorentzen, 2009). The urban environment is, of course, an important aspect of the activity.

By modelling an ontology, it supervises the model by structuring by annotating (Giri, 2011). As previously stated, the ontology teaches models that a rose is both a flower and a plant (Bekel, 2008; Slaghuis, 2009). Ontologies are built to create a clear consistent vocabulary that can be reused and published on the web. Thus, an ontology can be defined as *“a formal specification of a shared conceptualization”* and as a set of concepts and definitions (Gruber, 1993; Noy & McGuinness, 2001).

In this thesis, an ontology was created to capture the concepts of urban leisure activities. It captures possible activities in relation to leisure in an urban environment. The ontology is based on different (Dutch) activity categories and is expanded with particular leisure activities in Zwolle during the data collection. In this matter, the ontology modelling will answer the second sub-research question *“Which urban activity categories can be distinguished regarding urban leisure in the inner city?”*. The ontology ‘Urban Leisure’<sup>[2]</sup> was created with Protégé 5.2.0: an open-source ontology editor developed by Stanford researchers. In Protégé, an ontology is built on classes (descriptions of concepts), properties of each concepts describing features and attributes and sub-classes, that represent concepts that are more specific (Noy & McGuinness, 2001). Figure 3.2 visualizes the coherence between classes and the object properties. The class hierarchy was built on three main classes defining an activity, imported from the Place Activity ontology (Scheider, 2017).

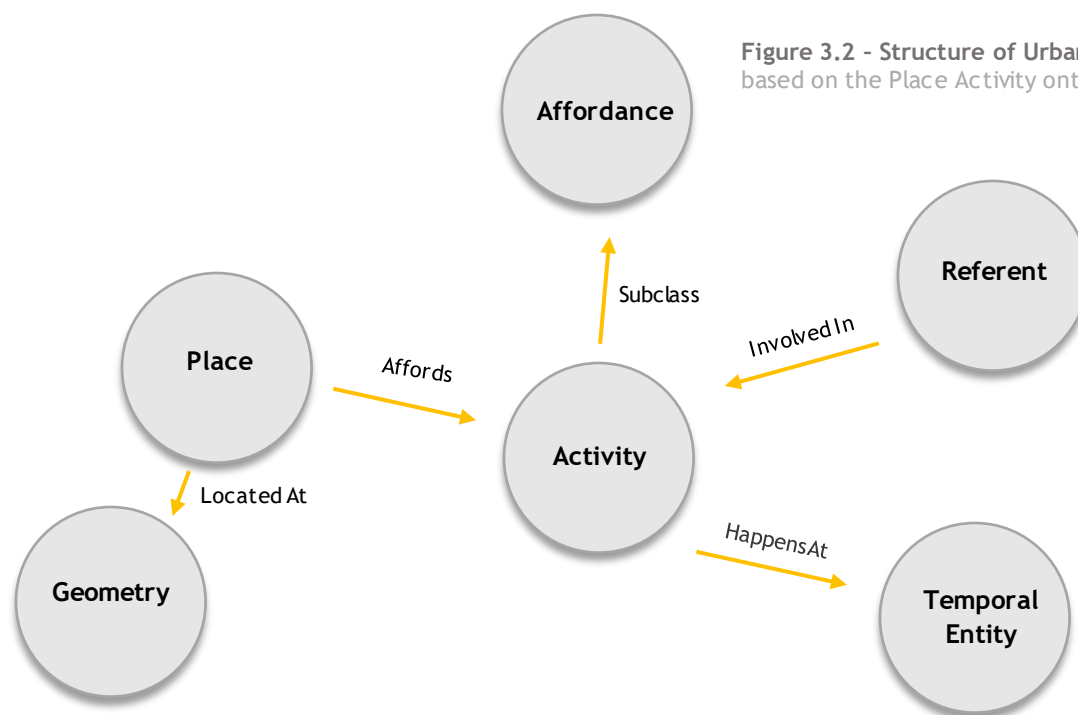


Figure 3.2 - Structure of Urban Leisure ontology based on the Place Activity ontology by Scheider (2017)

<sup>2</sup> See: <http://geographicknowledge.de/vocab/UrbanLeisure>

First, ‘*affordance*’ which describes the possible leisure-related activities that can be done in an urban environment such as sightseeing. This class defines the activity only, hence why it is chosen to use verbs to describe activities in the ontology. Second, the class ‘*place*’ describes a place that may afford an activity such as a restaurant. Some activities have a specific setting to conduct the activity while a place as a park potentially affords many activities. Place and activity are always tightly related as people *do* things in places (Hobel et. al., 2015; Scheider & Janowicz, 2014). ‘*Referent*’ is the third class which describes an object that is involved in an activity such as food is involved in eating at a restaurant. The classes further have different object properties as visualized in the figure and specified below (Scheider, 2017):

- **Affords:** affordance relation between a place and activity
- **InvolvedIn:** links the object that is involved in an activity to the activity
- **Involves:** the inverse of “InvolvedIn”, links activity to the involved objects
- **LocatedAt:** the location of a place
- **HappensAt:** the time when a place activity (might) happen(s)

Finally, the geometry and the temporal entity of an activity defines the space-time prism in which the activity potentially can take place (ibid, 2017).

Ashworth & Page (2011) have identified urban features in different clusters such as culture, history, business, sports, gastronomy, nightlife and shopping. These classes formed the starting point of collecting activity categories in an ontology, in order to capture all possible activities in relation to leisure. Collecting activity categories was done in an iterative process in which constantly subclasses were revised, refined and cleansed during the data collection.

As a starting point, activity categories were retrieved from studies as ‘*The Use of Time*’ of Szalai (1972) and the latest Dutch study ‘*Met het Oog op de Tijd*’ (2013) conducted by the *Sociaal Cultureel Planbureau* (SCP). Traditionally, leisure is seen as ‘free time’ when obligations such as work and household were fulfilled. A third used study, the *ContinuVrijeTijdsOnderzoek* (2015) by NBTC-NIPO Research. This study focus on different leisure activities and the type of activities that are conducted in the Netherlands. They consider leisure as a (day) activity conducted outside the home, and whereby people are minimum one hour away (including travel time). They have excluded visits to family and friends and activities during a holiday. Furthermore, search queries on travel review website TripAdvisor were added.

In all time-budget studies, a distinction is made between indoor and outdoor (leisure) activities. For instance, a movie can both be watched at home or in the cinema. The ‘*Urban Leisure*’ ontology takes into account activities that are done outside the house. The CVTO further considers that leisure activities do not include an overnight stay (such as camping) as only tourist will make use of that amenity. Also, visiting friends and family without any other actions were not considered. These assumptions were both taken over in the *Urban Leisure* ontology.

Building an ontology is an iterative process in which constantly subclasses are revised, refined and cleansed during the data collection. For instance, during the first LDA model output, it seemed that the quality of the model was better when for more top classes such as ‘food’ as referent was chosen instead of further specifying food by different cuisines such as ‘*sushi*’ or ‘*Argentinian cuisine*’. For these reasons, the ‘*Urban Leisure*’ ontology is never complete; there are other time-budget studies and data that could add on new activities, places or referents. Noy and McGuinness (2001, p.4) state that there is no one correct way to model a domain (such as urban leisure). As the ontology is built to describe the place affordance of Zwolle, this ontology describes activities in a middle-sized city in the Netherlands - but also could be reused for other cities.

### 3.3 Preparing and enriching web data

After describing the possible activities in the ontology, the next phase is to collect a dataset of web texts that describes the possible activities at different places. This paragraph focusses preparing the training dataset and on selecting web sources, which aligns the third sub-research question “*which web sources can be used to train a dataset for a web-text semantic model?*”. The first step is to collect different places in the inner city of Zwolle for which activity potential was described by different web sources. In chapter 2, a distinction was made between different sorts of information: web data that was created top-down by institutions, *user-generated content* (UGC) and newer sorts of big data such as sensor data. The potential use of all three sorts of data was investigated. Sensor data resources did not meet the assumptions to implement in this model because it was not able to regard this data as web texts. The selected web information was enriched by the classes as distinguished in the ‘*Urban Leisure*’ ontology and a place identifier in Open Street Map (OSM). The data collection was done by using Google Sheets and a corresponding survey to be able to adjust typing mistakes. Eventually, the dataset was enriched during the training phases with tags and reviews available of that place on Google.

#### Assumptions of defining places and collecting web data

To make sure that the model can be applied, the data selection had to include specific websites with a sufficient amount of texts describing the place and activity to parse. There are two main assumptions that influence the selection of web sources:

1. Places should be described uniquely by the web page
2. Places should be contained in Open Street Map (OSM)

The first assumption constrains selecting all possible websites that describe activities within the inner city of Zwolle. A lot of websites such as the official tourist info site and different blogs contain an overview of different possible activities. These are often themed as ‘the best addresses to shop in Zwolle’ or ‘must visit when you are in Zwolle’. This sort of web pages describes different places which cannot be applied to the model as they do not describe one place uniquely on the web page. However, some websites do describe multiple places on unique web pages and thus can be inserted into the data collection.

The second assumption relies on the place-boundedness of the activity. For the data collection, places with a name were selected to make the place identifiable. To identify and localize the places for the semantic model, the web text was cross-linked to an Open Street Map identifier: either a node or way (Table 3.1). Most places were identified as a node, corresponding to a single point in space. A few streets were identified as way which is the highest scale of places in this research.



osm	Node		Defining a point in space
osmw	Way		Defining a linear feature or area boundaries

Table 3.1 - Open Street Map identifiers

Wiki Open Street Map, Elements (2017)

Open Street Map (OSM) is a volunteered based and editable map of the world - a prominent example of VGI. The map is based on *open data*. The choice of using OSM over national geo-information databases (such as PDOK) is based on the international potential to reuse and the possibilities to download the data. However, geo-referenced information for Zwolle is often adjusted to the Basisregistratie Adressen en Gebouwen (BAG) geo-information. The screenshot of a part of the inner city of Zwolle shows how OSM further interpret different elements with different labels such as street, the name of the location, amenity and corresponding website (Figure 3.3). The screenshot shows how different buildings surrounding the Grote Markt in Zwolle have different amenities such as ‘worship’, ‘shopping’ and ‘café’ as is shown in different symbols. To conclude, users of OSM already have identified activity and place for lots of nodes on Open Street Map.



Figure 3.3 - Screenshot of the Grote Markt, Zwolle as shown on OSM OpenStreetMap (as screenshots in July 2017)

The aim is to collect as many places as possible: which resulted in 200 different places within the inner city as described on different web pages. Some places offer multiple activities - such as a park or a bakery that also has a café and sells coffee. To limit the model, the data collection is constrained by a maximum of five activities per place. Other constraints in the data collection are that the place should be described by its *current* activity and no overnight amenities are included as this is only done by a certain group of people. An example is the consequence of not adding the Pelsertoren into the data collection. It once functions as a historical building that could be visited - currently, the tower is a bed and breakfast.

#### Different sorts of web information: a consideration of place selection

As previously stated, the Web consists of different sorts of information which all can be collected as a data source. Web information is considered as the most important source nowadays (Fabrikant & Buttenfield, 2001; Wakamiya et. al., 2011). It is the source where people gaze upon, determining what they are going to do. At the same time, as consequence of *user-generated content* people describe their activities and rate them, which consequently is determining other people’s activities in the city based on the available web information. Therefore using web pages can be useful in capturing the perceptions of people as it resonates the voice of the consumer (Guo et. al., 2017). The encoding of places is based on different websites describing places in the inner city of Zwolle. Hereby websites were selected that are often used for planning activities or trips. These websites consist mostly of overview pages of ‘*what to do in Zwolle*’ and are linked to other web pages or the official web page of the place for further information.

The encoding was done in the summer months starting May until July 2017. Hereby must be acknowledged that some websites are time-dependent. Mostly institutional websites adjust their agendas and promotions based on the time of the year, which can determine that some activities such as sitting on a terrace or shopping fair are represented in the training data while a winter fair is missing in the training data.

The collection of places and their possible affordance were retrieved from different starting points. First, by selecting places retrieved from top-down based web data created by institutions. The website of the municipality is one of the most important sources to retrieve information about the city. On the official website, you find a tab 'vrije tijd' (English: leisure time), under which descriptions can be found in different subjects such as events, culture, sports, recreation and shopping. This redirects to different pages but also redirects to the official website of the Zwolle Tourist Agency (VV). This website describes the city in multiple themes and refers to places and activities to do in Zwolle. It consists of categories 'culture', 'tasting', 'activities' and 'shopping'. Each main category refers to different places and partners' own information pages. The tourist agency thus highlights different places of interest in the city. Furthermore, the list of Google search results was retrieved during the data collection. When looking for a certain city on Google, the search engine shows its own recommendations next to the hits of the related search query. These points are suggested when is searched for queries as 'Zwolle' or 'Zwolle Nederland interessante plaatsen'.

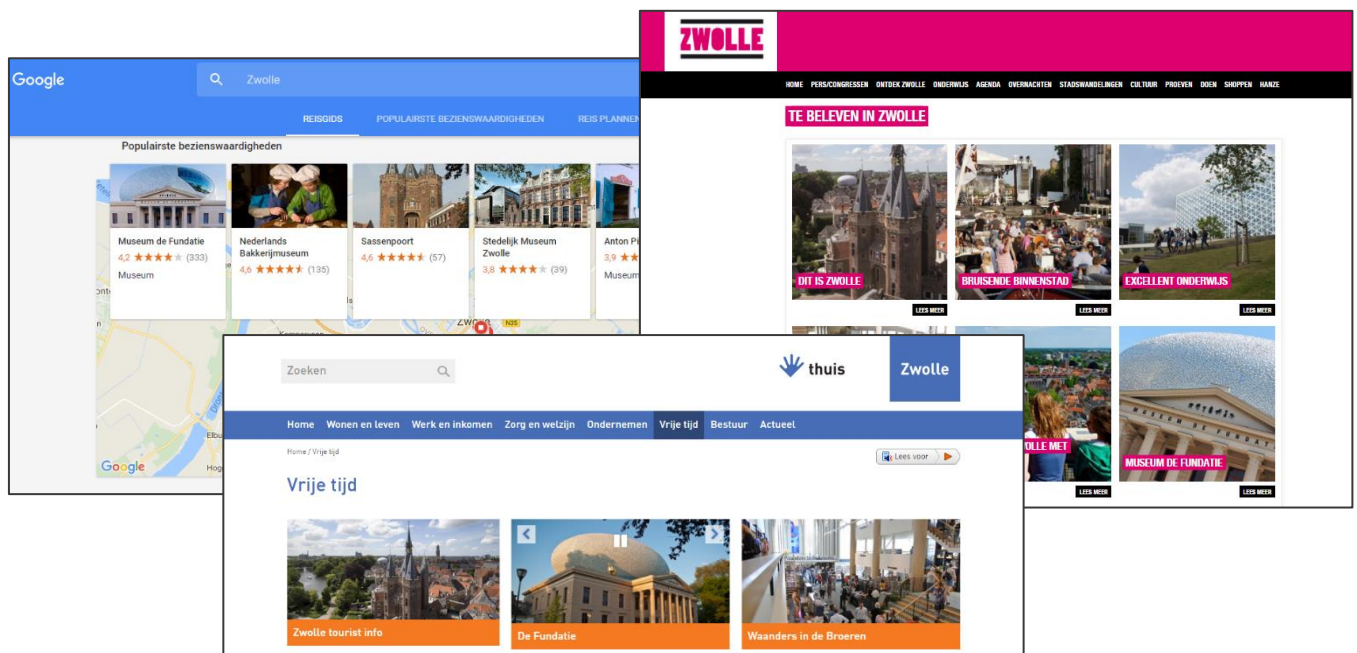


Figure 3.4 - Webpages describing the place affordance of the inner city |  
Obtained from: Zwolle.nl/vrije-tijd, zwolletouristinfo.nl, Google POI

Secondly, to acknowledge the rise of user-generated content, different UGC sources were retrieved. First the specific tourism website TripAdvisor, which is an often adopted example of UGC. The website enables people to exchange their opinions and recommendations about certain destinations and specific places, often with a rating (Akehurst, 2009). The tab 'sightseeing' describes different activities to do in the city, in multiple categories such as museums, recreation and outdoors. Another social media website that has been taken into account is Facebook - one of the most popular online social networking services. Commonly, service-related websites have their own Facebook page where Facebook users could leave recommendations for their friends. Each city has its own page with recommended places to eat, drink, doing sightseeing, hotels or addresses to shop. The list is extensive and refers to the official Facebook page of a certain amenity.

Next to social media - different sites in the *blogosphere* were retrieved. Blogs exist since 1997 but became more relevant in the last decade. Blogs provide primarily text information - but also photos, videos and audio. Akehurst (2009) describes blogs as ‘*personal thoughts and commentaries on say a specific destination, personal travel stories and details of trips*’ (p. 54). To retrieve blog commentaries of Zwolle was searched on Google with the query ‘tips Zwolle’. The suggested blog pages and related pages were selected from the first two pages of the search query. A development in the *blogosphere* is that not only blogs are written by and for consumers (*consumer to consumer*, C2C), but also includes *business to consumers* (B2C) and *government to consumers* (G2C) (Akehurst, 2009). An example of this fluid character of user-generated content is ‘In de Buurt Zwolle’ - a platform to inspire and inform citizens about the whereabouts in the city. The set-up is like a blog, but the platform is owned by De Persgroep. The tips are based on the people of Zwolle and describe different neighbourhoods, local news, tips for activities, shops and restaurants. But also businesses sponsor blog posts on this platform to advertise their business. ‘In de Buurt’ often publishes descriptions of places in a list of ‘*places where to eat the best bitterballen*’ or ‘*the best terraces in the sun*’. All posts in the months June and July on the subpages ‘*doen*’ and ‘*eten & drinken*’ were retrieved.

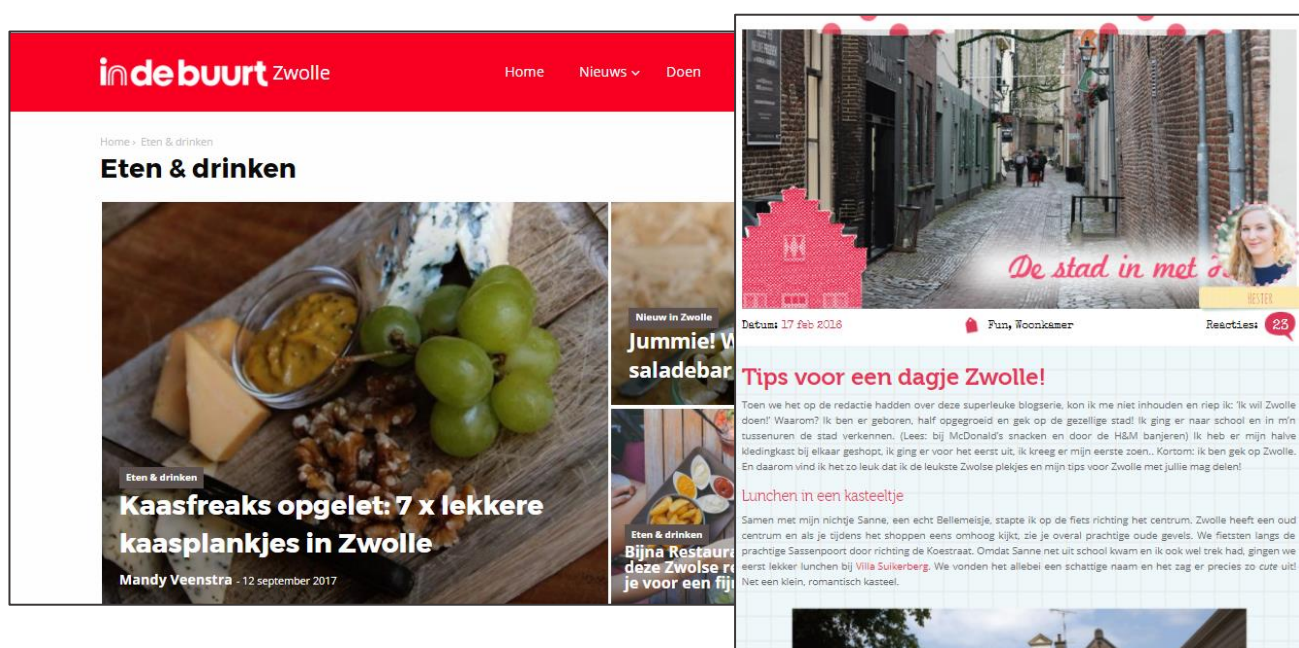


Figure 3.5 - Webpages describing the place affordance of the inner city II  
 Obtained from: In de Buurt Zwolle, blog Huis van Belle

Collecting places and selecting the most suitable website was an iterative process; in which the places that were found on a website caused a further search for the best site that describes the place with the most available text to parse. Often this was the official website of the place itself. Furthermore, during the data collection consequently, new places were added based on the descriptions of the website. The evaluation and further experiences with selecting places and activities will be discussed in the results. In this chapters, the difficulty to parse ‘sensor’ based web data and its potential will be further discussed.

### Google Sheet form

The data collection was sustained by a place coding form in Google Sheets, to prevent coding errors for the ontology (Appendix I). The activity label and the website URI were cross-linked with the identifier of the OSM feature, which was retrieved from Open Street Map Nominatim <sup>[3]</sup>, and the categories of place and referent based on the ontology (referred to as ulo).

<sup>3</sup> See: <http://nominatim.openstreetmap.org/>



In this way, the text can be joined back to a shapefile of the inner city of Zwolle. Table 3.2 indicates the seven elements of the Google Sheet form and an encoded example of music venue Hedon. The complete encoding sheet can be found in the Place LDA repository on GitHub <sup>[4]</sup> - a website for open source software.

OSM Identifier	Place name	ulo:Activity	ulo:Referent	ulo:Place	Website
- Node - Way - Relation	Name of the place	As in ontology	As in ontology	As in ontology	URL
osm:2500428169	Hedon	Ulo:Listening	Ulo:Concert	Ulo:Theatre	<a href="https://www.hedon-zwolle.nl/4/info">https://www.hedon-zwolle.nl/4/info</a>

Table 3.2 - Example of encoding of place affordance

### 3.4 Modelling LDA

To extract urban leisure activities with web texts, both a single and multi-label version of Latent Dirichlet Allocation (LDA) were trained to test whether urban leisure activities can be estimated as questioned in the fourth sub-research question: ‘to which extent can urban leisure activities be estimated, extracted from knowledge from a web-text semantic model?’. This paragraph zooms in on the used method, assumptions and parameters of LDA modelling. To run LDA in Python - a widely-used programming language - a few steps are required: preparing the documents by web enrichment, cleaning the texts and preparing a document-term matrix before actually running the model (Bansal, 2016). The script in Python - ‘Place Topic Modeller’ - is written to build the topic model. The encoded script and source code can be found in the repository on GitHub.

#### 3.4.1 Assumptions and supervision

LDA is a generative model, meaning that it is a statistical technique that tries to find *latent variables* (topics or documents) in a probabilistic process to *explain* the observed data. The task is to find topics that fit the model the best, given the data. The main assumption of using the Latent Dirichlet Allocation model is that documents are produced from a mixture of topics. Documents exhibit multiple topics and LDA tries to model how the topics are distributed within the text document. Weingart (2011) has explained LDA by the ‘*bag of words assumption*’: the words within texts are generated based on a probability distribution of topics. The order of the words in the document does not matter and are collected based on the topic distribution. The *bag of words assumption* considers a document as a bag of words that are randomly used from different topic ‘bags’. Generating the words throughout the document happens in a two-stage process (Blei, 2012, p.78):

1. Randomly choose a distribution over topics.
2. For each word in the document:
  - a. Randomly choose a topic from the distribution over topics in step 1;
  - b. Randomly choose a word from the corresponding distribution over the vocabulary.

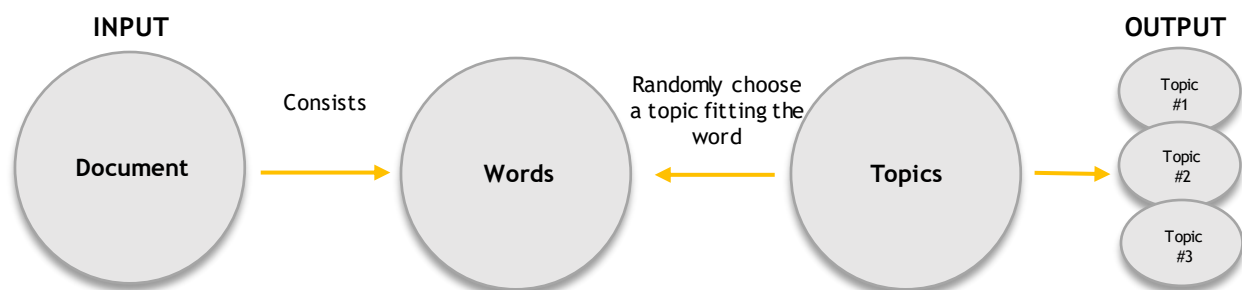


Figure 3.6 - The basic notions of LDA

<sup>4</sup> See: <https://github.com/simonscheider/PlaceLDA>

To concretise the use for this thesis: the semantic model traces back which activity categories (*topics*) define the web information about the inner city of Zwolle. The input is x number of documents describing the affordance of places in Zwolle. These documents consist words which are based on the probability distribution of topics. The output of the LDA model is a number of topics that describe the ‘themes’, thus the place affordance as described within the dataset. In this thesis, a supervised variant of LDA is used, in which it is supervised by the ontology, in combination with unsupervised machine learning classifiers (LDA + ML) (paragraph 3.4.5). However, this distribution is focussed on a single-label result while in documents multiple topics can reside. Often documents are built up from multiple topics which are all relevant to one subject: these can be considered as multi-labelled documents (Ramage et. al., 2008). Therefore, Ramage et. al. (2009) have introduced a *labelled Latent Dirichlet Allocation* (L-LDA). L-LDA models incorporate multi-labelling by constraining the topic model by only using topics that correspond to an observed document label in the training data (Ramage et. al., 2009, p. 249). This supervision is done by labelling the ontology to the web texts. L-LDA modelling is proven by Ramage and colleagues to address this credit problem, and proven to be more effective on tasks such as document visualizations. This thesis aims at both testing the single-label LDA with machine learning classifiers and the multi-label L-LDA model to see if any improvements can be made by the L-LDA model.

### 3.4.2 Web-based enrichment: constructing training data

The list of places, which were identified with OSM identifiers during the manual encoding, are imported and enriched by different sources such as several web information in the script such as web title, class and coordinates (see for details in table 4.3). One of used web information sources are the relevant keys of interest from OSM itself, by using OSM Overpass API <sup>[5]</sup>: [‘shop’, ‘amenity’, ‘leisure’, ‘tourism’, ‘historic’, ‘man\_made’, ‘tower’, ‘cuisine’, ‘clothes’, ‘tower’, ‘beer’, ‘highway’, ‘surface’, ‘place’, ‘building’]. These keys were indicated by OSM users; an example of *volunteered geographic information*.

Key	Description
<b>Class</b>	Activity class manually added in terms of ulo ontology. Format <i>ulo:Activity</i> and <i>ulo:Referent</i>
<b>ulo_Place</b>	Place type manually added in terms of ulo ontology
<b>Website</b>	URL of the website used to scrape place descriptions
<b>Web title</b>	Title of the website used to scrape place descriptions
<b>Web text</b>	Text of the website used to scrape place descriptions (cleaned with BeautifulSoup, see <i>placewebscraper.py</i> )
<b>Name</b>	Name of the place (manually added)
<b>Review text</b>	Text of Google Place reviews (if available). Google place information was added based on spatial distance and name similarity
<b>Google type</b>	Place tags from Google Places (if available). (in alphabetical order)
<b>Google ID</b>	Google Place ID (if available).
<b>Latitude</b>	WGS 84 Y Coordinate (taken from OSM, converted to centroid for ways) (if available)
<b>Longitude</b>	WGS 84 X Coordinate (taken from OSM, converted to centroid for ways) (if available)
<b>OSM key tags</b>	Open Street Map key tags containing their respective values, or ‘No’ if not present at OSM [‘shop’, ‘amenity’, ‘leisure’, ‘tourism’, ‘historic’, ‘man_made’, ‘tower’, ‘cuisine’, ‘clothes’, ‘tower’, ‘beer’, ‘highway’, ‘surface’, ‘place’, ‘building’]

Table 3.3 - Keys of enriched data file  
‘Place Topic Modeller’ by Scheider (2017)

<sup>5</sup> See: <https://overpass-turbo.eu/>

All places on the OSM map can be labelled with extra information such as amenity, opening hours, cuisine and a web address (see figure 3.7). Not every place is evenly extensively labelled - some places miss opening hours and other places miss contact information.

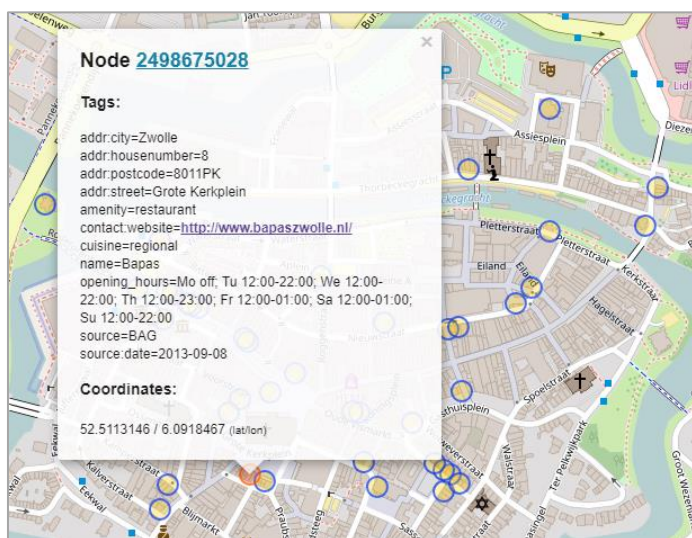


Figure 3.7 - Extra information on OSM for restaurant 'Bapas'

By using the search query 'amenity = restaurant' in OSM Overpass. This indicates all places that have been indicated as a restaurant by VGI.

OSM Overpass Turbo (2017)

Furthermore, information from Google Places was added based on the web address. Linking Google information with the OSM identifier was based on both place name similarity and a spatial distance of maximal 300 metres. The information on Google Places consists texts of available reviews of users of a certain amenity or place and a web address. Finally, the OSM objects were enriched by the manually selected and scraped web information in the form of a text-file CSV. To scrape the web texts, the Python's module '*BeautifulSoup*' by Richardson (2017) [6] was used to clean up HTML texts.

### 3.4.3 Cleaning text and model preparation

To train an LDA topic model on web texts, we have used the Python module '*lda*' [7]. Before actually modelling, an important step is to clean the available texts. Cleaning in any text mining task consists of removing punctuations, stop words and normalize the set of documents (corpus). To do so, the NLTK (Natural Language Toolkit) library was added for natural language pre-processing. In the theoretical framework, it was discussed how the prevalence of geographic web information mostly consists of humanly designed formats which cannot be easily grasped by machine learning (Adams & Janowicz, 2014). By using NLTK's word tokenizer the data collection is cleaned of punctuation, stop words and stemmers. The process of tokenizing is converting a sequence of characters (words on a web page) into a sequence of tokens (strings with their identified meaning). Both the English and Dutch libraries were used since the data collection consists of mostly Dutch web pages.

Furthermore, English or Dutch stemmers were used: a normalizing method to identify the many variations of one word and meaning. Afterwards, the document was turned into term-document matrices using Scikit-Learn's count vectorizer with a minimum document frequency of 1. The following document-term matrix (table 4.4) shows an example of a distribution over a collection of documents (corpus) D1, D2 and D3 ... and the frequency count of a certain vocabulary within a topic (W1, W2 or W3 ...). Each entry (i, j) represents the frequency of term  $W_i$  in  $D_j$ . The topic document distribution is shown by  $N$  (Bansal, 2016).

<sup>6</sup> See: <https://pypi.python.org/pypi/beautifulsoup4>

<sup>7</sup> See: <http://pythonhosted.org/lda/>

	Word 1	Word 2	Word 3	N
Document 1	0	2	1	3
Document 2	1	4	0	0
Document 3	0	2	3	1
N	1	1	3	0

Table 3.4 - Example of a document-term matrix

Source: Bansal (2016)

LDA aims to improve the distribution of documents by repeating and sampling this document-term matrix until any convergence occurs. This process is converted by two lower matrices: the document-topics matrix and the topic-terms matrix. These matrices provide distributions over topic words and document topics. By repeating this for each word for each document, the probability is being recalculated during a number of iterations. The LDA model was trained with 18 topics iterating 600 times over the documents. Hereby was chosen for the most frequent class label within the training data for multi-encoded places.

The LDA model enables to work with different parameters to distinguish the training data and its enrichment. Besides the difference in set language, three other parameters need to be set. First, if 'useTypes' is indicated as true, the LDA also adds the tags from OSM and Google Place into the feature vector, in addition to the topics. We also tested different semantic levels of tag features. When 'actlevel' is marked as true, it restricts the model to only use activity classes and not the referent classes. At last, the 'minclasssize' filters out classes with too few instances in the data with less than five instances.

### 3.4.5 Classification: cross-validation

To classify the selected places by activity classes, ten standard machine learning classifiers were used, including logistic regression. All classifiers are briefly explained in Appendix II. The online user guide of Scikit-Learn provides more detailed information about each method (Scikit-Learn, 2017a).

- \* Logistic regression
- \* Nearest Neighbours
- \* Linear SVM
- \* RBF SVM
- \* Gaussian Process
- \* Decision Tree
- \* Random Forest
- \* Neural Net
- \* AdaBoost
- \* Naïve Bayes

The classifiers are retrieved from the Scikit-Learn Python module (Skikit-Learn, 2017a). This module is a simple machine learning tool for data mining and analysis. The classifiers were trained and tested on feature vectors including topics, tags, topics + tags which were extracted from the manually selected web texts, Google web texts, Google reviews or OSM/Google tags by using all ten classifications. During the training, a 10-fold cross validation was used (Scikit-Learn, 2017b). Cross-validation divides data into different subsets and takes one subset as a *validation set* while the remaining subsets are considered as *training data*.

### 3.4.6 Substituting labelled LDA

While in the LDA was chosen for the most frequent class for each place, this does not match the multifunctionality of the reality. Therefore, as final step a labelled Latent Dirichlet Allocation topic model was produced in order to compare which improvements could be made by using multi-labelling. *Labelled* Latent Dirichlet Allocation is based on the paper of Ramage et. al. (2009) in which they further problematize the problem of credit attribution arising from LDA models: a text is not evenly built on a number of topics but is often distributed. The supervision is done by only using the activity labels deriving from the *Urban Leisure* ontology.

The labelled version is based on a submodule L-LDA implemented like Sci-kit Machine Learning in Python (Adams, 2017) <sup>[8]</sup>. It makes use of the same enriched file as was used in LDA. L-LDA generates a description of the activity labels in stemmed words and creates a confusion matrix directly on the documents. However, the predictability of L-LDA is different than previously discussed in LDA topic modelling. The probability factor of L-LDA is directly based on the given class in a certain text by giving a distribution over labels. Therefore, L-LDA is a multilabel indicator which is making comparing to the single label classifiers of LDA not straightforward (ibid, 2017). In the following chapters, the differences are shown and discussed in relation to the difficulty of comparing multilabel and single label classifiers on the training data

### 3.5 Evaluation: validity, fitting and reliability

The Web is acknowledged as an endless source of information as roughly 90 percent of the data is proven not to be older than two years (van Oortmarssen et. al., 2014). The use of LDA in various disciplines has proven LDA to be a powerful content analysis technique designed for extracting and summarizing large sets of unstructured documents (Jacobi et. al., 2015). Considering the explorative character of this research, evaluation of the used procedure is essential. Does the model allow extracting and analysing leisure activities and actually describes urban areas of interest? A commonly cited aphorism in statistics is “*all models are wrong, some are useful*” (quote of George Box in 1976). The use of models is often debated on their correctness and validity while this quote states that the focus should rely on whether something can be applied to everyday life in a useful manner. In this light, discussion and evaluation of the results is an important part of this thesis thus a specific chapter will discuss the trained dataset.

A common pitfall is that despite that topic modelling is useful to summarize shared opinions, the resulted topics from LDA modelling are not necessarily intuitive, human concepts (Campbell et. al., 2014). The topics are derived from a random word distribution - which could not match human perception. On top of this pitfall, its internal validity relies on independent topics derived from a random distribution. Sometimes it is desirable to overlap different topics, such as ‘*sport*’ and ‘*news*’ could overlap in ‘*sports news*’ (Campbell et. al., 2015). To overcome this issue, the supervised model intends to distribute pre-defined, thus human interpreted topics. Nevertheless, the confidence level of the used ontology to define the topics is never outright and completed. Modelling an ontology is an iterative process in which another activity, referent or place class could be missing in the dataset because it is based on the training data and not on all possible activities and therefore only relevant for the provided corpora.

Yet, in machine learning, an important question is whether the resulted topics *fit* the model: the goodness of fit. The main aim of all machine learning and topic models is to learn from a certain dataset that enables to predict future and new data (Domingos, 2012). Therefore the consideration of *overfitting* and *underfitting* is important to evaluate how well the data generalize to new data because the used dataset is always a sample. In other words, how well can the learned topics be applied to specific examples and new data? Overfitting is when the model adapts the data within the sample too well. As consequence, it also learns the noise in the data. When adding new data, the model does not fit so the model is not generalizable. Underfitting is when neither the model can fit the data nor can the model generalize new data. To limit overfitting, it is important to focus on cross-validation. This way the overall validity and quality of the predictions can be quantified - as some data was considered as ‘new’ data that is added. This way can be predicted how the model generalizes beyond training (Domingos, 2012; Ellis, 2017; Scikit Learn, 2017). Domingos (2012) explains the relevance of cross-validation as securing making an accurate prediction in the ‘real world’.

The ten classification scores used in this thesis (see paragraph 3.4.5), based on *accuracy* and *weighted precision*, were compared to a naïve classifier, the majority vote. All classification scores are based on a classification of *precision* and *recall* (Scikit-Learn, 2017d). This is based on the principle that all

<sup>8</sup> See: <https://github.com/TaskeHAMANO/LLDA/tree/a25ec0f66c48e9b4d21109e342da6cbd6e67764a>

predicted instances can be predicted either positive or negative and either true or false (figure 3.8). Klintberg (2017) explains this by using a sample of pictures of *hot dogs*, can be predicted as a hot dog or as something else (positive or negative). However, a prediction can be true or false. For example, this could mean that a picture of two tanned legs on a beach is identified as a picture of a hot dog, while the picture clearly does not show hot dogs. This prediction is *false positive*. When a picture of a hot dog is not predicted as hot dog, then this prediction is *false negative*. When the prediction is true it means that the prediction is correctly done (*true positive*) and of course vice versa (Klintberg, 2017; Ting, 2010; Scikit-Learn, 2017d).

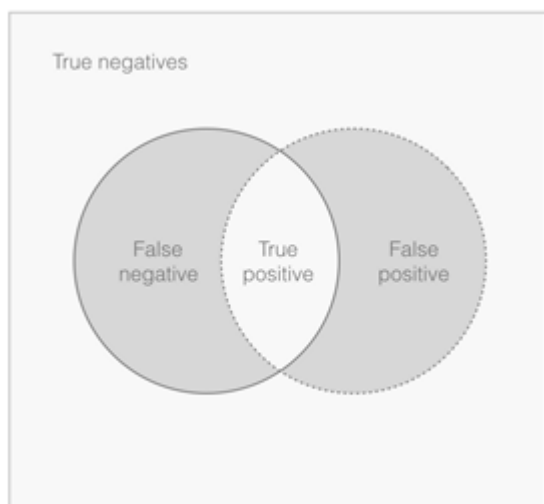


Figure 3.8 - The principle of precision and recall  
Klintberg (2017)

Precision and recall are used to measure how relevant the outcome is. A high precision means that a model has more relevant results than irrelevant ones: a low false-positive rate. While a high recall means more relevant items were retrieved: a low false-negative rate. High scores for both shows that the classifier is returning accurate results (high precision) as well returning positive results (high recall). When the classifier has a high recall but a low precision, which returns many results but many of them are incorrect. A high precision and low recall mean the opposite - few returned results but most of them predicted correctly. An ideal situation is thus a model with a high precision and a high recall (Scikit Learn, 2017d). The classification performance is summarized in a confusion matrix, which shows which errors are made or which classes are predicted sufficiently.

In this thesis, a ten-fold cross-validation is implemented. This allows training the model ten times on different subsets of the data to estimate the performance of newly added data (ibid, 2012; Brownlee, 2016). The used dataset is based on the places that people are describing '*what to do in Zwolle*' - the '*voice of the consumer*' (Guo et. al., 2016) - and not on *all* possible activities in the inner city of Zwolle. Considering all activities ask for a different encoding strategy.

The consequence of the chosen data selection strategy is that not *all* places within the boundaries of the inner city of Zwolle are considered. By using a different strategy, for instance by selecting all existing places in the city, influences the output. Therefore, Campbell et. al. (2014) states that it is important that LDA studies report their parameters for the reliability of the used method. However, the reliability is also based on the data collection. Akehurst (2009) describes the pitfall for user-generated content that this content may be '*limited in value, reflecting incoherent, unstructured and random ramblings of individuals*' (p. 52). It is important to acknowledge that the results of the model are influenced by the input of the model.

The research design is composed for a case study of a Dutch middle-sized city and therefore the findings are not generalizable to cities, especially to larger cities. The activity potential of every city differs which influences the used topics in the model. Alternatively, the elements of this research design are designed for reuse. The ontology is categorized into different main categories as some classes might be too specific for a city. The publication of the '*Place Topic Modeller*' on GitHub invites other researchers to read and reuse the code while the publication of the visualisation including metadata encourages others to adjust the visualisation or add data.

### 3.6 Visualising data

Given the potential of data, it encourages to visualize the results in order to reveal and communicate resulting trends in the data. To visualise the results, in this research three kinds of visualizations are used. First, the topics can be described by word clouds (Adams & Janowicz, 2014). Therefore the topics will be visualised with the online word cloud generator Wordle. The word cloud represents a probability distribution over all words, but the proportional distribution of words has no further value in the visualisation. The word clouds are used as a semantic analysis to say something about the validity of the topic distribution (Kling & Pozdnoukhov, 2012).

The second visualisation relates to evaluating the model, namely by creating a decision tree to show which decisions were made during modelling the results. A decision tree is a human-readable tree from which you can read the reliability of the model (Scikit-Learn, 2017e). It checks for different rules if a certain feature (either a topic or type) has met the condition and shows for how many items within the training data this condition applies. The *Place Topic Modeller* script automatically creates a decision tree which can be visualised with Graphviz editor. The advantages of decision trees are that they are simple to follow and to interpret and require little data preparation. However, the decision tree can create too complex trees that fit the data too well and do not generalize the results. Furthermore, they can be unstable by small variations in the dataset which causes some bias. Because decision trees are human readable, errors and biases can be identified (Scikit Learn, 2017e).

As a final step in the data analysis, the outcomes of the model need to be visualised in order to present the potential geographical use in urban policy. In the digital realm, Rob Kitchin describes the use of navigating through data to enable users to gain an overview of the dataset but also enables users to zoom in, select or filter groups of data (Kitchin, 2014, p. 106). Its applicability is constructed by visualizing areas of interest, which indicates spots within the city of high interest. The *Place Topic Modeller* script includes a script '*exportSHP*' that exports a shapefile '*placetopics*' to map the results. The script is based on the Fiona Python package <sup>[9]</sup>.

The resulting shapefile '*placetopics*' and its data was further prepared in ArcMap: a programme to view, edit, create and analyse geospatial data. The shapefile consists of georeferenced data about the encoded places and a distribution over the resulted topics derived from the model. In the Editor, the attribute table was cleansed: unnecessary columns were deleted and a new column with the final assigned topic was created. Next, a selection was made for each topic and saved in a new shapefile. All these shapefiles are eventually merged into a new shapefile '*topics*' which shows for each topic which place corresponds to that specific topic.

Afterwards, the new shapefile was exported into Tableau 10.1: a data visualisation programme. In Tableau, the data was visualised by plotting the longitude and latitude of the encoded places in a symbol map. To show the different topics in the Tableau visualisation, every topic was clustered to create *activity clusters* which correspond to a topic.

---

<sup>9</sup> See: <https://pypi.python.org/pypi/Fiona>





## 4 | Zwolle: a vibrant, leisure and data-driven city

Many governmental branches have adopted feeds of data in favour of their management practices and research in favour of the everyday life of their citizens (Kitchin, 2014b). The chosen case-study of this thesis is the inner city of Zwolle. The city of Zwolle claims that the city “*celebrates life in the historic inner city*” and is a perfect place to “*discover the city, experience the historic inner city and celebrate (...) during countless festivals and events*” (Zwolle Tourist Info, 2017). Zwolle focuses on enhancing the experience economy in the city, like most, if not all, cities have done or are doing. The municipality has set an agenda for a lively city with keywords such as *hospitable* and *surprising*. What makes this case-study is interesting is that Zwolle aspires to do so with a data-based policy. In this chapter will be zoomed in on this city as the chosen case study, its leisure function and their ambition of being a data-driven city.

### 4.1 The vibrant city

Zwolle is a middle-sized city, situated in the eastern part of the Netherlands, in the province of Overijssel. The city counts 126 thousand inhabitants (CBS, 2017a). The inner city is characterized by the *Netherlands Environmental Assessment Agency* (In Dutch: PBL) as a vibrant city within a strong region (Everts et. al., 2015). This characterization distinguishes itself by its cultural history and atmosphere. Elements of this ‘vibrant city’ are a relatively young demography with a lot of students, children and few elderly. The economy is relatively diverse with few vacant buildings in the streetscape. Other similar Dutch cities are Haarlem and Utrecht, according to the PBL (ibid). Zwolle functions as a major regional economic hub between the north of the country and the west. The most common economic sectors in Zwolle’s inner city are trade, business services and the food service industry, respectively 36.8%, 20.9% and 14.5% of the city’s economy (Zwolle Buurtmonitor, 2015).

Zwolle draws relatively young citizens who are attracted by education and job opportunities and the urban amenities of the city (CBS, 2017b). Meanwhile, in 2016 the city has welcomed over 1.3 million visitors, both citizens and tourists, making use of amenities in its old inner city centre. Many 15<sup>th</sup> century buildings remaining from the former Hansa settlement are preserved and attracts people from outside the city (Gemeente Zwolle, 2017).

Zwolle and its city centre are doing pretty fair according to the *Inner City Strategic Report* of the municipality (Gemeente Zwolle, 2017). In 2016, the city attracted ten percent more people than the previous year. Regarding the national rankings, Zwolle marks being the 6<sup>th</sup> city of welcoming the most people in the Netherlands. The city centre is seen as a lively and dynamic place for living, shopping, meeting and working according to the municipality. Although the most important reason to visit the city is still shopping (36 percent), other reasons to visit the inner city of Zwolle are the various cafés and restaurants, museums, events, theatre and preserved city sights (ibid). The main shopping areas clusters are around the Grote Markt, Sassenstraat, Diezerstraat, Oude Vismarkt and the Lutkekestraat while squares as Melkmarkt, Grote Kerkplein and Rodetorenplein offers a variety of cafés and restaurants.

The highlights of the city vary from places where to shop, restaurant to eat and culture to see. From the 15<sup>th</sup> century Hanseatic architecture such as the gatehouse Sassenpoort and the Peperbus tower to two top museums: Stedelijk Museum Zwolle and the art collection of Museum de Fundatie which welcomed over 284.000 visitors in 2016. Furthermore, the Grand Church on the like-wise named square surrounded by cafés. Another acknowledged sights are Waanders in de Broeren, which is more than just a bookstore in a church, which attracted 467.000 visitors in 2016. At last, the city hosts a three-Michelin star restaurant, ‘De Librije’, which is indicated as one of the best restaurants in the country. Museum de Fundatie, De Librije and Waanders in de Broeren all recognize a national reputation and are part of the city its branding strategy (Gemeente Zwolle, 2017). These seven highlights of the inner city are indicated in an overview map of the city (figure 4.1).

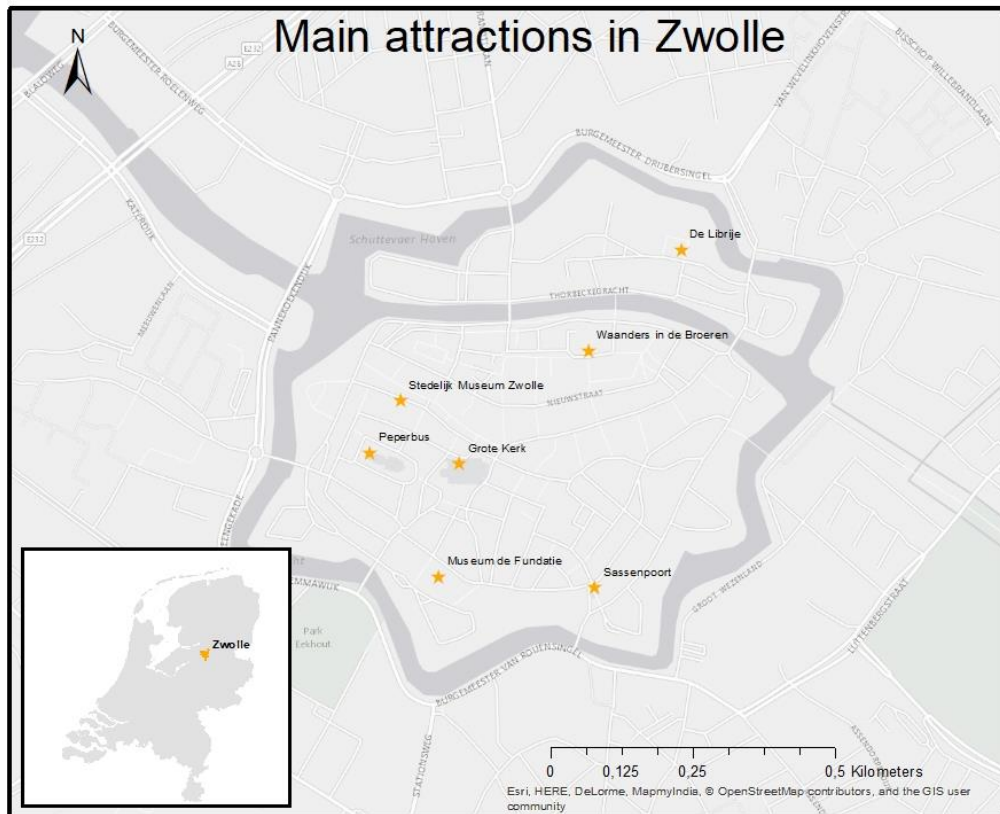


Figure 4.1 - Overview of Zwolle's main attractions in the inner city

The 1.3 million people that have visited the city did not only come for leisure shopping purposes but most importantly wanted to see and to be entertained (Gemeente Zwolle, 2017). Despite that people spends more (105 million euros in 2016), they tend not to stay that long. One of the challenges is to translate the appreciation that the city gets into longer visits and returning visitors. The *Inner City Strategic Report* states the ambition of Zwolle is to intrigue people, so they will come back (Gemeente Zwolle, 2017, p. 18-19).

## 4.2 The leisure city

As leisure plays an important role as an economic driver of urban growth - the question is how to transform (inner) cities into a vital and attractive place to be. In almost every city raises the question how to be a vital and attractive city - where can be lived, worked and people can be entertained. Zwolle has indicated its ambition to be (and stay) the most attractive inner city of the northeast of the Netherlands. Therefore the city branding's campaign is heavily focused on the image for people. The title of the cited *Inner City Strategic Report* is 'Zwolle Bruist'<sup>[10]</sup> - a place with 'experiences' for everyone. From a historic inner city to shops, cultural facilities and events: Zwolle offers different 'experiences'. In the City's agenda, frequently used words are 'hospitable', 'quality', 'surprising', 'offer' and 'accessibility' as keywords for the future city development (Gemeente Zwolle, 2014a). Within the current state of the experience economy, it is acknowledged that the city no longer relies on its main shopping function to be an attractive (inner) city. In the previous chapter, it has been stated that leisure is a post-modern model of urban development. Cities focus on leisure as branding strategy by creating new and enhancing old consumer spaces where people can consume experiences. Therefore leisure industries have become a major aspect of economic development and governmental responsibility (Jayne, 2006; McLean & Hurd, 2014; Thibault & Lavigne, 2014).

<sup>10</sup> The English translation is "a sparkling or exiting city"

Hence why the city marketing campaign of Zwolle is focussing on culture, events and the food service industry as part of a ‘sparkling’ leisure city: the city agenda positions the inner city as an ‘experience’ itself. Figure 4.2 shows a visual representation of the city development programme (Gemeente Zwolle, n.d.). This programme is themed as ‘Zwolle Gastvrij’, indicating that the city wants to enhance their experience by improving hospitality for their *guests*. The waiter is serving different possible activities within the inner city (on his plate), indicating the centre as a multifunctional attractive area with different opportunities. This variety of regions of activity suits the assumption of linkages between user, resources and region of activity of Burtenshaw (1991) and adopted by Hall and Page (2006), explained in chapter 2. For instance, different monuments and historic buildings such as the Peperbus and the Sassenpoort indicates the *Historic City*; Museum de Fundatie and cultural related images implies the *Cultural City* but music and the food industry could also be related to events and restaurants within the *Nightlife City*; and a shopping bag expresses the *Shopping City* region of Zwolle.



Figure 4.2 - Gastvrij Zwolle  
Ontwikkelingsprogramma Binnenstad 2020

All resources combine and overlap, expressing the *Leisure City* of Zwolle as a whole (Ashworth & Page, 2001, p.10). To enhance the experience economy of Zwolle, the focus of the municipality is on strengthening the food serving industry, telling the story of the city its historic culture but also enhance the accessibility of the city.

To develop the inner city in respect of the experience economy, the municipality focusses on enhancing its role as a *hospitable* city. To be a hospitable city, it seems important to investigate the current activities and perceptions of citizens and visitors of the inner city (Gemeente Zwolle, 2017). Therefore the municipality is already doing extensive research to the viability of the inner city neighbourhoods in the Buurt-voor-Buurt survey (Gemeente Zwolle, 2016). The inner city scores positive on recreational facilities such as shops, food services and culture. The recreational function of the inner city means, more spatial pressure on the centre, indicated by the lack of parking space and the lack of green (Gemeente Zwolle, 2016). Nevertheless, citizens rate their inner-city neighbourhoods on 7.6 average, which is already considered as ‘very good’ (Gemeente Zwolle, 2016).

### 4.3 The data-driven city

The agenda of the municipality of Zwolle is set on a fun, vital, accessible, livable and an attractive city - but also on doing this in a 'smart' and data-driven way. In chapter 2, it was described how the volume, velocity and variety of data have caused another shift in city governments (Batty, 2013; Kitchin, 2013; Marshall, 2012). The impact of data has also gained interest in Zwolle. The municipality of Zwolle has two main reasons to pursue data-driven policies: it could benefit societal value and they argue that cities who monitor and intervene based on new city intelligence do better (Vergeer & van Capelleveen, 2017). City intelligence based on (big) data can be referred as acting based on knowledge and insights based on data and information (see figure 1.1.) The data-fied city is a city with abstracted data about the everyday life within the city. This data can be processed into information and when analysed and applied to eventually wisdom.

By collecting data and analysing information about the inner city gains insights about the city which can be further adopted within the 'hospitable city' strategy. In a conversation with the head of Research and Information van Capelleveen and geo-information consultant Broekhaar, the city is currently experimenting with WiFi trackers that measure walking routes in the city centre (personal communication, 13 March 2017). This experiment observes people in the main shopping area and whether they walk to other streets or just stay on the main shopping streets. The goal of sensing and data collecting is to pursue a more pleasant city by having more insights where the people are. Another example of doing that is by counting algorithms in cameras. In the context of data-driven policy, there is more need for higher frequency of monitoring the city. Capelleveen and Broekhaar have stated to know a lot about the city - however, they know less about their guests and the city as a *place to be*. How do people experience the city? And when does the municipality know when they are acting right? And do the people experience what the municipality have intended? They distinguish the use of data in *hard data* and *soft data* resources, in which hard data refers to counting and tracking sensors while furthermore the municipality is interested in qualitative data, referred as *soft data* to answer those particular questions.

The main objective of this thesis is to test a topic model to analyse leisure activities of people by combining existing place information with web-based information. There is countless information available on the Web about the inner city of Zwolle - indicating places to see or activities that can be done at a certain place. This indicates which places people are talking about in Zwolle and which places attract people based on web data? The applicability of using Zwolle as a case study is fitting within its ambition of enhancing the leisure industry - such as attracting and entertaining more people so they come back or stay longer to spend more money. In respect of the developments around data and an *evidence* and *fact*-based policy, with more detailed information about the whereabouts in the city - asks for the search for the potential use of various data sources.

Web data could indicate a valuable source next to sensing, real-time, methods. Web information about places is currently the most important source to plan the next activity and forms a valuable form of expressing individual perceptions of places (Akehurst, 2009; Költringer & Dickinger, 2015). The enrichment of the inner city provides them new insights how to understand activity potential in public space. It could give valuable information of leisure behaviour based on new information sources, increasingly written by the people it selves. The insights provided could assist in the enrichment of public policy decisions regarding leisure space (and planning): but mostly reflects on whether web data is a useful source to research how people engage with their urban environment in light of the city as an attractive place to be.



# CHAPTER 5



## 5 | Results

In this chapter will be elaborated on the findings during the process of importing data and training the model to understand the activity potential of a place. The results eventually will answer the main question: ‘*how does a web-based semantic model allow to extract and analyse urban leisure activities of people, combining place information with web texts, in the inner city of Zwolle*’. The encoded and trained dataset consists of 200 unique place names within Zwolle’s inner city, 189 unique OSM identifiers, 20 activity classes, 105 sorts of referents and 62 different types of places. Because of the explorative character of this research, this chapter will focus on findings and faced constraints during training the different models.

### 5.1 Collecting place affordances and data sources

Urban areas are characterized by the perspective of the crowd, indicating places where to work, drink, eat, live, shop, sightseeing or doing other activities (Wakamiya et. al., 2011). In the *Urban Leisure* ontology different activities, places and referents are adapted to define leisure-related activities within the features of the urban environment. These classes derived from different studies in time-use and recreational activities and was further added on during the data collection. This has resulted in a long list of different classes describing activities, referents and places.

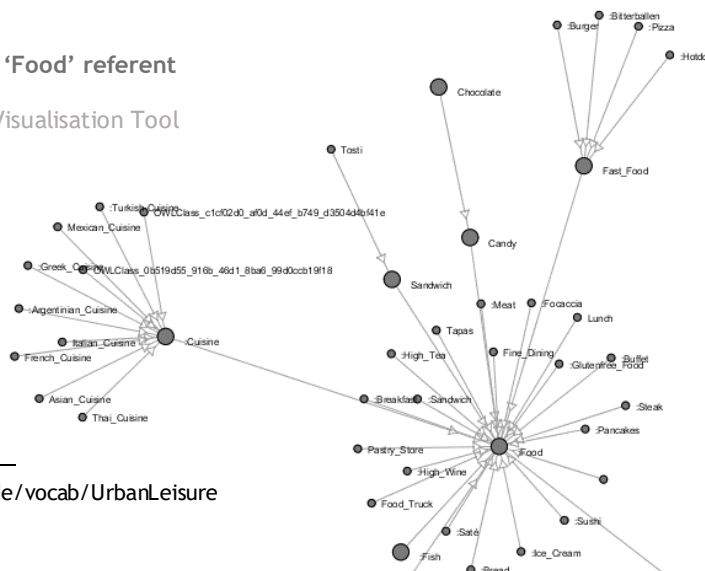
#### 5.1.1 Creating an ontology to describe urban leisure

The collection of activities in the ontology is abbreviated to fifteen main categories: ‘*Climbing*’, ‘*Cooking*’, ‘*Dancing*’, ‘*Drinking*’, ‘*Eating*’, ‘*Listening*’, ‘*Playing*’, ‘*Reading*’, ‘*Relaxing*’, ‘*Shopping*’, ‘*Sightseeing*’, ‘*Sporting*’, ‘*Strolling*’, ‘*Touring*’ and ‘*Watching*’. In this matter, the activities are involved in leisure-time, done for enjoyment in an urban surrounding. Hence, some activities can also be done without a leisure purpose such as listening to a lecture, doing some run shopping like doing the groceries or do sports professionally. The ontology further describes a range of places, described in 29 main classes, that afford some activity - such as places such as a nightclub, restaurant, café, theatre or park has an affordance which could involve drinking. Finally, the ontology consists of a list of referents in 19 main categories that are involved in pursuing an activity. Such as *festival* is a referent of a place, such as a park, when there is an event organised. This place affords other activities such as dancing and listening than the original place affordance of the park.

During the training phase of the model, the ontology was still under revision. For instance, it seemed that the probability improved when the ontology existed of more top classes such as *food* as a referent instead of further specifying food. Food is within the *Urban Leisure* ontology an extensive referent class that describes different cuisines that are served in restaurants and different dishes and snacks such as *sushi* and *focaccia* (see figure 5.1). Most of the categories are only once in the dataset, thus created more noise during training the model. The final ontology *Urban Leisure* is for further use published on the Web<sup>[11]</sup>.

Figure 5.1 - Visualisation of the ‘Food’ referent in Urban Leisure ontology

made with NavigOWL, Ontology Visualisation Tool by Hussain et. al. (2014)



<sup>11</sup> See: <http://geographicknowledge.de/vocab/UrbanLeisure>

### 5.1.2 Describing the training dataset

The ontology is used as a tool to ‘teach’ the model different activity classes, referents and places that could occur in the dataset and is used to encode places by affordance. The encoded dataset (N=326) consists of 200 unique place names within the inner city of Zwolle. Meanwhile, the dataset only exists of 189 unique Open Street Map identifiers; which means that some locations on the map share different ‘place-names’. The encoded web data described 20 unique activities in the city, including 105 referents and 62 different sorts of places. Two activities, namely *eating* (29.8%) and *shopping* (20.6%), takes up for almost half of the activity instances within the dataset. This means that about half of the dataset describes a place in which you could eat or shop. This reflects the conclusions of the report ‘Zwolle Bruist’ that the focus of the possible activities in the inner city is still primarily on shopping and the food serving industry (Gemeente Zwolle, 2017). Respectively, the third and fourth category in size are *drinking* (15,6%) and *watching* (10,7%). Watching as activity is often related to referents as *architecture* such as a monumental building or watching an *exhibition* in a gallery in the city. Other described activities have a relatively small share of the dataset (as shown in figure 5.2)

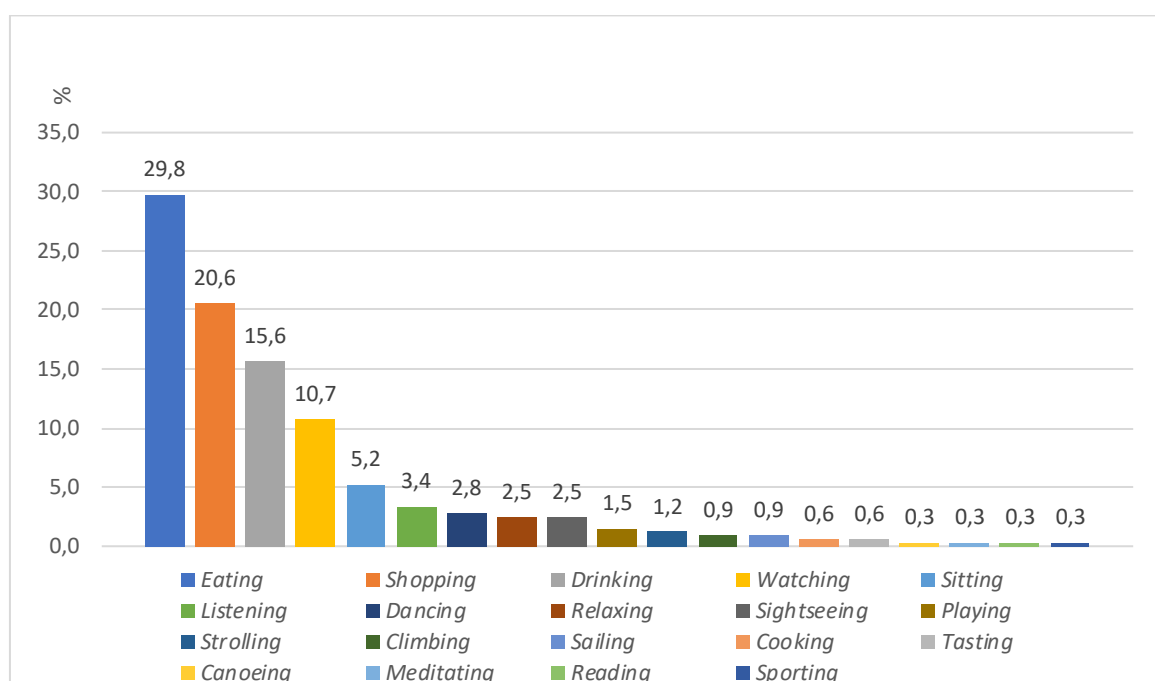


Figure 5.2 - Relative count of described activity classes within the dataset (N=326)

Also for the encoded places in the training data is seen that the largest share (37.1%) within the training dataset consists of descriptions of the food-serving industry such as *restaurant* with 23.9 percent and *café* with 13.3 percent (see table 5.1). This relates to the great share of the *Eating* category within the encoded data. It can be expected that places related to *Shopping* also has a high share of the training data. In the top 10 of the most counted places, only *fashion store* and *bookstore* are respectively ninth and tenth with a very small share between 1.8 and 2.5 percent. This can be explained when looking back at the ontology, in which 27 different specific stores are distinguished such as a *deli* and *flower store*. Just as was shown for the *Food* referent category, the *Shops* category is encoded very extensively.

While the top ten greatest place classes within the data take account for more than half of the dataset (61.7%), the share within the referent class is much more dispersed (table 5.1). The ten most encoded referents take account of 41% of the total 105 different referents in the dataset, which are often encoded just once or twice. Also here *food* is the most encoded class (7.4%). Shopping can be regarded as the second activity category in the dataset, *gifts* are the most encoded class of buying goods during shopping.



To conclude this description of the collected dataset, the food-serving industry forms the most described sector of Zwolle with almost 30 percent. Most websites describe places where you could eat some food. The image of Zwolle as ‘*Shopping City*’ seems less visible according to this dataset description, but this can be explained by the fact that websites about shops mostly are web shops and do not describe the shop itself sufficiently and therefore were not encoded.

#	Place	%_Place	Referent	%_Referent
1	Restaurant	23.9	Food	7.4
2	Café	13.2	Architecture	6.4
3	Park	4.0	Coffee	4.3
4	Square	4.0	Lunch	4.0
5	Bar	3.4	Terrace	4.0
6	Church	3.1	Gifts	3.4
7	Museum	3.1	Exhibition	3.1
8	Historic Building	2.8	Fashion	3.1
9	Fashion Store	2.5	Beer	2.8
10	Bookstore	1.8	Wine	2.8
<b>Total</b>		<b>61.7%</b>		<b>41.1%</b>

Table 5.1 - Top ten place and referent classes within dataset, absolute and relative

### 5.1.3 Reflection on collecting data

During the data collection, places and their corresponding activity and referent were retrieved from both top-down based websites as user-generated content. The methodology chapter described how this iterative process of collecting places and their affordance took place while in this paragraph will be focussed on encountered difficulties during data collection. These should be considered when discussing the results and the use of topic modelling to describe the activity potential in the city. In total, 219 unique web pages were selected which describe different activities in the inner city of Zwolle.

During the data collection, it became clear that some activities that are typically conducted in urban areas - and often described on web pages - cannot be applied and trained in the model. The problem is that some urban activities are generalized by human interpretation such as ‘*shopping in the city*’ or doing a ‘*walking tour through the city*’ do not take place in *one* unique place. Not all possible activities in the city are specifically place-bound, as is assumed in the model. These kind of activities involves wandering around the city visiting multiple places during one activity. For instance, leisurely shopping means strolling around the shopping streets without a specific shopping goal (Kim, 2006). The Diezerstraat - one of the major city streets in Zwolle - was described multiple times as a good place to go shopping but cannot be considered during the encoding process because the leisurely shopping is not as place-bound as doing shopping in a particular store on a particular street. Another important activity for people is taking a walking tour or doing sightseeing from a boat or tourist vessel. Web pages describe the possibility of the activity or tour by addressing the starting location or the rental place. But the actual activity is not bound by coordinates, or in this case by an Open Street Map identifier. Hereby the model is limited to using different scales in which ways people perceive different regions. This can be explained because the model mostly uses encoded OSM nodes, which are specific points in space (such as *point of interests*).

Another issue that occurred during the data collection, which is important to take in consideration for the representativeness of the applied model, is that not all sights in the city have a website that describes the activity in a sufficient way nor is identified as a point of interest on Google or Open Street Map. An example is the old city hall at the Grote Kerkplein. The city hall is still part of the municipality thus the official website refers to visiting hours in the city hall. Although, this is not enough text to parse that describes the place in unique fashion. On the opposite, some places are

indeed that multifunctional that it affords even more than the maximum assumed activities which means that the place affordance will not be considered in the model in its entirety. Therefore the human interpretation of a particular place could be different than is encoded in the data selection: it is possible that one associates the place with an activity that is not encoded in the used data-set.

The encoding is heavily dependent on the website and on the content of the website. The most known attractions and places are often described on multiple websites but it seemed that the information slightly differs or adds on the already encoded information. As we have restricted our model to have 5 entries each this cannot always be added to the encoded dataset. Commonly, user-generated content focus on a more specific theme such as 'hidden terraces' which you should discover in the city or the best places to eat *bitterballen*. Furthermore, some places have non-existing websites or websites that limitedly describe the certain place. They have for example a webshop as a website: the human eye can see what they can buy there but the website does not have a sufficient amount of text that describes the place to parse and is therefore not selected.

To conclude, the assumptions of the topic model causes some limitations in selecting places - while the context of the website may be good enough to interpret by the human understanding. One major issue between human and machine learning seems to be the place-bound assumption of the topic model which excludes the human perception of actions such as shopping in the city. Also, the web data itself caused difficulties to parse due non-existing websites, wrong links between places and descriptions and a number of texts available on the website to parse.

## 5.2 Training and evaluation of LDA with classifiers

In order to train a Latent Dirichlet Allocation model, an supervised LDA topic model was combined with ordinary unsupervised machine learning. For this purpose, the topic model needs to be run first. This paragraph describes the preliminary results of the first part of the modelling phase, in which summaries of topics were modelled in terms of a probability distribution over words for each topic (Ramage et. al., 2009).

### 5.2.1 Missing values and sample quality

The resulting '*training\_train\_u.json*' file contained all scraped texts generated from different ontological classifications of activities for 153 places in Zwolle. This means that the web scraping was unsuccessful for 36 selected places. This means that these web pages encountered some restriction and could not be retrieved automatically. Even fewer places obtained a review text on Google (n=66). Web scraping tools have a varying quality depending on the quality of service of the scraped website, especially from complex and dynamic websites consists of texts in different formats which sometimes cannot be scraped with the used scraping tool (PromptCloud, 2017). In addition, extracting usable data depends on a well-chosen URL. The main URL of the web page is often not leading to the most useful texts. Sometimes was therefore suggested to use one URL of a sub-page. Another explanation of web scraping failures could be related to the type of the website which gives, for example, a denial of service. Some websites have access restrictions by setting up barriers to prevent web scraping. This could be related to ethical or security reasons (ibid, 2015). To compensate for unsuccessful scraped websites, a frequently used method is to build in sleep times and to join different scraping results. Still, the training dataset consists of many missing web texts. In addition to enrichment with web texts, the training data was also prepared with different keys (table 4.3). The entities which enrichment was unsuccessful were left out of further data analysis in the training data.

### 5.2.2 Running the single-label LDA model

The modelling results of the LDA model consists of different runs with different parameters. The modelling results for each run can be found in the ‘PlaceLDA’ GitHub repository in the folder *models* [12]. This folder contains the classification results for each model which were evaluated by 10-fold cross validation, a print of the decision tree and a shapefile ‘*placetopics*’ with a subset of places taken from ‘*model1allclass*’.

In fact, two different single-label LDA topic models were trained. The first model was based on web texts while the second model was based on review texts retrieved from Google. The trained dataset based on web texts is bigger in size (n=153) than in comparison to the dataset of Google review texts (n=66). In total, a set of five models for each LDA model were trained with each focusing on different parameters (table 5.2). All models were evaluated on prediction quality in order to choose which parameters improved the model the most. Based on the evaluation of the resulting models with each different parameters, it was chosen to focus on the parameters as set in model 1 and model 2 (see for values in Appendix III, table A.1).

	Trained data	Language	useTypes	Actlevel	minclasssize
<b>Model 1</b>	<b>Web text</b>	<b>Dutch</b>	<b>True</b>	<b>True</b>	<b>5</b>
‘ <i>allact</i> ’	Web text	Dutch	True	False	5
‘ <i>allactallclass</i> ’	Web text	Dutch	True	False	0
‘ <i>allclass</i> ’	Web text	Dutch	True	True	0
‘ <i>wouttypes</i> ’	Web text	Dutch	False	True	5
<b>Model 2</b>	<b>Review text</b>	<b>English</b>	<b>True</b>	<b>True</b>	<b>5</b>
‘ <i>allact</i> ’	Review text	English	True	False	5
‘ <i>allactallclass</i> ’	Review text	English	True	False	0
‘ <i>allclass</i> ’	Review text	English	True	True	0
‘ <i>wouttypes</i> ’	Review text	English	False	True	5

Table 5.2 - Different set parameters for each trained topic model

#### 5.2.2.1 Resulting topics of model 1

Model 1 runs on ‘*training\_train\_u.json*’ with using the following parameters as ‘*web text*’ for generating 18 topics with LDA, retrieving words from the Dutch language, by using tags from OSM and Google Places as features in addition to the topic probabilities (usetypes = true), constraining the class labels to only ulo:activity classes (actlevel = true), and constraining the size of classes to contain at least 5 instances (minclasssize = 5). The class frequency distribution consists of 13 classes, but only four classes were big enough: ‘*eating*’ (70 instances), ‘*shopping*’ (33 instances), ‘*drinking*’ (18 instances) and ‘*watching*’ (15 instances). The other classes were removed.

```
trainLDA('training_train_u.json', 'webtext', language='dutch', usetypes=True, actlevel=True, minclasssize=5)
```

The trained dataset existed of 153 documents, a vocabulary size of 7067, and 34326 words. After 600 iterations, the LDA model summarizes eighteen topics (figure 5.3). Each topic represents the five most frequent words in stemmed Dutch. This means that the words are defined by their root, therefore different variations of one word are possible. For example, in topic 5 the stemmed word ‘*smak*’ could indicate ‘*smaakt*’, ‘*smaakvol*’, or ‘*smakelijk*’ (English translation: tasty). In combination with ‘*bier*’, ‘*caf*’, ‘*wij*’ and ‘*wijn*’, this topic could be interpreted as a topic that defines a place where they serve tasteful beers and wines in a café setting.

<sup>12</sup> See: <https://github.com/simonscheider/PlaceLDA/models>

Another topic that includes ‘*bier*’ in combination with ‘*brouwerij*’, ‘*blond*’, ‘*huis*’ and ‘*brij*’ summarizes the described place as a beer brewery. So a few topics can be interpreted quite well, such as topic 1: ‘*lekker*’, ‘*zwoll*’, ‘*restaurant*’, ‘*koffie*’ and ‘*lunch*’ are all words that indicate that you could have a nice lunch or coffee in a certain restaurant in Zwolle. Also, topic 0 could be interpreted as a certain place affordance namely watching movies (‘*film*’). Two words, ‘*path*’ and ‘*unlimited*’ are traced back to the biggest cinema in the city (Pathé Zwolle), where a cinema subscription is called ‘*Pathé unlimited*’. Beside all movie theatres in this category, it is notable to see other places categorized as this topic. This also includes squares such as the Grote Kerk and wellness stores.

Not every topic can be that well interpreted by the selected stemmed words. In some topics, such as in topic 8, 9, 10 and 14 it is clear that stemmed variations of the name of the retrieved website (www.leuketip.nl, Facebook or In de Buurt Zwolle) were included. These words do not say anything about the potential place affordance. Topic 8 consists of stemmed words ‘*wwwleuketipnl*’, ‘*zwoll*’, ‘*stadsgid*’, ‘*gratis*’ and ‘*verbeter*’ which describes the scraped website (where you can download free city guides of Zwolle) instead of the defined entities categorized by the LDA model. Table 5.3 shows how different restaurants and cafés are described by this certain topic 8. It is possible that these places are described in the free city guide, but this topic does not describe the activity potential of all places - which seems to be different for each place.

Vaca Negra	Doppio
Aan de Stadsmuur	Derksen & Derksen
Harm Smeengekade	Brasserie Het Vliegerhuys
AINZ	Restaurant UNO
De Gillende Keukenmeiden	Café 't Beugeltje
t Kerkbrugje	Pogo-Designshop

Table 5.3 - Places described by topic 8 in model 1

Another example is topic 12 that only consists of words that are related to opening hours. This is relevant information for people to retrieve but this topic does not say much about the place affordance of Zwolle. The resulting topics of model 1 show that artefacts have taken over in the training data which indicates that the training dataset is too small. It is expected that the evaluation by cross-validation also will show that certain noise takes over the document which can be shown by overfitting or low precision and recall scores. Therefore the fit of model 1 probably will be not satisfactory.



Figure 5.3 - Visualisation of 18 topics derived from training 'model 1'

### 5.2.2.2 Resulting topics of model 2

The second model was run on *'training\_train\_u.json'*, making use of the *'review text'* scraped from Google Places to generate 18 topics with LDA, retrieving words from the English language, by using tags from OSM and Google Places as features in addition to the topic probabilities (*usetypes = true*). The model further constrained the class labels to only the *ulo:activity* class (*actlevel = true*), and constrained the size of classes to contain at least 5 instances (*minclasssize = 5*). The class frequency distribution consists of 5 classes, but only three were big enough: *'eating'* (41 instances), *'shopping'* (10 instances), *'watching'* (10 instances). Note that the total number of instances is smaller than the number of model 1.

```
trainLDA('training_train_u.json', 'reviewtext', language='english', usetypes=True, actlevel=True, minclasssize=5)
```

The trained data existed of only 66 documents, a vocabulary size of 2155, and 5124 words. After 600 iterations, the LDA model summarizes eighteen topics in both English, Dutch and Spanish (figure 5.4). Each topic exists again of five stemmed words which were used in reviews to describe certain places. For example, topic 15 can be interpreted as describing a place where they serve some good coffee and reviews found the staff friendly. At first sight, all words within the topics seem cohesive and understandable. Also, the two encoded coffee bars *Doppio* and *Espresso Bar Maling* are actually in this topic. Nevertheless, because this dataset is actually quite small, other places are categorized within this topic while these places are less likely to be coffee bars. *Blue Sakura*, *Bar & Brood* and *Ingebugerd* are restaurants, where they likely will serve coffee and the reviews will likely be about the quality of the serving staff. The same observation can be seen in topic 6, which is very specific about a 'beautiful' church in the city. It is expected that, if people like the churches in Zwolle, this topic consists of a list of churches in the city. The *Onze Lieve Vrouw Basiliek* is the only church in this category. Topic 6 further consists of a cinema, theatres and a park.

In comparison to model 1, the resulting topics of model 2 are slightly better to interpret because there are fewer artefacts included into the topics. This might be explained as it seems that most of the training data for model 2 contain some text which reviews places related to the food serving industry, for example, good coffee or great service. Because of the small number of the dataset, which only consists of 66 documents, more than half of them are related to food reviews. This is of course quite predictable as it is more likely that people review a restaurant than a clothing store (for example). Nevertheless, as the resulting model 2 is the most unequal distributed between classes, the accuracy of this LDA model will probably lower than model 1.



Figure 5.4 - Visualisation of 18 topics derived from training 'model 2'

### 5.2.3 Evaluation of machine learning classifiers

The fit of the models was further evaluated by ten-fold cross-validation based on *precision* and *recall*, compared to the naïve classifier which is the most frequent class (see Appendix III, table A.2). Hereby the hypothesis is that the model has a good fit when the model has a high precision and a high recall. The results of all classifiers can be found in Appendix IV for model 1 and Appendix V for model 2.

#### 5.2.3.1 Evaluation of model 1

The accuracy of model 1 based on the naïve classifier is moderate [acc. = 0.518, s.d. = 0.043]. This accuracy is based on the biggest class - in this case *eating* with 70 instances. This supports the observation of the topics in figure 5.3 that the most topics contain text about places related to the food serving industry. If the classes were more equal, the naïve classifier would have been worse. Based on the naïve classifier, model 1 has a moderately weighted recall and a low weighted precision (see Appendix III, table A.2). The model selects half of the correct items, but also many of the selected items selected are not correct. In comparison to fishing, the classifier has a net that catches a lot of fish, but also quite some other things than fishes.

Attached in Appendix IV is an overview of the cross-validated results of all ten classifiers for model 1. Eight of the classifiers shows significant improvement in quality and fit. Compared to the naïve classifier, the Linear SVM classifier does not show any improvement. The only classifier showing a decrease is Naïve Bayes classification which is a very overfitted model (table 5.4). Without cross-validation, the accuracy of the model is about 0.85 (see avg/total in the second part of the table), while the weighted accuracy for this classifier is 48.6% with a standard deviation of 16% [acc. = 0.486, s.d. = 0.160]. Because each class its estimation differs because of size, the weighted precision and recall form a more reliable estimate based on cross-validation. When all classes would be more equal, the accuracy of the classifier will drop as there will be less chance of overfitting.

<b>Naïve Bayes</b>	
<b>Accuracy</b>	0.486
<b>Standard deviation</b>	+/- 0.160
<b>Weighted precision</b>	0.620
<b>Weighted recall</b>	0.486
<b>F-measure</b>	0.468

	<b>precision</b>	<b>recall</b>	<b>f1-score</b>	<b>support</b>
<b>Shopping</b>	0.89	0.52	0.65	33
<b>Eating</b>	1.00	0.51	0.68	70
<b>Watching</b>	0.74	0.93	0.82	15
<b>Drinking</b>	0.29	1.00	0.45	18
<b>Avg/total</b>	0.85	0.62	0.66	136

Table 5.4 - Cross-validated results and fit for Naïve Bayes classifier (model 1)

The two best fitted and cross-validated models are based on the classification of Logistic Regression and Neural Net. These classifiers made the most improvement on the weighted precision in comparison to the naïve classifier of model 1, which means that the model selects more correct items. By using a logistic regression as a classifier, the accuracy of the model goes up to 60.9%, which is already a considerable effect [acc. = 0.609, s.d. = 0.126] (table 5.5). Both weighted recall and specially weighted precision have increased to an also considerable score. In comparison to the naïve classifier, the F-measure shows more balance between precision and recall. [ $r = 0.609$ ,  $p = 0.638$ , F-measure = 0.491]. The model has improved both in selecting more correct items and in selecting more of the correct items.



The second part of the table shows that the model fits very well to the data for each class with an average precision and recall of 0.96. For the class *Watching*, the score shows a perfect fit while *Drinking* has the lowest recall of each class, which means not all correct items corresponding to a drinking activity are selected.

CV Logistic Regression	
Accuracy	0.609
Standard deviation	+/- 0.126
Weighted precision	0.638
Weighted recall	0.609
F-measure	0.491

	precision	recall	f1-score	support
Shopping	0.94	0.97	0.96	33
Eating	0.96	1.00	0.98	70
Watching	1.00	1.00	1.00	15
Drinking	1.00	0.78	0.88	18
Avg/total	0.96	0.96	0.96	136

Table 5.5 - Cross-validated results and fit for Logistic Regression classifier (model 1)

The confusion matrix (table 5.6) explains the second part of table 5.5, in which is shown which instances of a class in the training is classifier whereby the model. The classes in the column describe the predicted classes. This means that class *Shopping* assigns 34 instances - of which 32 are indeed about shopping and two instances are actually about drinking activities. For the class *Watching*, the model predicts all 15 instances correct which leads to a perfect precision and recall. All instances in the confusion matrix for *Drinking* are predicted correctly based on the actual class the prediction is missing instances which causes a lower recall because not all correct items are selected.

		Predicted class				
		Shopping	Eating	Watching	Drinking	Actual total
Actual class	Shopping	32	1	0	0	33
	Eating	0	70	0	0	70
	Watching	0	0	15	0	15
	Drinking	2	2	0	14	18
	Predicted total	34	73	15	14	

Table 5.6 - Confusion matrix for Logistic Regression classifier (model 1)

Using Neural Net as classifier resulted in the highest accuracy rate of 68.5% (table 5.7). This classifier was not optimized yet by the maximum iterations of the model. This means that a longer iteration might produce an even better model. This can be explained by the quite high standard deviation of 13% [acc = 0.685, s.d. = 0.130]. The weighted precision and weighted recall of this classifier are quite good. The F-measure gives a moderate score, which means the harmony between precision and recall improved [r = 0.663, p = 0.625, F-measure = 0.537]. In comparison to the logistic regression classifier, the second part of the table shows a lower average recall and precision [avg. total r = 0.85, avg. total p = 0.84]. This indicates that the fit of this model is slightly low. Especially the '*drinking*' class shows a lower precision and recall, meaning this class does not fit the data that well.

CV Neural Net	
Accuracy	0.685
Standard deviation	+/- 0.130
Weighted precision	0.625
Weighted recall	0.663
F-measure	0.537

	precision	recall	f1-score	support
Shopping	0.80	0.85	0.82	33
Eating	0.86	0.94	0.90	70
Watching	1.00	0.80	0.89	15
Drinking	0.75	0.50	0.60	18
Avg/total	0.84	0.85	0.84	136

Table 5.7 - Cross-validated results and fit for Neural Net classifier (model 1)

The confusion matrix (table 5.8) shows how the Neural Net classifier predicted instances for each class. Just like shown in the confusion matrix of logistic regression, the *Watching* class shows that all predicted instances are correct which gives a perfect precision. Not all correct items are selected which means the recall of this model is less satisfactory. Both *Shopping* and *Eating* has predicted instances that actually belongs in the other three classes which causes a lower precision. It also predicts too many instances, which influence the recall. The *Drinking* class misses instances and not all instances are predicted correctly. Because this class is smaller compared to the *Shopping* and *Eating* class, this influences precision and recall more which is the lowest as shown in table 5.7.

#### Predicted class

	Shopping	Eating	Watching	Drinking	Actual total
Shopping	28	3	0	2	33
Eating	3	66	0	1	70
Watching	2	1	12	0	15
Drinking	2	7	0	9	18
Predicted total	35	77	12	12	136

Table 5.8 - Confusion matrix for Neural Net classifier (model 1)

The accuracy of both Logistic Regression and Neural Net are already quite good - they are the best fitted cross-validated classifiers. When new data will be added, these classifiers have the best quality to anticipate on that data. The improvements merely focus on the weighted precision, which means that the classifiers improved with respect to a naïve model by selecting more items that are correct. Both the weighted precision and the weighted recall are around 60% to 70% which is a good score. The F-measure of Neural Net is improved to a moderate score above 50%. The Decision Tree for model 1 can be used to check whether the model makes plausible decisions (Appendix VI). The decision tree starts checking for the 136 samples in the training data if the rule for topic 15 applies. This topic describes words such as *café*, *shop* and *special* and classifies this topic within the *Eating* class. When this rule applies, the condition has met for 90 items within the sample and it checks for the following condition. Again for the *Eating* class, the rule checks if it applies to topic 0 which describes a cinema theatre in the city. When this condition is not met, this rule is considered as false. You can see the distribution of the sample becomes smaller and more unequal. At this point, there are no drinking items in the tree anymore and a small number of the other three classes. The next rule checks the conditions for topic 4 which consists of adjectives such as *small* and *beautiful*. When this rule does

not apply, the next rule is again about the *Eating* class which is logical because of the great share of *Eating* related instances in the training data. The next condition checks for topic 16 for four instances in the sample. Topic 16 describes foundation Langhuis which is a cultural institute with art expositions. When this condition is met, only one instance has been selected within the Shopping class.



Figure 5.5 -Topic 15, topic 0, topic 4 and topic 16 as discussed in the Decision Tree (from left to right)

A decision tree visually represents the decisions made by the model by using a ‘flipping coin’ decision method. Fundamentally, the more unequal the sample eventually is, the better the made decisions are. A risk of decision-tree learners is that they could not generalise the data well. The described example showed a somewhat biased result: an instance which was related to *shopping* was true to the condition of a topic which describes a place related to culture and art. In human interpretation, it would be more likely that people visit this kind of places to watch art, instead of shopping. It could be concluded that this item is a false positive selected item. Nevertheless, the fit of the model is not yet yielding a “satisfactory” prediction quality of above 80%. The classifiers still select many items that are not correct or do not capture many of the correct items in their predictions.

#### 5.2.3.2 Evaluation of model 2

Also, model 2, which is based on review texts, is evaluated by ten machine learning classifiers. The accuracy of the naïve classifier is moderate [acc. = 0.671, s.d. = 0.014]. Again, this naïve classifier is based on the biggest class which in this case is *Eating* (see Appendix III, table A.3). Based on the naïve classifier, the model has a moderate weighted recall and a low weighted precision [ $r = 0.671$ ,  $p = 0.451$ , F-measure = 0.26]. The model selects quite some items that are correct but also many of the selected items are not correct.

Attached in Attachment V is an overview of the cross-validated results of all ten classifiers, this time for model 2. Four of the ten classifiers shows an improvement in quality and fit. It is noticeable that the classifiers Linear SVM, RBF SVM and Gaussian Process give the exact same result as the naïve classifier. The weakest classifier is this time Decision Tree, which shows the biggest decrease in accuracy in comparison to the naïve classifier (table 5.9). The accuracy drops to 63.8% [acc. = 0.638, s.d. = 0.127]. The weighted precision slightly improves to a moderate score. Nevertheless, the weighted recall slightly drops. The model became better in selecting items that were correct but became worse in selecting correct items. In comparison to the naïve classifier, the F-measure is quite improved because precision and recall are more in harmony.

CV Decision Tree	
Accuracy	0.638
Standard deviation	+/- 0.127
Weighted precision	0.585
Weighted recall	0.655
F-measure	0.655

	precision	recall	f1-score	support
ulo:Eating	0.98	1.00	0.99	41
ulo:Shopping	1.00	0.90	0.95	10
ulo:Watching	1.00	1.00	1.00	10
Avg/total	0.98	0.98	0.98	61

Table 5.9 - Cross-validated results and fit for Decision Tree classifier (model 2)

Based on accuracy and the principle of precision and recall, the classifiers Nearest Neighbours and Neural Net seem to be the best fitted cross-validated classifiers. Based on Nearest Neighbours, the accuracy of the model improves to 76.6 percent [acc. = 0.766, s.d. = 0.133]. Also the weighted recall and the weighted precision both improves considerably [r = 0.767, p = 0.699, F-measure = 0.593]. The second part of table 5.10 shows that this classifier is more strict on precision and recall for evaluating model fit for each class. The classes *Eating* and *Shopping* are predicted quite satisfactory while *Watching* now has a moderate recall which means that less of the correct items are selected in comparison to the naïve classifier.

CV Nearest Neighbours	
Accuracy	0.767
Standard deviation	+/- 0.133
Weighted precision	0.699
Weighted recall	0.767
F-measure	0.593

	precision	recall	f1-score	support
ulo:Eating	0.86	0.93	0.89	41
ulo:Shopping	0.78	0.70	0.74	10
ulo:Watching	0.75	0.60	0.67	10
Avg/total	0.83	0.84	0.83	61

Table 5.10 - Cross-validated results and fit for Nearest Neighbours classifier (model 2)

When looking at the confusion matrix of the Nearest Neighbours classifier, it shows for each class that the most instances are predicted correctly which influences the precision (table 5.11). Not all instances are predicted correctly and also *Shopping* and *Watching* are missing instances. That means that not all correct items are selected by the Nearest Neighbours classifier.

		Predicted class			
		Eating	Shopping	Watching	Actual total
Actual class	Eating	38	2	1	41
	Shopping	2	7	1	10
	Watching	4	0	6	10
	Predicted total	44	9	8	61

Table 5.11 - Confusion matrix for Nearest Neighbours classifier (model 2)

The best classifier is Neural Net, which shows an overall good improvement in comparison to the naïve classifier (table 5.12). The accuracy rates increases to 78%, with a 16% standard deviation [accuracy = 0.783, s.d. = 0.167]. The accuracy of the classifier is considerable. Based on a moderate weighted precision and a satisfactory recall, the topic model selects more often the correct items [ $r = 0.767$ ,  $p = 0.66$ , F-measure = 0.560].

CV Neural Net	
Accuracy	0.783 (+/- 0.167)
Weighted precision	0.66
Weighted recall	0.767
F-measure	0.560

	precision	recall	f1-score	support
ulo:Eating	0.95	1.00	0.98	41
ulo:Shopping	1.00	0.90	0.95	10
ulo:Watching	1.00	0.90	0.95	10
Avg/total	0.97	0.97	0.97	61

Table 5.12 - Cross-validated results and fit for Neural Net classifier (model 2)

The second part of the table shows the fit for each classifier. For both *Shopping* and *Watching* the precision, the precision is perfect which means all selected items are correct. This can be confirmed by the confusion matrix of the Neural Net classifier (table 5.13). All items are predicted correctly as *Shopping* or *Watching*. The recall is not perfect because both classifiers miss instances. The biggest class, *Eating* has a perfect recall that means that all correct items are selected. In the confusion matrix is seen that all 41 instances are within the correct category. Nevertheless, this time the precision is not perfect because the classifier also predicted two instances as *Eating* while they are actually in the other two classes.

		Predicted class			
		Eating	Shopping	Watching	Actual total
Actual class	Eating	41	0	0	41
	Shopping	1	9	0	10
	Watching	1	0	9	10
	Predicted total	43	9	9	61

Table 5.13 - Confusion matrix for Neural Net classifier (model 2)

The accuracy rates of Nearest Neighbours and Neural Net for model 2 are more satisfactory already than the two classifiers for model 1. The improvements focus on overall improvement in accuracy, weighted precision, weighted recall and the F-measure. This means that both classifiers succeed in selecting more correct items and that the selected items are more often correct. The accuracy rates are between 75-80% which is close to a satisfactory prediction quality. Again can be looked at the decision tree for the quality of the decisions made by model 2 (Appendix VII). The decision tree for model 2 consists of topics and tags derived from OSM and Google tags. As an example, the outer right line of the tree will be discussed. The decision tree starts checking for a condition whether the items within the training data involves a shop. When this condition is considered as false, it checks for the remaining 50 samples for the condition of topic 6, which consists of adjectives as *beautiful* and includes *church* within the *Eating* class. When also this condition is considered as false, it checks for topic 16 - about a *nice building* or *museum* - if the activity *Watching* can be applied.

Also when this condition is considered as false, it checks whether the activity class is related to a place of worship amenity. At this phase, only three items are in the sample. When this rule again did not apply, the model considers the remaining item to the shopping class. Somewhat unlogic is that when the activity involves the place of worship amenity, it will be considered within the *Eating* class.



Figure 5.6 - Topic 6 and topic 16 as discussed in the Decision Tree (from left to right)

The decision tree proves that when evaluating model 2, you should be aware of overfitting due a small testing sample (N=66). It could lead to improbably made decisions and to a decision tree with a not satisfactory quality. During the evaluation, it seems when the sample is starting to be very small it is likely to be sensitive for overfitting. Both *Eating* or *Shopping* do not derive from a place of worship amenity. In a matter of fact, eating would not be accepted in such places. The accuracy rate of model 2 is close to satisfactory, but in this case should be considered that model 2 is based on a small training set (N=66) which quicker causes noise within the prediction quality.

### 5.3 Training and evaluating L-LDA

In order to advance the found results, the current Latent Dirichlet Allocation topic model was improved by adding *labelled* supervision on the topic model that models web texts. The L-LDA topic model does not use types, which means that the tags from Open Street Map and Google Places were not taken into the feature vector. This can be explained by the fact that L-LDA only focuses on the distribution of topics within the text and gives an idea which topic is the most important in the text. Therefore it is not relevant to enrich the text by tags from OSM and Google Places. The language is set as Dutch as the web texts are in Dutch. L-LDA is restricted to activity classes only and uses all activity classes to look which activity class is described within the text.

```
trainL-LDA('training_train_u.json', 'webtext', language='dutch', usetypes=False, actlevel=True, minclasssize=0)
```

The L-LDA produces labelled topics with a top twenty of word stems from 18 activity classes. All labelled topics are summarized in Appendix VIII. The resulting topics differ from the previous LDA model in terms of that the topics are directly created to its corresponding activity class (Ramage et. al., 2009). Therefore each activity class consists of words that describe a certain activity and its relation to the city of Zwolle. Some of the topics do not make sense because in a small training dataset, noise can be easily picked up. Also, the very uneven distribution of all activities within the training data, as seen in figure 5.2, causes not evenly strong topics.

One of the classes with a few instances is *Canoeing* (0.3%). Figure 5.7 shows the top twenty word stems associated with the activity Canoeing. The word stems do not say anything about the activity itself. Instead, the topic describes words in relation to a (children's) birthday party such as '*verjaardagsfeestje*' and '*kinderfeestje*'. Furthermore, it describes facilities such as '*toilet*' and '*kiosk*' which could relate to the rental facility but are not specifically related to the activity itself. When looking at the encoded sheet which was the input for the training data - only one entity was encoded in the canoeing activity class. The selected website was an event on Facebook inviting people to paddle around the moats in Zwolle. By only selecting one source, the L-LDA topic model is very sensitive to pick up irrelevant words. This is due to a very small sample size which restricts the model and causes substantial noise. The unequal distribution of activity classes in the encoded dataset shows

that this can easily take over in the training data and causes irregularities in the results. It can be concluded that a certain activity class should have a minimum size to have a quite satisfactory prediction quality.



Figure 5.7 - Labelled topic: Canoeing

*Eating, Drinking, Shopping* and *Watching* are the biggest activity classes in the training data with respectively 29.8%, 20.6%, 15.6% and 10.7% of the total dataset (see figure 5.2). These activity classes are the only classes with more than five entities as seen in the results of testing the prediction quality of the LDA topic models. *Shopping* is the second biggest category while the top twenty word stems associated with *Shopping* does not say much about the activity (see figure 5.8). The topic includes 'winkel' which is Dutch for a store and describes a certain warehouse 'Pistache' in the city. Beside those words, this topic does not exactly describe shopping activity. The topic caught up the website [www.leuketip.nl](http://www.leuketip.nl), as also was seen in the LDA topics. Also, this labelled topic has some substantial noise caused by the sample.



Figure 5.8 - Labelled topic: Shopping

Contradicting to *Shopping*, the class *Eating* shows a quite cohesive topic (figure 5.9). The topic has related words and includes the verb eating itself. It mostly refers to review words related to eating - such as 'heerlijk' and 'lekker' which both means delicious. It further includes 'restaurant': the place where could be eaten. Also 'Italiaanse' is included which refers to the Italian cuisine. This could make sense that there are a lot of Italian restaurants in the city. Also, the activity label *Watching* is pretty good described by the results of the L-LDA topics. Figure 5.10 shows both words referred to a gallery (such as *galerie, exposities, kunst*) and to a cinema (*path, unlimited*). Both going to the cinema and gallery are considered as a watching activity.



Figure 5.9 - Labelled topic: Eating



Figure 5.10 - Labelled topic: Watching

Both *Shopping* and *Eating* make more sense than for topics with a very small instance such as *Canoeing*. Hereby can be concluded that to restrict substantial noise, the activity classes should have a larger sample. Not only the number of entities matter, as *Shopping* is second biggest class in this dataset it is not as satisfactory compared to *Eating* and *Watching*. The sample is also independent of the quality of the selected website. As stated previously, the *Shopping* labelled topic does not include a lot of word stems that refer to the activity. An explanation could be that websites that describe a store often also includes a webshop or other non-relevant web artefacts that are being picked up while scraping the website. The errors of non-relevant word stems and the small sample size restrictions should be averaged out in a larger sample consisting of up to thousands of documents.

Just as the single-label LDA model, the next step should include an evaluation of the results of labelled LDA by ten-fold cross-validation in order to compare both models. While the single-label LDA model was approached with machine learning classifiers on top (LDA + ML), with the multi-label L-LDA model we have found difficulties in working with single-label classification predictors (Scikit-Learn, 2017c). Because the resulting L-LDA still works as a multi-label class, comparing the results of LDA and L-LDA are concluded not to be straightforward in order compare because of a difference in difficulty level of the output. The single-label LDA + machine learning classifiers modelled all data into eighteen different topics. Hereby it reasons back to the bag of words assumption: each topic stands for a different 'bag' with words describing a specific class. While L-LDA reasons from multi-labelled topics which do not directly support for a confusion matrix to evaluate the results based on precision and recall. Therefore a comparison cannot be made.

The original paper in where the L-LDA method was proposed (Ramage et. al., 2009), the evaluation was mainly done by human interpretation of the topics, just as was done with the word visualisations in this thesis. Translating the results of L-LDA in a single-label classifier in order to make it work in a confusion-matrix will lose the relevance of L-LDA, that trains data directly on documents. In machine learning, this problem is referred as the *multi-label classification* or *multi-class classification* problem (Jain, 2017; Sci-Kit Learn, 2017). Multi-label classifiers such as L-LDA assign to a document a set of labels, in this case, activity classes. They predict properties of the labels in a document which are not mutually exclusive or equally distributed. Without cross-validation, the results of L-LDA are over 90%, which seems pretentious high. Cross-validation is a method to measure the accuracy of the predicted set of labels and how well they match with the true set of labels. There are two ways to adapt to the problem of multi-label classification. One is by transforming the method or by adapting the algorithm (Sci-Kit Learn 2017, Jain, 2017). Transforming indicates transforming the data into a set of binary classifiers which can be handled similar like using single-classifiers. Algorithm adaptation tries to directly perform multi-label classification than rather simplifying the problem. However, even by addressing this problem, the difficulty remains that the results of L-LDA still needs to be meaningful as multi-label predictor. Hence can be understood that machine learning still needs some improvement to grasp human understandings of place.



## 5.4 Visualising place affordance

In order to show the geographic potential of using web semantics, the results of the LDA topic model were visualized on an interactive map. As discussed in the methodology chapter, the Python script also included a script that creates a shapefile. The used shapefile is based on the modelling results of *model 1* with the *all class* parameter set. The parameters enable the model to only use the activity classes. As no minimum class size was set, the model does not filter out any activity class.

```
trainLDA('training_train_u.json', 'webtext', language='Dutch', usetypes=True, actlevel=True, minclasssize=0)
```

The resulting shapefile from the Python script consists of a description of 84 places including geo-references. Considering the  $n=153$  documents, the model could not retrieve coordinates for each place within the training dataset. Consequently, topic 0 referring to *path, unlimited, film, nam, zwoll* has not any references to places, meaning the topic is left out of the visualisation. The remaining places, including deriving topic, are shown in an interactive map which is published in the *Tableau Public Gallery* <sup>[13]</sup>. It is recommended to view the map online. A screenshot of the visualisation is shown in figure 5.11. The map is interactive in such way that when hovering over an item on the map, more details of the selected point becomes visible. It shows the place name as encoded in the data collection and it shows the description of its place affordance based on the topics that have derived from the LDA model. Furthermore, in the legend, the different activity clusters are all shown (which corresponds to a topic). When clicking on a cluster (topic), the map highlights each place that matches this certain cluster (topic) according to the results of the topic model.

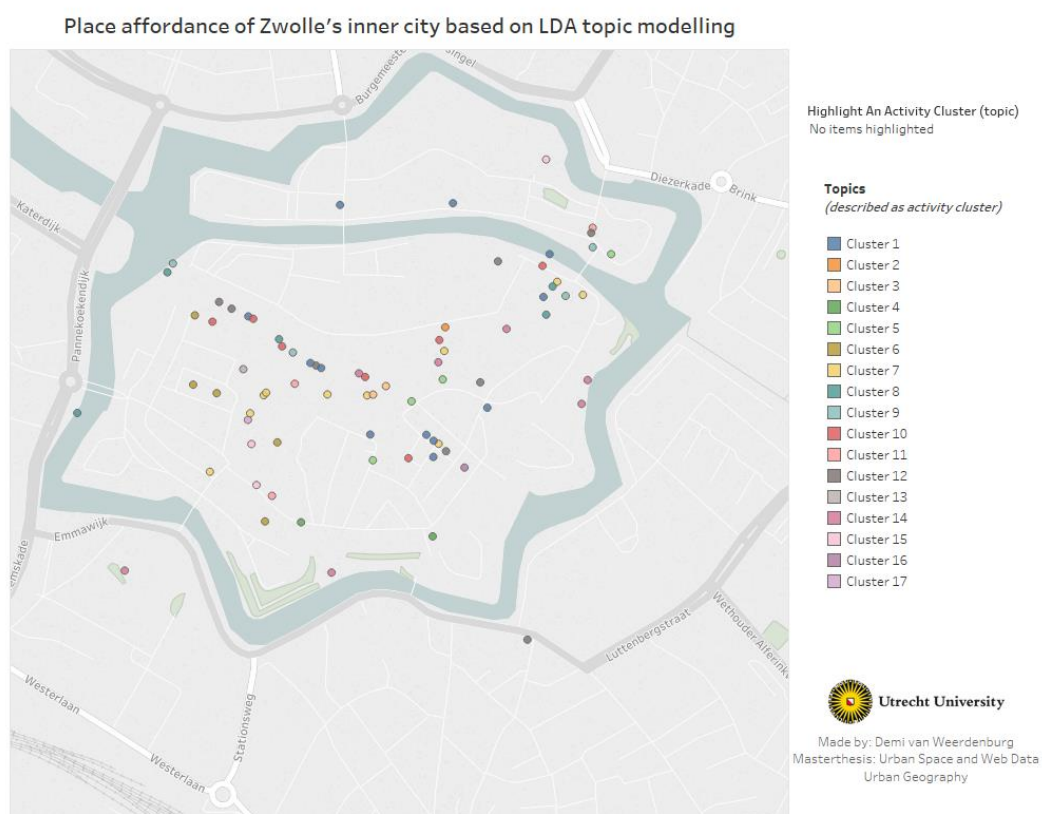


Figure 5.11 - Map of the place affordance of Zwolle's inner city based on LDA topic modelling  
Interactive map is published in the Tableau Public Gallery

Although the map shows some errors as the accuracy of LDA was not yielding to satisfactory because not all topic descriptions make 'sense' to understand the affordance of the place. The deriving description based on the topics does not suit all related places when observing the place name. Nevertheless, this kind of visualization shows its potential use for combining data resources and for understanding urban areas of interest.

The map represents a set of web texts which describes the inner city in all kinds of ways written by different sources. The map shows points close to streets such as the Melkmarkt, Diezerstraat and the Grote Markt surroundings. This result is not surprising as this area forms the core of the inner city: the Diezerstraat is, for example, the main shopping street with multiple chain stores. This confirms its popularity among the public based on web information. As a point of saturation was reached within the boundaries of selecting web texts, it shows the places that are discussed and experienced by people.

The resulting dashboard is published as visual on Tableau Public: an online gallery service of Tableau which makes it possible to embed visualisations (called "vizzes") into web pages, blogs, social media, email or can be made available for download to other users. It fits the explorative use of this thesis to make the data and script publicly available (both on GitHub and Tableau) in order to encourage others to work with the dataset. The results eventually can encourage to be re-used in different ways. In light of urban big data and the open data movement, data should be available and re-used by its citizens and others to give new value, insights and knowledge to the existing data. This fits within the ideology of *smart cities* in which people consequently becoming more aware of the data layer in the city and demand access. Cities respond by building platforms and open dashboards which both improves the municipal service but also the understanding of the urban quality of life (iBestuur, 2016; Kitchin).

The resulting visualisation still seems quite static and limited to show certain *areas*. Hobel et. al. (2015) have argued that an activity cannot be seen as a single point in space but involves a cluster of places in relation to an activity. *Areas of interest* contain multi co-located features based on similarity, proximity or related activities. This can be explained by the limited dataset that is used. The current size of the dataset, including the results are not that satisfactory because the size is relatively small. Next to adding more data, the results potentially could be combined with more real-time statistics and data to become a true dynamic dashboard that monitors real-time dynamics of the city over time and space (Kitchin, 2014, p. 106).





## 6 | Conclusion

This final chapter discusses the four underlying sub-research questions which were composed in order of the sequence within the research project and data analysis. These sub-research questions together will eventually answer the following main research question:

*How does a web-based semantic model allow to extract and analyse urban leisure activities of people, combining place information with web texts, in the inner city of Zwolle?*

The main objective of this thesis was to test a method that allowed enrichment of urban leisure-related activities based on combining place information with web texts. Therefore a probabilistic topic model, Latent Dirichlet Allocation, was used to extract web text, enrich this with other Web data and a model was trained that could estimate any activity class based on those web text - which were made and used by the consumers themselves.

### 6.1 Answers to the research questions

#### 6.1.1 Sub-research question I

The first sub-research question is: *how can web-based text semantics be used for understanding and analysing activity potentials of people in certain urban spaces*. Web-based texts semantics is a growing field in which different methods are explored, trying to grasp and understand natural language in order to represent this in a digital context. By (probabilistic) topic modelling, the structure of large samples of documents with thematic information can be revealed (Blei, 2012). The tool is already widely used to describe shared opinions, life, topics or social incidents and used to assist in better decision-making (Huang & Li, 2016).

The gap this research is trying to fill is between the production of web-based geographic information and its potential to enrich our knowledge about the relevance of places in the digital sphere. For leisure purposes, web information has become one of the most important sources to search for, plan and share activities and perceptions about the city (Költringer & Dickinger, 2015). Due to developments of user-generated content (UGC), the Web increasingly voices the experiences of people itself - described by Guo et. al (2017) as representing the '*voice of the consumer*'. Therefore its value seems unprecedented as it voices the activities of people and determines them what to do (Adams & Janowicz, 2014; Hobel et. al., 2015 & Hu et. al., 2015).

Web semantics, which can be defined as '*learning the meaning of*', can be used to retrieve those perceptions based on the web texts as it makes use of natural language that deals with the ambiguity of places. Topic modelling can retrieve a big volume of different documents - webpages where people are writing about a certain city - and could reveal the voice of the consumers in understanding where they are writing about (Guo et. al., 2016). Hereby it deals with multiple activities and perceptions. By using these web pages, which describes activities in the city, and by using a topic model such as Latent Dirichlet Allocation, we could describe activities by the words that the consumers themselves are using to describe the city and/or activity.

This information deriving from LDA enrich the characterization of urban areas as it bundles different web pages and shared opinions on different topics that describe the city (figures 5.3; 5.4; 5.7; 5.8; 5.9; 5.10 and Appendix VIII). Knowing the resulting topics are useful for city planners. As increasingly more people will use cities for leisure purposes and the leisure industry becomes more relevant for urban policies, it could optimize the understanding of the place representations of people and human behaviour in place as it summarizes how people write about a certain place. It could help understanding, analysing or generalising urban areas by bigger datasets based on geographical data. However, as has been shown in the results section, web semantics are not robust for artefacts on websites.

### 6.1.2 Sub-research question II

The second sub-research question is as follows: *which urban activity categories can be distinguished regarding urban leisure in the inner city?* The urban environment enables lots of different activities but these have to be acknowledged in topic modelling, therefore this question focusses on an important methodological step during the data analysis. As result, the ontology *Urban Leisure* was created and published on the Web. Urban leisure was defined as *a (meaningful) activity done in the free-time, after obligations such as work and the household, outside the house in a state of enjoyment, and obviously making use of the features of the urban environment.*

The ontology captures three concepts related to an urban leisure activity; the *affordance* itself, the *referent* that is involved in the activity and the *place* that affords the activity (Scheider, 2017). As discussed in paragraph 5.1, the resulting ontology *Urban Leisure* consists of fifteen main categories that describes an activity that could be performed in an urban environment. These activities derived from different (traditional) time-budget studies and were added on during the data-collecting. This indicates that creating an ontology is an iterative process: when you add new data which describes something new, you should teach the model as well.

This further means that the ontology was constantly revised, refined and cleansed and moreover means that the ontology is never in its final version implying that it completely describes urban leisure as a concept. Noy and McGuiness (2001) have stated that creating an ontology is very dependent on the scope of the ontology and the determined use. As this ontology is created for a data collection and data modelling for a middle-sized Dutch city, it is merely focussed on the affordance of like-wise cities.

The ontology further described a range of places in 29 main categories, that could afford some activity and a list of referents in 19 main categories that are involved in pursuing an activity. It was concluded that in order to use an ontology in topic modelling, the ontology needs to be a cleansed version with very straightforward classes, as was explained for the *Food* class in figure 5.1. In relation to the quote by Adams and Janowicz on page 16, it seems that in topic modelling we cannot come explicitly close to a human perception of place as the topic model only could adapt straightforward classes while in human thinking there is, for example, a difference between different cuisines or meals.

### 6.1.3 Sub-research question III

The third sub-research question asks *which web sources can be used to train a dataset for a web-text semantic model?* There are endless sources of Web data available. The importance of web information is that it is suggested to be the major source where people gaze upon that determines what they are going to do (Fabrikant & Battenfield, 2001; Wakamiya et. al., 2001). Magasic (2016) explained that the image or perception a certain area has is influenced by what is shared on the Web. People are inspired by activities they read about online. At the same time, as consequence of user-generated content, when people write about their perceptions, they inspire others to do the same (Költringer & Dickinger, 2015). Eventually, this creates a vicious circle of sharing and retrieving web information. Furthermore, increasingly more 'things' become connected to the Web, feeding amounts of data and streams of potentially usable information. However, not all data available is suitable for use in a Latent Dirichlet Allocation topic model. There are a few assumptions that restrict selecting some data sources while they remain relevant in urban analysis.

First of all, logically, the website should consist of enough *text* to parse. Currently, cities all over the world are experimenting with using different sets of data and sensors in public space. Big datasets deriving from smartphones, wearables, GPS and other digital sensors are used in urban policies (Gao et. al., 2017). This data can be considered as quantified (*hard*) data and do not consist of any words so is unable to parse. Hence the focus on during the data collection is on *soft*, qualitative data sources. Nevertheless, this can be seen as an interesting data source that seems to be underexposed in the data-driven approach of cities. It expresses the perception and experiences of people or institutions who wrote texts and reviews in which they describe a certain activity at a certain place. Kitchin (2013) stresses that not only algorithms and measurements could describe the city but that human responses on the Web create a valuable source while this is yet quite neglected in current big data research and use.

Other assumptions further restrict the selection of web pages as a place should be described uniquely on the webpage and the place should be contained in an Open Street Map identifier as well.

In this thesis, a further distinction was made between institutional web information which was created top-down by institutions as the municipality itself and user-generated content (UGC) as a valuable source as being an interactive source with rich geographic information (Goodchild, 2007b). Different authors even stress how UGC can rethink the city from the bottom up (Goodchild, 2007a; Townsend, 2013). However, the information both sources generate can be biased in both a positive or negative way. As institutional data is used to *promote* the city and users who write reviews out of *frustration* could cause differences in perception per website. Furthermore web texts could provide fake information or are computer-generated. Nevertheless, in the light of the aspiration of being fun, vital, accessible, liveable and attractive, it remains relevant to grasp this kind of data sources within (big) data analysis as well to measure their activities and the experiences of people.

During the encoding, some difficulties were encountered because of the restrictions in the LDA topic model, which further indicates a problem between human understanding and machine learning. An example that has been given is the activity *leisurely shopping in the city*, whereas there is no specific place-boundedness to this activity. The other way around, some places contained more activities that actually modelled which were restrained by the encoding. This implicates that the data used in the assumed LDA model is very restricted and implicated different kinds of web sources and interpretations.

As result, the collected dataset consists of 326 documents that describe activities in 189 unique OSM identifiers. Because the encoding was done manually in an iterative way, this causes a scalability problem as we will later see that the encoded training dataset is too small for accurate results. On the one hand, as it seems that a point of saturation was reached in the way the encoding took place it can be questioned if there are any more places to parse. Eventually, more sorts of web data can be used to address this problem. But even if we can address this problem, as the encoding of web pages was done manually this leads to a very time-consuming process.

#### 6.1.4 Sub-research question IV

The final sub-research question is: *to which extent can urban leisure activities be estimated, extracted from knowledge from a web-text semantic model*. The modelling phase consisted of a single-label and multi-labelled variant of Latent Dirichlet Allocation. The single-label variant was run both on web texts and review texts only, which showed both a quite good prediction quality for some machine learning classifiers, as discussed in paragraph 5.2. When focussing on the web texts model, it showed an accuracy between 60.9% and 68.5%, which is a considerable effect. The weighted precision and recall resulted in a score of 60% to 70%. This means that the model was considerable good at selecting the correct items. LDA model 2 even showed a higher accuracy rate between 75-80%, but hereby must be implicated that this was based on a very small dataset ( $n = 66$ ), indicating a great chance of overfitting because of it easily takes over noise within the prediction quality. When looking at the resulted 18 topics of the web texts model, it showed some good interpretable topics such as topic 5 but also some topics with non-related artefacts such as topic 8 shows a less interpretable topic (figure 6.1).



Figure 6.1 - Topic 5 and 8 (model 1)  
From left to right

While topic 5 clearly describes a café where beer and wine are served, topic 8 is heavily influenced by noise within the data such as ‘*wwwleuketipnl*’ or ‘*stadsgids*’, which does not say much about any particular activity and cannot be interpreted by human understanding. The relatively small training dataset of model 1 (n=153) shows that topic modelling is quite sensitive to any noise within the data which eventually causes overfitting on the training data. To conclude, the LDA method was quite adequate but a too small training dataset causes too many irregularities in the model.

Next, to the single-label LDA model, the multi-label L-LDA model should show improvements by restricting the model and *teaching* the model with the created *Urban Leisure* ontology. The labelled LDA model produced labelled topics (activity classes) which each consisted of words that describe the activity and its relation to the city. Some of the topics were also not human readable because of it has picked up a lot of noise or irrelevant information. This is also related to the small number of data and an unequal distribution of activity classes as encoded in the training data (see figure 5.2). The relative count of the described activities within the collected dataset is not equally distributed. This shows clearly a difference between unfrequent encoded activity classes such as Canoeing and frequently encoded activity classes such as Eating and Watching. The four activity classes with more than five instances, *Eating*, *Shopping*, *Drinking* and *Watching* take up more than 75 percent of the training data set. It can be concluded that, in order for the model’s prediction quality, the obtained information in the trained dataset is too limited for detailed classes. Nevertheless can be questioned if within a middle-sized Dutch city all activity classes as encoded in the ontology are evenly prominent. As in Zwolle the major economic and leisure functions are shopping and the food-serving industry (Gemeente Zwolle, 2014a), it is not too surprising to see this categories stand out.

In order to answer the question how good the presented models are in estimating activities for human interpretation, this cannot be easily compared. A major issue that we have encountered is the multi-class classification problem. The multifunctionality of a place is difficult to take into account when evaluating the results with machine learning classifiers. The difficulty remained that even by the suggested solution of simplifying the results of L-LDA, the results are questioned to be meaningful, indicating that it needed to acknowledge in machine learning that places are multifunctional areas and people associate places differently based on different spatial footprints (Hobel et. al., 2015). Both single-label and multi-label LDA models have its own restrictions in doing so. Therefore can be concluded that the gap between human interpretation and computational representation still remains unsolved as cities are mostly vibrant and multifunctional areas.

#### 6.1.5 Main research question

The four sub-research questions together answer the main research question: *how does a web-based semantic model allow to extract and analyse urban leisure activities of people, combining place information with web texts, in the inner city of Zwolle?* The prevalence of geographic information and the need for more intuitive methods of representing place have increased the demand for machine understanding of place (Jones et. al., 2008). The used Latent Dirichlet Allocation topic model was based on both institutional web texts and user-generated content, to retrieve the place perception of the people about different leisure activities. It was argued that this indicate the perceived space of the inner city of Zwolle by using the ‘*voice of the consumers*’. By exploring these data sources and using topic modelling was sought how the urban areas of interest can be conceptualised by machine learning using this data sources.

Despite the relatively small training dataset and implications, its prediction quality and the resulted topics have already shown potential for using web semantics in enriching our knowledge about urban space and leisure, as some of the topics were quite good to be humanly interpreted. The potential of web-based semantic modelling is proven, especially when is acknowledged that a bigger dataset will eventually filter out the noise within the topics. This is shown by differences in the results of the activity classes whereas bigger classes seemed to be more cohesive. Furthermore, this unequal distribution does tell something about the affordance of the entire inner city of Zwolle which is mainly focussed on the food-serving industry and shopping, which is not a surprise (Gemeente Zwolle, 2017).



As Zwolle wants to become a vibrant *'place to be'*, it is interesting to see that the perception of the people still voices heavily on food, drinks and shopping and less on culture and events. Therefore it shows a need to use the Web as a measurement of the current state of the leisure city but as well to get more insights how to influence the focus of the perceptions of the crowd.

Nevertheless, the used method is still focussed on points of interests and therefore the result is a static interpretation of the question *'what can be done there?'*. The visualisation shows how the LDA model reasons from a certain point and connects this to a prediction of a topic (figure 5.11). The map showed a various amount of points on the map instead of areas of interest, as how humans frequently perceive parts of the city. Hobel et. al. (2015) have argued that an activity cannot be seen as a single point in space but involves a relation to other places. Therefore an (urban) area of interest is distinguished by its multi co-located features such as similarity, proximity or related activities.

The goal of this was to test a method that allowed enrichment for the leisure city of Zwolle by using web texts describing urban areas. It was an explorative research in how qualitative data could enhance the data-driven approach of the city of Zwolle, in order to pursue their ambition to be a fun, vital, accessible, liveable and an attractive city. Web-semantics have proven to be a potential source of enriching our knowledge of places, although in its current state somewhat biased and limited in use. Topic modelling is a valuable method to summarize the descriptions of the crowd in relation to leisure - which can be used in various ways and be combined with other data in city intelligence. As the city of Zwolle's ambition to be a *'place to be'* that offers multiple consumer spaces and different experiences on culture, events and the food industry as part of a leisure city - the results of the topic model shows that not all facets are characterizing the city on the Web. The results of this thesis show that the ambitions of the city as leisure city do not follow how visitors and citizens perceive and experience city. As it is the city's role to be a hospitable city - the use of web-semantics give insights and reflect on the perceptions that exist of the city on the Web.

## 6.2 Discussion

The potential of (big) data seems unprecedented. Kitchin (2013a) describes the opportunity of big data as a *"deluge of rich, detailed, timely and low-cost data - that can provide much more sophisticated, wider scale, finer grained understandings of societies and the world we live in"* (p. 263). It provides new ways of looking at massive amounts of unstructured data which were previously difficult to access. Nevertheless, scholars as Batty et. al. (2012), Adams & Janowicz (2014), Hobel et. al. (2015) acknowledge that grasping these unstructured data sets still causes challenges in translating human representations of a place into a computational representation of a space. The remainder of this chapter discusses what the results of the LDA models mean and how we can use topic modelling in the context of geographical analysis and urban policies.

### 6.2.1 Volume, variety and velocity

Current streams of data, or big data, are often characterized by three V's: *volume*, *variety* and the *velocity* of the data (Kitchin, 2013a, p. 262). As discussed in the previous paragraph, the training dataset implies a too small sample and too limited obtained information for detailed classes. The size and the topics of the training data have influenced the results of the (labelled) Latent Dirichlet Allocation topic models. When using a small number of documents, it substantially causes noise within the topics. Irregularities in the dataset, such as names of websites as [www.leuketips.nl](http://www.leuketips.nl) which showed up commonly, easily take over in topics. This kind of irregularities is very probable to be averaged out when a larger sample is used. As seen in the L-LDA topic model, activity classes with a bigger number of instances (*Eating* and *Watching*) resulted in more cohesive topics. Not only the count of instances is important. Despite being the second biggest activity class in the dataset, the class *Shopping* did not include many relevant words. It can be concluded that the quality of the website is important in LDA modelling. Does the website have sufficient text to parse? And does the website does not contain too many artefacts? In a small dataset, these irregularities are much more sensitive to modelling results.

Usually, it takes up to thousands of documents before the topics deriving from a topic model will be stabilized. This causes a scalability problem in which can be questioned whether it is feasible to encode a sufficient amount of documents. As the encoding was done manually, this leads to a very time-consuming process. Furthermore can be questioned if it is possible to encode that many different places for a middle-sized city as Zwolle is. The iterative way of encoding in this thesis already has led to point of saturation, implying that almost every possible activity within the inner city will be covered in the encoding.

The restrictions encountered during the data collection caused that some places could not be encoded. Two main assumptions influenced the website selection. First, the place should be described uniquely by the web page itself. This means that the selected web page can only describe one activity at a time otherwise it would cause difficulties by encoding the activity and referent of the place. The encoding scheme only could take one entry at a time. For places with multiple activities, a restriction of five activities was made. The second assumption relied on the identification of Open Street Map. By selecting uniquely described places by the website and within Open Street Map, it was not able to select some events that take place in different locations or to encode shops. Shopping was expected to be the biggest category because Zwolle was preliminary seen as a shopping city for years (Gemeente Zwolle, 2014a). However, most websites of shops do not describe a specific store but describe the brand or show an online webshop, which did not contain a sufficient amount of web text to scrape. Therefore the scalability problem works twofold: the encoding definitely miss some activities that describe the leisure city of Zwolle but within the assumptions, it is questioned how many more places can be encoded to improve the scale of the training dataset. It seemed that the point of saturation was reached in the way web pages were encoded. For a middle-sized city as Zwolle, it seems not suitable to encode eventually thousands of websites in the collection within the assumptions and stated restrictions of the data collection.

To conclude the discussion of the volume of the training data, it should be acknowledged that just adding *more* data does not automatically leads to more *relevant* information (figure 6.2). The figure shows the relation of sources and statistics: when some collected data is able to compute statistics with. To collect sufficient relevant information, data researches requires tremendous bigger datasets than for instance with survey data to be accurate and relevant. Big data is considered as a *dirty data* source, with a lot of noise. To control for that noise, a lot of data is needed. Hence why volume can be considered as a *consequence* instead of a characteristic of (big) data.

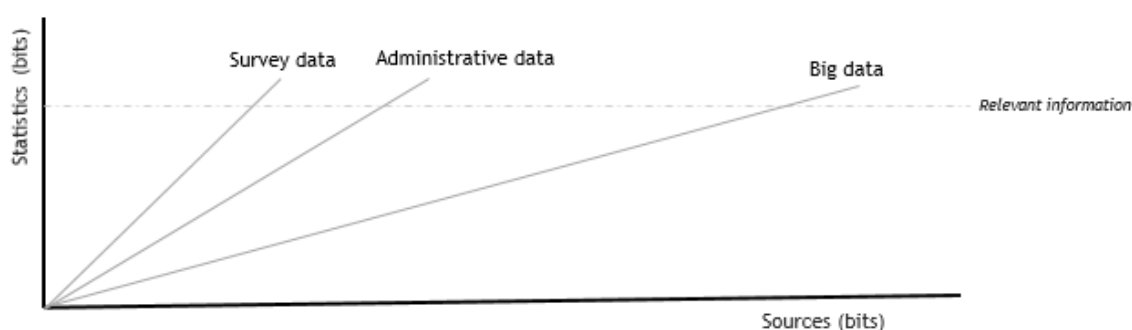


Figure 6.2 - Relevance of information resources in statistics

Obtained from Lena Tichem (CBS), personal communication, a presentation on October 11, 2017.

The noise of data is often a consequence of the *variety* of data within the dataset. Often different sorts of data are added to a 'pile' to extract information. Relevant information occurs from the right type of data. In the case is sought for a method to combine different web texts to conceptualize the leisure city. Therefore institutional web data was combined with different *user-generated contents* such as available on blogs, Facebook pages and TripAdvisor. Next to the improvement in the volume of the used data, also the variety of used data resources could be extended. There are various other

social media and resources available which contains text, such as Twitter, which forms a valuable source for destination information retrieval (Akehurst, 2009). And of course there are various other forms of information resources that could be relevant to obtain place information from, for example, sensor data. However, the possibilities and assumptions of the topic model cannot consider all kinds of data. As sensor data is quantified data, it cannot be scraped as texts to gain topics from. Akehurst (2009) questions the emerging amount of information available on the internet: the volume of information increases but which information is accurate, up-to-date and useable? By encoding manually, each website could be observed on its accuracy. For instance, only web texts that describe the current place affordance were encoded. The focus of this thesis was mainly on descriptive texts but a possible and interesting case would be to add more web data that consists of some text based on reviews such as on social media in order to focus on experiences of people of a certain place instead of what can be done where. Related to the aim of the municipality of Zwolle to use more *soft data* in data analysis, using more sources based on written text are useful in understanding dynamics in the city.

In other terms of variety, the multifunctionality of place could not easily be encoded and evaluated in this thesis. On different levels, the (inner) city functions as a multifunctional place. In the theoretical framework was the work of Burtenshaw discussed in figure 2.5, which showed that different activity profiles such as the *Nightlife City* overlaps with other profiles. Here was concluded that the city has overlapping resources and has a multifunctional use. On a smaller scale, within the latest years multifunctionality have become popular within shops, libraries and so on. For example, one of the “*problematic places*” within the encoded dataset was bookstore Waanders in de Broeren. Its main function is to sell books, but the store also sells gifts and you are able to drink a cup of coffee. Because the ontology *Urban Leisure* distinguishes activities very explicit, you cannot learn the machine to understand that one single place affords both buying books and drinking coffee. The *labelled* Latent Dirichlet Allocation, on the other hand, predicts the distribution of which activity class is the most important in one document. In this case, can acknowledge that a single document about a single place consists of multiple activities. As a fictional example, the website of bookstore Waanders in de Broeren could exist for 80% about shopping books and other gifts and for 20% about drinking coffee in the store. This means that by supervising the topic model, the result gives a better understanding of the perceived space of people instead of linking a topic to a certain point in space.

The last characteristic, or perhaps a consequence of big data, is the velocity of data. Encoding webpages, let them scrape by a Python script and evaluate the topics and the model by ten-fold cross-validation is not presumably fast - as big data consequently is considered as making the city plannable and predictable by minutes (Batty, 2013, p. 276). The strengths of using web semantics lie in the volume and variety of the data. Nevertheless, the results of topic modelling place affordance can be used in further city intelligence as well.

### 6.2.2 Topic modelling and city intelligence

The volume, velocity and variety of data drives city governments to work data-driven (Batty, 2013; Kitchin, 2013a, Marshall, 2012). The municipality of Zwolle argued that by using more (big) data its monitoring and intervention would do better (Vergeer & Capelleveen, 2017). Currently, a lot of statistics on the whereabouts in the city are gained from the Buurt-voor-Buurt survey. It is used to measure the viability of the inner city neighbourhoods in Zwolle. Referring to figure 6.2, which shows the relevance of information resources in statistics, indicates that gaining the same insights with (big) data rather than using a survey means a need for an increasingly bigger sample of data to avoid noise influencing the results. It is necessary to remain critical in what kind of data you want to use to answer the relevant questions. Big data asks for a new conceptualization of existing and new theories about how to use data in urban challenges. The emergence of data is unprecedented, nevertheless, it is important to acknowledge that data is not superior and still a sample (Batty, 2013; Kitchin, 2013a).

By restricting the training data it has caused a relatively small dataset which is in comparison to other big data resources quite static as it predominantly makes use of one source: web texts. The potential use of this LDA topic model in explaining the city as a fun, vital, accessible, liveable and attractive place is enriching existing knowledge by making use of the voice of the crowd online. The results of LDA describe Zwolle by the terms of the crowd in different topics. As place is deeply rooted in human perception (Scheider & Janowicz, 2014), LDA summarizes perceptions into different themes and connect them to a certain point in Zwolle. As municipality, it is valuable information to know how the city is currently described. Edwards et. al. (2010) acknowledges this by explaining that the urge for data by planners and policy-makers derives from a need for more detailed information about the whereabouts in the city.

The interactive visualisation made based on the results of LDA topic model 1 shows its potential to use this data to inform people about the leisure city of Zwolle. Wakamiya et. al. (2011) have argued that defining urban areas and giving information about the characteristics is a crucial for making (geographic) decisions as everything we do takes place in space. Information retrieval on the Web became one of the most important sources of deciding what to do in a certain space (Akehurst, 2009; Fabrikant & Battenfield, 2001; Költringer & Dickinger, 2015). Interactively mapping could inform people about the 'topics' and what to do in the city. As this information is based on web data - which becomes more based on *user-generated content* - it both summarizes the perceived space of the crowd who already have been there and could influence new people by informing them the existing assessments of leisure activities in space and time.

Hobel et. al. (2015) have argued that an activity cannot be seen as a single point in space in the perceived space of a person. It involves a cluster of related places to a certain activity based on similarity, proximity or related other activities. The visualisation shows indeed similar places based on the topics and as the data has been plotted on a map proximity can be assumed by human interpretation. However, the model and visualisation do not show *areas* based on the topics. Because the data collection and encoding was focussed on Open Street Map node identifiers, it immediately causes emphasis on places as a point. Mapping places in such way do not always match the perception of users, as people typically conceive places as a region and different users tend to associate places differently (Hobel et. al., 2015, p. 20). The resulting visualisation seems therefore quite static in representing place perceptions and affordance.

To make sense of the results, its improvement lies in adding *more* data sources to improve the overall quality of the model. The resulting prediction quality of LDA already lies around 70 to 80 percent - so it is probable that by adding more web data the prediction quality further improves. Next, to add more web data which is similar to the already used data, its potential further lies combining results of other data sources. The value of city intelligence lies in using different sources: combining different kinds of insights in explaining the city. But as it is very probable that within twenty years the city will be mainly understood by digital sensors and real-time data (Batty, 2013; Kitchin, 2014), it is important to show how other measurements enhance sensing measurements. Sensor data is focusing on quantitative data of the city. Sensing data in the city already occurs in various forms such as measuring crowd activities with smartphones, WiFi data, Bluetooth, aggregated camera images and counters. They measure crowd activity, walking routes and counting how many people are at a certain time at a certain place. The added value of topic modelling is understanding place perception of the crowd - something sensors cannot measure.

Combining research about the city, such as the potential of leisure-related activities in the inner city with crowd dynamics such as walking routes leads to valuable new information. Where are the people and why are they there? Crowd management data is often more used to redirect visitor flows in order to avoid overcrowding. Using web semantics could potentially further explain why people are visiting those specific places, for example by aggregating places into hot- and cold spots of certain activities in the city. This could be of valuable potential for further evidence-based policies. Combining different datasets, enables the city to reflect on how people engage with their urban

environment in respect to place affordance. The city will know what makes them attractive, and in order to stay that vital, it could use the insights to improve policies and public services. And then data, if representative and reliable enough, could form the core of explaining events and behaviour in the city.

### 6.3 Implications and recommendations

The potential of using (big) data in city intelligence almost seems undoubtful - however, the challenge remains how data can grasp and translate human representations of a place in a computational representation of space (Adams & Janowicz, 2014; Batty et. al., 2012; Hobel et. al., 2015). Therefore this paragraph discusses the implications involved in this research and gives recommendations for further use and research.

It is important to remain critical in data-analysis, as what is not seen in this data does not mean it is not there. The data-driven approach in this research heavily influences the methodology and can 'make or break' the results. Hereby can be implicated that using web-semantic still cause some problems that need to be solved in order to use web-semantic to its full potential. Campbell et. al. (2014) have stated that despite topic modelling is indeed useful to summarize shared opinions, a common pitfall is its construct validity, implying that the topics from LDA modelling are not per definition intuitive, human ideas or concepts. The main conclusion based on the LDA model about the leisure-city of Zwolle is that there is not much focus on culture and events. However, due to the limits in assumptions, it can also be explained that, for example, events cannot be taken into account in topic modelling as events are not always place-bound and identifiable with an OSM identifier. Therefore it seems that machine learning for now only can adapt straightforward classes and places while in human thinking there is a certain difference in place perception of as in the ontology between different cuisines and meals. Furthermore, the multi-labelling issue showed difficulties in how a place cannot be evaluated as a multifunctional place with different activities. The single-labelled LDA model reasoned from the most frequent activity class for each place while even more difficulty appeared in the multi-labelled LDA in cross-validating the results.

The ambition of Zwolle to use a data-driven approach to enhance their city as a vital and attractive *place to be* is very focused on *how* people perceive their city. In this case, there is certain a mismatch between the assumptions used in the topic model, such as the restriction to a certain location in OSM and a unique description of a place only, and the perceived space of people in the city. Therefore Zwolle's ambition to replace the Buurt-voor-Buurt survey with data-driven research was not met specifically.

Therefore is the first recommendation that the research agenda on web semantics should focus on overcoming these describing issues to get closer to intuitive human perception of place. Hereby is recommended, if possible, to use a bigger dataset with over a thousand documents to limit noise that possibly can take over. When a bigger dataset is used, potentially more classes are big enough to model instead of the four biggest classes that were used in the LDA model as presented in this thesis.

Another recommendation is in light of the growth towards Semantic Web. Using semantics and linked data definitely should enhance the model. To start, the used Urban Leisure ontology can be used to create links between affordance and place to learn models how places can be multifunctional places. The used ontology is never outright and complete but becomes by linking concepts more meaningful. In going towards a real semantic Web, using semantics could help solving the multi-labelling issue at least on a semantic level in machine learning.

Nevertheless can be questioned if only using a bigger dataset would improve the modelling results as this implies a scalability problem. It can be questioned if there are more than thousand different places with leisure-related activities in the inner city of Zwolle or another middle-sized city. To use topic modelling in data-driven use - which implicates effective and efficient policies and making decisions based on data - it is valuable to further look at the *frequency* how place descriptions of the

web. Next to the need of representing space as an *area of interest*, this could further help in hot- and cold spot analysis in cities. As people frequently see the same information, it is very probable they would visit the related place instead of going elsewhere (Költringer & Dickinger, 2015). Another valuable recommendation is to take reviews into account in encoding web pages, to not only measure behaviour but also the expected experience. In relation to the concepts of perceived and conceived space - using reviews can help getting a more valuable overview of the experiences of the inner city. The interactive visualisation has shown its potential that much more data can be added and be retrieved to shape a representation of the city created by data.

The used method has restricted the use of various data sources which are valuable in data-driven approach in cities. The societal relevance of this research was to enhance the ambition of data-driven urban policy in cities. Despite the challenges web data brings, it showed potential in how city intelligence not only should rely on quantified data such as sensor data as frequently is done in practice. Considering web data as a qualitative data source shapes the city and its intelligence by citizen created information. It emphasizes the importance of people in data analysis and creates a bottom-up created layer of data on the city. In its full potential, when all describes issues can be overcome, web-semantic research could describe the city by capturing concepts as emotions, values, opinions and the way people interact and make sense of our cities.



## References

- Adams, B. & Janowicz, K. (2014). Thematic Signatures for Cleansing and Enriching Place-Related Linked Data. *International Journal of Geographical Information Science*, 29(4), pp. 556 - 579.
- Adèr, H.J. (2008). Chapter 14: Phases and initial steps in data analysis. In H.J. Adèr & G.J. Mellenbergh (eds.). *Advising on Research Methods: A consultant's companion* (pp. 333 - 356). Huizen: Johannes van Kessel Publishing.
- Aguiar, M. & Hurst, E. (2007). Measuring Trends in Leisure: The Allocation of Time Over Five Decades. *The Quarterly Journal of Economics*, 122(3), pp. 969 - 1006.
- Akehurst, G. (2009). User Generated Content: the use of blogs for tourism organisations and tourism consumers. *Service Business*, 3(51), pp. 51 - 61.
- Alazzawi, A.N., Abdelmoty A.I. & Jones, C.B. (2012). What can I do there? Towards the automatic discovery of place-related services and activities. *International Journal of Geographic Information Science*, 26(2), pp. 345 - 364.
- Ashworth, G. & Page, S.J. (2011). Urban tourism research: Recent progress and current paradoxes. *Tourism Management*, 32, pp. 1 - 15.
- Bansal, S. (2016; August 24). Beginners Guide to Topic Modeling in Python [Article]. Retrieved from: <https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-python/>.
- Batty, M. (2013). Big data, smart cities and city planning. *Dialogues in Human Geography*, 3(3), pp. 274 - 279.
- Batty, M., Axhausen K.W., Giannotti, F., Pozdnoukov, A., Bazzani, A., Wachowicz, M., Ouzonis, G. & Portugali, Y. (2012). Smart Cities of the Future. *The European Physical Journal*. Special Topics 214, pp. 481 - 518.
- Bekel, P. (2008, 11 April). De betekenis van Web 3.0 en het semantic web. [Article]. Retrieved from: <https://www.frankwatching.com/archive/2008/04/11/de-betekenis-van-web-30-en-het-semantic-web/>.
- Blei, D.M., Ng, A.Y., Jordan, M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(2003), pp. 993 - 1022.
- Blei (2012). Probabilistic Topic Models. *Communications of the ACM*, 55(4), pp. 77 - 84.
- Boersma, O. & A. Raatgever (2017; February 9). Nieuwe data voor de nieuwe binnenstad. [Blog]. Retrieved from: <http://www.platform31.nl/blogs/blogs-platform31/nieuwe-data-voor-de-nieuwe-binnenstad>.
- Brownlee, J. (2016; March 21). Overfitting and Underfitting With Machine Learning Algorithms [Article]. Retrieved from: <https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/>.
- Campbell, J.C., Hindle, A. & Stroulia, E. (2014). Latent Dirichlet Allocation: Extracting Topics from Software Engineering Data. In: Bird, C., Menzies, T., Zimmerman, T. (eds.) *The Art and Science of Analyzing Software Data* (pp. 140 - 157). Amsterdam: Morgan Kaufmann.
- Cambridge Semantics (n.d.). Introduction to the Semantic Web. Retrieved from: <http://www.cambridgesemantics.com/semantic-university/introduction-semantic-web>.



- Centraal Bureau voor de Statistiek, CBS (2017a; 29 June). Factsheet Urban Data Center Zwolle [Factsheet]. Retrieved from: <https://www.cbs.nl/nl-nl/achtergrond/2017/26/factsheet-udc-zwolle>.
- Centraal Bureau voor de Statistiek, CBS (2017b; 29 June). Zwolle blijft relatief jong [News Article]. Retrieved from: <https://www.cbs.nl/nl-nl/nieuws/2017/26/zwolle-blijft-relatief-jong>.
- Crouch, D. (2016). Geographies of Leisure. In: Rojek, C., S.M. Shaw & A.J. Veal (eds.). *A Handbook of Leisure Studies* (pp. 125-139). New York: Palgrave Macmillan.
- Crosbie, T. (2006). Using Activity Diaries: Some Methodological Lessons. *Journal of Research Practice*, 2(1), pp. 1 - 13.
- Cormode, G. & Krishnamurthy, B. (2008). Key differences between Web 1.0 and Web 2.0. *Peer-reviewed Journal of the Internet*, 13(6). Retrieved from: <http://www.ojphi.org/ojs/index.php/fm/article/view/2125/1972>.
- Delen, D., & Demirkan, H. (2013). Data, information and analytics as services. *Decision Support Systems*, 55(1), pp. 359- 363.
- Dempsey Morais, C. (2012, October 28). Where is the Phrase “80% of Data is Geographic” From? [Article]. Retrieved from: <https://www.gislounge.com/80-percent-data-is-geographic/>.
- Dhar, V. (2013). Data Science and Prediction. *Communications of the ACM*, 56(12), pp. 65 - 73.
- Dijst, M. (2009). Time geographical analysis. In: R. Kitchin and Thrift, N. (eds.). *International Encyclopedia of Human Geography* (pp. 266 - 287).
- Domingos, P. (2012, n.d.). A Few Usefule Things to Know about Machine Learning. *Communications of the ACM*, 55(10), October 2012.
- Edwards, D., Dickson, T., Griffin, T. & Hayllar, B. (2010). Tracking the Urban Visitor: Methods for Examing Tourist’ Spatial Behaviour and Visual Representations. In: Richards, G. & Munsters, W. (eds.). *Cultural Tourism Research Methods* (pp. 104-114) . Oxfordshire: CABI.
- Ellis, P. (2017; January 5). Cross-validation of topic modelling [Article]. Retrieved from: <http://ellis.github.io/blog/2017/01/05/topic-model-cv>.
- Evans, J. & Jones, P. (2011). The walking interview: Methodology, mobility and place. *Applied Geography*, 31, pp. 849 - 858.
- Evers, D., Tennekes, J. & Van Dongen, F. (2015). *De veerkrachtige binnenstad*. Den Haag: Planbureau voor de Leefomgeving.
- Fabrikant, S.I., Buttenfield, B.P. (2001). Formalizing Semantic Spaces for Information Access. *Annals of the Association of American Geographers*, 91(2), p. 263 - 280.
- Farrahi, K. & Gatica-Perez, D. (2011). Discovering routines from large-scale human locations using probabilistic topic models. *ACM Transactions on Intelligent Systems and Technology*, 2(1), pp. 3:1 - 3:27.
- Gao, S., Li, L., Li, W., Janowicz, K. & Zhang, Y. (2017). Constructing gazetteers from volunteered Big Geo-Data based on Hadoop. *Computers, Environment and Urban Systems*, 61(2017), pp. 172 - 186.
- Gangemi, A. (2013). A Comparison of Knowledge Extraction Tools for the Semantic Web. In: Cimiano, P., Corcho, Ó., Presutti, V. Hollink, L. & Rudolph, S. (eds). *The Semantic Web: Semantics and Big Data*, volume 7882 of *Lecture Notes in Computer Science*, pp. 351 - 366.

- Gemeente Zwolle (2014a). Coalitieakkoord 2014 - 2018. Agenda voor een levendige stad. Retrieved from: [https://www.zwolle.nl/sites/default/files/coalitieakkoord\\_2014-2018.pdf](https://www.zwolle.nl/sites/default/files/coalitieakkoord_2014-2018.pdf).
- Gemeente Zwolle (2014b). Zwolle bij de hand 2014. Kerncijfers. Retrieved from: [https://www.zwolle.nl/sites/default/files/zwolle\\_bij\\_de\\_hand\\_2014.pdf](https://www.zwolle.nl/sites/default/files/zwolle_bij_de_hand_2014.pdf).
- Gemeente Zwolle (2016). Buurt-voor-Buurt Onderzoek 2016. Retrieved from: <https://www.zwolle.nl/actueel/cijfers-kaarten-en-onderzoeken/onderzoeksdatabank/buurt-voor-buurt-onderzoek-2016>.
- Gemeente Zwolle (2017). Zwolle Bruist. De Strategische Agenda Binnenstad 2017 - 2022. Retrieved from: <https://www.zwolle.nl/sites/default/files/strategische-agenda-binnenstad-2017-samengevat.pdf>.
- Gemeente Zwolle (n.d.). Ontwikkelingsprogramma Binnenstad 2020. Retrieved from <https://www.zwolle.nl/binnenstadsprogramma>.
- Giri, K. (2011). Role of Ontology in Semantic Web. *Journal of Library & Information Technology*, 31(2), pp. 116 - 120.
- Gruber, T.R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2), pp. 199 - 220.
- Goodchild, M.F. (2007a). Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. *International Journal of Spatial Data Infrastructure Research*, 2007(2), pp. 24 - 32.
- Goodchild, M.F. (2007b). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69, pp. 211 - 221.
- Gregory, D., Jonston, R., Pratt, G., Watts, M.J. & Whatmore, S. (2009). *The Dictionary of Human Geography*. 5<sup>th</sup> edition. West-Sussex: Wiley - Blackwell. Conducted definition: *leisure*.
- Guo, Y., Barnes, S.J. & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, 59 (April, 2017), pp. 467 - 483.
- Hägerstrand, T. (1970). What about people in regional science? *Papers of the Regional Science Association*, 24, pp. 7 - 21.
- Hall, M.C., & Page, S.J. (2014). *The Geography of Tourism and Recreation*. Abington, Oxon: Taylor & Francis Ltd.
- Hassan, S. & Ukkusuri, S. (2014). Urban activity pattern classification using topic models from online geo-location data. *Transportation Research Part C*, 44 (2014), pp. 363 - 381.
- Heywood, I., Cornelius, S. & Carver, S. (2011). *An Introduction to Geographical Information Systems*. Essex: Pearson Education Limited.
- Hobel, H., Abdalla, A., Fogliaroni, P. & Frank, A.U. (2015). A Semantic Region Growing Algorithm: Extraction of Urban Settings. In: F. Bação et. al. (eds.), *AGILE 2015, Lecture Notes in Geoinformation and Cartography* (pp. 19 - 33). Springer International Publishing.
- Hollenstein, L. & Purves, R.S. (2010). Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science*, 1(2010), p. 21 - 48.
- Hu, Y., Gao, S., Janowicz, K., Yu, B. Li, W., & Prasad, S. (2015). Extracting and understanding urban areas of interest using geotagged photos. *Computers, Environment and Urban Systems*, 54, pp. 240 - 254.

- Huang, W. & Li, S. (2016). Understanding human activity patterns based on space-time semantics. *ISPRS Journal of Photogrammetry and Remote Sensing*, 121 (2016), pp. 1 - 10.
- Hulstaert, L. (2017; October 19). Data Camp Tutorial: LDA2vec: Word Embeddings in Topic Models [Tutorial]. Retrieved from: <https://www.datacamp.com/community/tutorials/lda2vec-topic-model>.
- iBestuur (2016; 24 november). Onderweg naar Smart Zwolle. *iBestuur*, 20, pp. 26 - 27.
- Jacobi, C., Van Atteveldt, W., & Welbers, K. (2015). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 4-2016, pp. 89 - 106.
- Jayne, M. (2006). *Cities and consumption*. London: Taylor & Francis Ltd.
- Johnson, A.J. & Glover, T.D. (2013). Understanding Urban Public Space in a Leisure Context. *Leisure Sciences*, 35(2), pp. 190 - 197.
- Jordan, T., Raubal, M., Gartrell, B., Egenhofer, M.J. (1998). An Affordance-Based Model of Place in GIS. *Conference Proceedings*. Retrieved from: [http://www.raubal.ethz.ch/Publications/RefConferences/2894\\_sdh98\\_Place.pdf](http://www.raubal.ethz.ch/Publications/RefConferences/2894_sdh98_Place.pdf).
- Jones, C.B., & Purves, R.S. (2008). Geographical information retrieval. *International Journal of Geographical Information Science*, 22(3), pp. 219 - 228.
- Jones, C.B., Purves, R.S., Clough, P.D., & Joho, H. (2008). Modelling vague places with knowledge from the Web. *International Journal of Geographic Information Science*, 22(10), pp. 1045 - 1065.
- Kemperman, A.D.A.M., Borgers, A.W.J., Timmermans, H.J.P. (2009). Tourist shopping behaviour in a historic downtown area. *Tourism Management*, 30, pp. 208 - 218.
- Kling, F., Pozdnoukhov, A. (2015). When a city tells a story: urban topic analysis. *Proceedings of the 20<sup>th</sup> International Conference on Advances in Geographic Information Systems*, pp. 482 - 485.
- Klintberg, A. (2017; May 22). Explaining precision and recall [Website]. Retrieved from: <https://medium.com/@klintcho/explaining-precision-and-recall-c770eb9c69e9>.
- Kitchin, R. & Dodge, M. (2011). *Code/space: software and everyday life*. Cambridge, Massachusetts, MIT Press.
- Kitchin, R. (2013a). Big data and human geography: opportunities, challenges and risks. *Dialogues in Human Geography*, 3(3), pp. 262 - 267.
- Kitchin, R. (2013b). The real-time city? Big data and smart urbanism. *GeoJournal*, 2014(79), pp. 1 - 14.
- Kitchin, R. (2014). *The Data Revolution: big data, open data, data infrastructures & their consequences*. Los Angeles: SAGE.
- Kang, E., Kim, H. & Cho, J. (2006). Personalization Method for Tourist Point of Interest (POI) Recommendation. In: Gabrys, B., R.J. Howlett & L.C. Jain (eds). *KES 2006, Part I, LNAI* (pp. 394 - 400). Heidelberg: Springer-Verlag Berlin.
- Költringer, C. & Dickinger, A. (2015). Analyzing destination branding and image from online sources: A web content mining approach. *Journal of Business Research*, 68(9), pp. 1836 - 1843.
- Kwan, M.P. (2002). Time, information technologies, and the geographies of everyday life. *Urban Geography*, 23(5), pp. 471 - 482.
- Lefebvre, H. (1991). *The production of space*. Malden, MA: Blackwell.

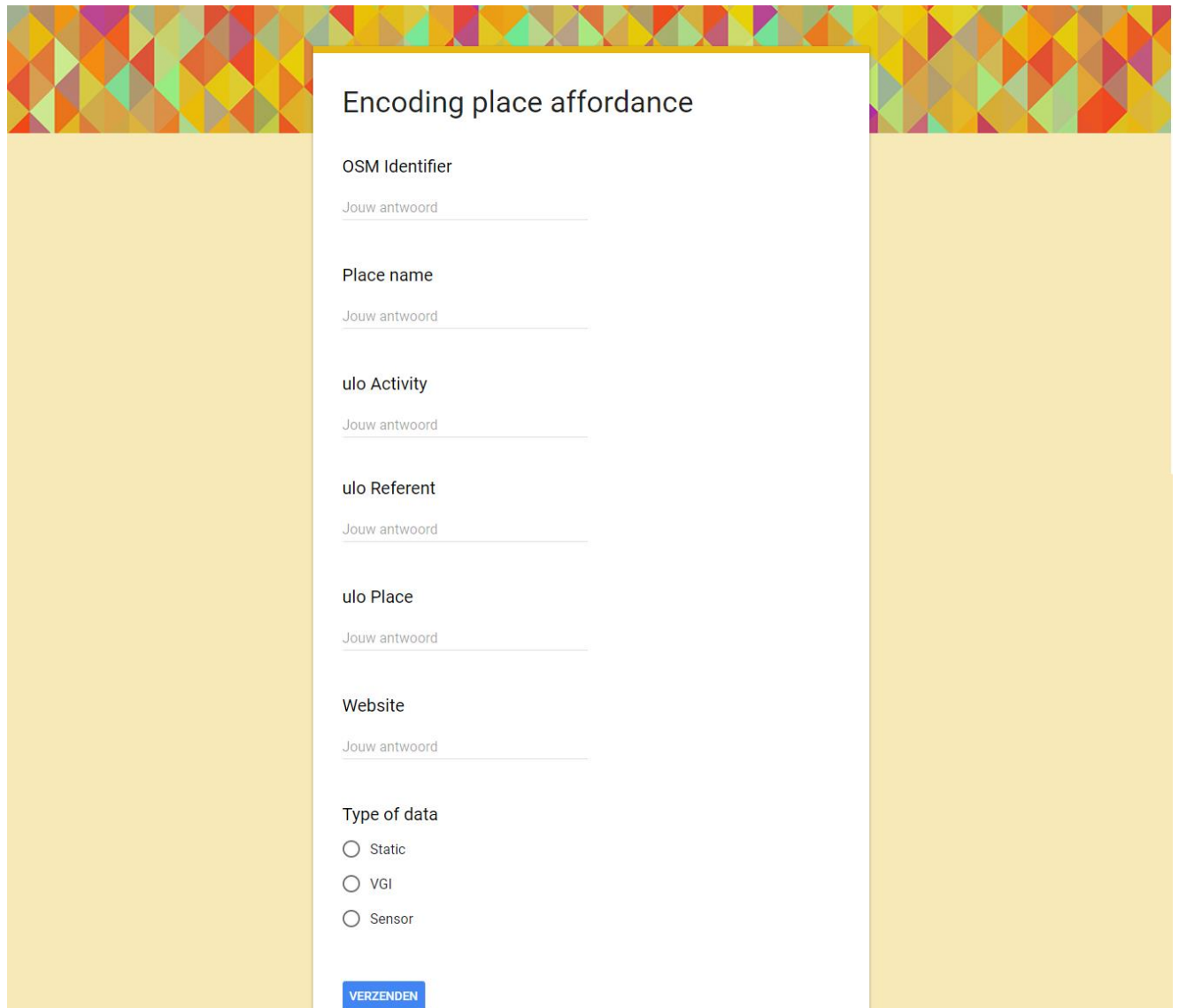
- Lorentzen, A. (2009). Cities in the Experience Economy. *European Planning Studies*, 17(6), pp. 829 - 845.
- Magasic, M. (2016). The 'Selfie Gaze' and 'Social Media Pilgrimage': Two Frames for Conceptualising the Experience of Social Media Using Tourists. In: Inversini, A. & R. Schegg (red.) (2016). *Information and Communication Technologies in Tourism 2016*. Springer International Publishing Switzerland, pp. 173 - 182.
- Marshall, C. (2012). Big Data, the crowd and me. *Information Services and Use*, 32(3-4), pp. 215 - 226.
- McLean, D.D. & Hurd, A.R. (2014). Recreation and Leisure: An Introduction. In: McLean, D.D. & Hurd, A.R. (eds.). *Kraus' Recreation and Leisure in Modern Society* (pp. 1 - 10). Burlington, MA: Jones and Bartlett Learning.
- Meijer, A. & Thaens, M. (2016). Urban Technological Innovation: Developing and Testing a Sociotechnical Framework for Studying Smart City Projects. *Urban Affairs Review*, first published date: September 30-2016.
- Montello, D.R., Goodchild, M.F., Gottsegen, J., Fohl, P. (2003). Where's Downtown?: Behaviour Methods for Determining Referents of Vague Spatial Queries. *Spatial Cognition and Computation*, 3(2&3), pp. 185 - 204.
- Nair, G. (2016, July). Text Mining 101: Topic Modelling [Tutorial]. Retrieved from: <http://www.kdnuggets.com/2016/07/text-mining-101-topic-modeling.html>.
- Nouwens, H. (2016; 4 October). Smart City trends in Nederland 2016 [Article]. Retrieved from: <http://slimstebinnenstad.nl/?p=1511>.
- Noy, N.F., & McGuinness, D.L. (2001). Ontology Development 101: A Guide to Creating Your First Ontology. Retrieved from: <https://protegewiki.stanford.edu/wiki/Ontology101>.
- Oh, H., Fiore, A.M. & Jeoung, M. (2007). Measuring Experience Economy Concepts: Tourism Applications. *Journal of Travel Research*, 46(2), pp. 119 - 132.
- Open Street Map (n.d.). [www.openstreetmap.org](http://www.openstreetmap.org).
- Pine, B.J. & Gilmore, J.H. (1998). Welcome to the Experience Economy. *Harvard Business Review*, 76(4), pp. 97 - 105.
- PromptCloud (2017, June 8). The Limitations of Web Scraping Tools [Presentation]. Retrieved from: <https://www.slideshare.net/promptcloud/limitations-of-web-scraping-tools>.
- Raatgever, A. (2017). Programma Vitale Binnensteden 2017, Platform 31. Retrieved from: <http://www.platform31.nl/wat-we-doen/kennisdossiers/vitale-binnensteden/programma-vitale-binnensteden-2017>.
- Ramage, D., Hall, D., Nallapati, R. & Manning, C.D. (2009). Labelled LDA: A supervised topic model for credit attribution in multi-labelled corpora. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 248 - 256.
- Roberts, K. (2006). *Leisure in contemporary society*. Oxfordshire: CABI.
- Robinson, J.P., Converse, P.E., Szalai, A. (1972). Everyday life in twelve countries. In: Szalai . A. (eds.). *The Use of Time* (pp. 113 - 144). The Hague - Paris: Mouton.
- Richardson, L. (2017, May 7). Python BeautifulSoup Package 44.6.0 [Script]. Retrieved from: <https://pypi.python.org/pypi/beautifulsoup4>.

- Rudinac, S., Zahálka, J., Worring, M. (2017). Discovering Geographic Regions in the City Using Social Multimedia and Open Data. In: Amsaleg, L. et. al. (eds.). *MMM 2017, Part II. LNCS 10133*, pp. 148 - 159.
- Scikit-Learn (2017a). User Guide Python Module [Guide]. Retrieved from: [http://scikit-learn.org/stable/user\\_guide.html](http://scikit-learn.org/stable/user_guide.html).
- Scikit-Learn (2017b). Cross-validation: evaluating estimator performance [Guide]. Retrieved from: [http://scikit-learn.org/stable/modules/cross\\_validation.html](http://scikit-learn.org/stable/modules/cross_validation.html).
- Scikit-Learn (2017c). Model Evaluation - quantifying the quality of predictions [Guide]. Retrieved from: [http://scikit-learn.org/stable/modules/model\\_evaluation.html#classification-metrics](http://scikit-learn.org/stable/modules/model_evaluation.html#classification-metrics).
- Scikit-Learn (2017d). Precision-Recall [Guide]. Retrieved from: [http://scikit-learn.org/stable/auto\\_examples/model\\_selection/plot\\_precision\\_recall.html](http://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html).
- Scikit-Learn (2017e). Decision Trees [Article]. Retrieved from: <http://scikit-learn.org/stable/modules/tree.html>.
- Scheider, S. (2017). Place LDA. Retrieved from: <https://github.com/simonscheider/PlaceLDA>.
- Scheider, S. & Janowicz, K. (2014). Place reference systems. *Applied Ontology*, 9(2014), pp. 97 - 127.
- Shim, C. & Santos, C.A. (2016). Urban Tourism: Placelessness and Placeness in Shopping Complexes. *Tourism Travel and Research Association: Advancing Tourism Research Globally*. 78.
- Slaghuis, L. (2009, 18 February). Semantic Web - Hoe werkt het nou echt? [Article]. Retrieved from: <https://www.frankwatching.com/archive/2009/02/18/semantic-web-hoe-werkt-het-nou-echt/>.
- Spierings, B. (2009). Producing Urban (Dis)similarity: Entrepreneurial Governance, Consumer Mobility and Competitive Consumption Spaces: The Case of the Enschede Region. In: Arts, B., Lagendijk, A., Houtum, H. (eds). *The Disoriented State: Shifts in Governmentality, Territoriality and Governance*. Environment & Policy, volume 39. Dordrecht: Springer.
- Szalai, A. (1972). *The Use of Time*. The Hague - Paris: Mouton.
- Ting, K.M., (2010). Precision and Recall. *Encyclopedia of Machine Learning*, p. 781. Retrieved from: [https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8\\_652](https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8_652).
- Thibault, A. & Lavigne, M.A. (2014). The “leisure city”: what are we talking about? *World Leisure Journal*, 56(1), pp. 2 - 5.
- Townsend, A.M. (2013). *Smart Cities: big data, civic hackers and the quest for a new utopia*. New York: W.W. Norton & Company.
- Tuan, Y-F. (1977). *Space and place: the perspective of experience*. University of Minnesota Press.
- Vaid, S., Jones, C.B., Joho, H. & Sanderson, M. (2005). Spatio-textual indexing for geographical search on the web. *Advances in Spatial and Temporal Databases*, Volume 3633 of Lecture Notes in Computer Science, pp. 218 - 235.
- Van der Drift, S. (2015). *Revealing spatial and temporal patterns from Flickr photography. A case study with tourists in Amsterdam*. Master thesis, Wageningen University.
- Van Oortmarssen, A., De Vries, M. & Van Loenen, B. (2014). *Privacy op zijn plaats*. Amersfoort: Geonovum.

- Van Duppen, J. & Spierings, B. (2013). Retracing trajectories: the embodied experience of cycling, urban sensescapes and the commute between 'neighbourhood' and 'city' in Utrecht, NL. *Journal of Transport Geography*, 30, pp. 234 - 243.
- Vergeer, M. & van Capelleveen, P. (2017; 17 May). *City Intelligence in de gemeente Zwolle*. Overheid 360 Congres, Utrecht. Retrieved through intranet - not publicly available.
- Vrolijk, D. (2017; April 12). Binnenstad: ken uw gasten en uzelf. [Blog]. Retrieved from: <http://www.platform31.nl/nieuws/binnenstad-ken-uw-gasten-en-uzelf>.
- Wakamiya, S., Lee, R. & Sumiya, K. (2011). Urban Area Characterization Based on Semantics of Crowd Activities in Twitter. In: C. Claramunt, S. Levashkin & M. Bertolotto (eds.) *GeoSpatial Semantics - Proceedings of the 4<sup>th</sup> International Conference GeoS 2011* (pp. 108 - 123). Berlin Heidelberg: Springer-Verlag.
- Weingart, S. (2011). Topic Modelling and Network Analysis [Tutorial]. Retrieved from: <http://www.scottbot.net/HIAL/index.html@p=221.html>.
- Zwolle Buurtmonitor (2015). Cijfers over Zwolle [Monitor]. Retrieved from: <https:// zwolle.buurtmonitor.nl/>.
- Zwolle Tourist Info (2017). Ontdek Zwolle [Tourist Webpage]. Retrieved from: <http://www.zwolletouristinfo.nl/>.



## Appendix I - Encoding Sheet

A form titled "Encoding place affordance" with a decorative geometric pattern at the top. The form contains several input fields and a radio button group, all set against a light yellow background.

**Encoding place affordance**

**OSM Identifier**  
Jouw antwoord \_\_\_\_\_

**Place name**  
Jouw antwoord \_\_\_\_\_

**ulo Activity**  
Jouw antwoord \_\_\_\_\_

**ulo Referent**  
Jouw antwoord \_\_\_\_\_

**ulo Place**  
Jouw antwoord \_\_\_\_\_

**Website**  
Jouw antwoord \_\_\_\_\_

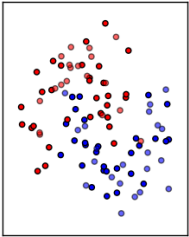
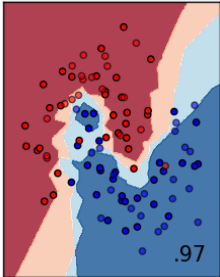
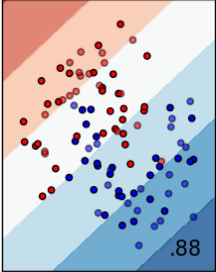
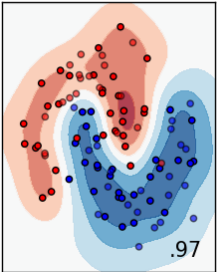
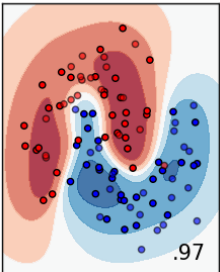
**Type of data**

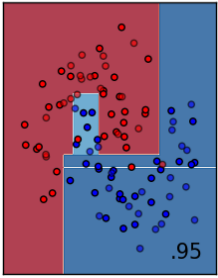
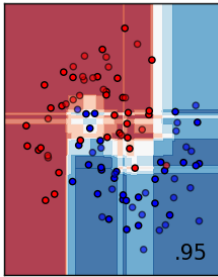
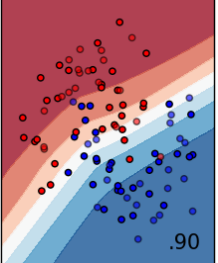
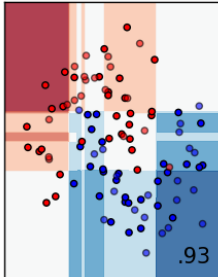
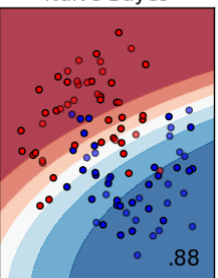
- Static
- VGI
- Sensor

**VERZENDEN**



## Appendix II - Machine Learning Classifiers

	<p>Input data</p> 	
<p><b>Logistic regression</b></p>		<p>Logistic regression estimates the probability of the observation's being part of one of the classes</p>
<p><b>Nearest Neighbours</b></p>	<p>Nearest Neighbors</p> 	<p>The principle is to find a training sample in its closest distance to a new point and predict the label from these to spatially categorize the points.</p>
<p><b>Linear SVM</b></p>	<p>Linear SVM</p> 	<p>Uses a subset of training points in the decision function (support vectors)</p>
<p><b>RBF SVM</b></p>	<p>RBF SVM</p> 	<p>Uses a subset of training points in the decision function (support vectors). This</p>
<p><b>Gaussian Process</b></p>	<p>Gaussian Process</p> 	<p>The prediction interpolates the observations based on a probabilistic prediction (Gaussian).</p>

<p><b>Decision Tree</b></p>	<p>Decision Tree</p> 	<p>It predicts the value of a target variable by learning simple decision rules inferred from the data features. The deeper the tree, the more complex the decision rules and the fitter the model. However, this classifier can be over-fit and unstable by small variations in data.</p>
<p><b>Random Forest</b></p>	<p>Random Forest</p> 	<p>An estimator that fits several decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.</p>
<p><b>Neural Net</b></p>	<p>Neural Net</p> 	<p>Attempts to mimic the learning pattern of natural biological neural network.</p>
<p><b>AdaBoost</b></p>	<p>AdaBoost</p> 	<p>A meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted</p>
<p><b>Naïve Bayes</b></p>	<p>Naïve Bayes</p> 	<p>A method based on Bayes' theorem with the 'naïve' assumption of independence between every pair of features. The assumptions are over-simplified. Known as a decent classifier but a bad estimator.</p>

Retrieved from Scikit Learn's user guide

Obtained from <http://scikit-learn.org/stable/documentation.html> on 2 August 2017

## Appendix III - Prediction quality of parameters and naïve classifiers for LDA

	Accuracy	Standard deviation	Weighted precision	Weighted recall	F-measure
<i>Model 1</i>	0.518	0.043	0.270	0.518	0.170
<i>Model 1 allact</i>	0.249	0.092	0.066	0.241	0.084
<i>Model 1 allactallclass</i>	0.208	0.123	0.059	0.208	0.077
<i>Model 1 allclass</i>	0.480	0.094	0.239	0.480	0.137
<i>Model 1 wouttypes</i>	0.518	0.043	0.270	0.518	0.170

Table A1 - The prediction quality of each ran model based on web texts

CV naïve classifier	
Accuracy	0.518
Standard deviation	+/- 0.043
Weighted precision	0.270
Weighted recall	0.518
F-measure	0.170

	precision	recall	f1-score	support
Shopping	0.00	0.00	0.00	33
Eating	0.51	1.00	0.68	70
Watching	0.00	0.00	0.00	15
Drinking	0.00	0.00	0.00	18
Avg/total	0.26	0.51	0.35	136

Table A2 - Cross-validated results and fit for the naïve classifier of model 1

CV naïve classifier	
Accuracy	0.671
Standard deviation	+/- 0.014
Weighted precision	0.451
Weighted recall	0.671
F-measure	0.268

	precision	recall	f1-score	support
ulo:Eating	0.67	1.00	0.80	41
ulo:Shopping	0.00	0.00	0.00	10
ulo:Watching	0.00	0.00	0.00	10
Avg/total	0.45	0.67	0.54	61

Table A3 - Cross-validated results and fit for the naïve classifier of model 2

## Appendix IV - Classification Results LDA model 1

Overview of cross-validated results of all ten machine learning classifiers

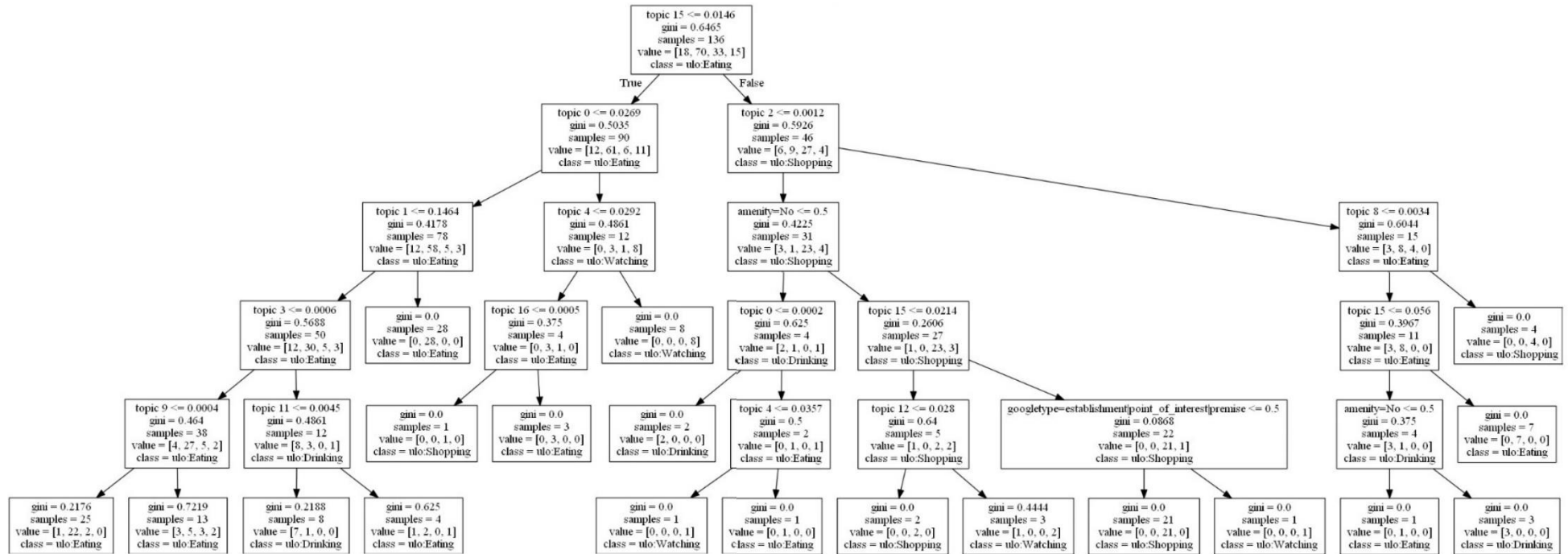
	<b>Logistic Regression</b>	<b>Nearest Neighbours</b>	<b>Linear SVM</b>	<b>RFB SVM</b>	<b>Gaussian Process</b>
<b>Accuracy</b>	0.609	0.582	0.518	0.559	0.602
<b>Standard deviation</b>	+/- 0.126	+/- 0.128	+/- 0.042	+/- 0.080	+/- 0.103
<b>Weighted precision</b>	0.638	0.605	0.270	0.507	0.455
<b>Weighted recall</b>	0.609	0.582	0.518	0.559	0.602
<b>F-measure</b>	0.491	0.477	0.170	0.393	0.315
<b>Improvement compared to naïve classifier?</b>	Yes	Yes	No	Yes	Yes
	<b>Decision Tree</b>	<b>Random Forest</b>	<b>Neural Net</b>	<b>AdaBoost</b>	<b>Naïve Bayes</b>
<b>Accuracy</b>	0.575	0.542	0.685	0.555	0.486
<b>Standard deviation</b>	+/- 0.071	+/- 0.098	+/- 0.130	+/- 0.109	+/- 0.160
<b>Weighted precision</b>	0.554	0.384	0.625	0.577	0.620
<b>Weighted recall</b>	0.597	0.569	0.663	0.555	0.486
<b>F-measure</b>	0.452	0.208	0.537	0.482	0.468
<b>Improvement compared to naïve classifier?</b>	Yes	Yes	Yes	Yes	No

## Appendix V - Classification Results LDA model 2

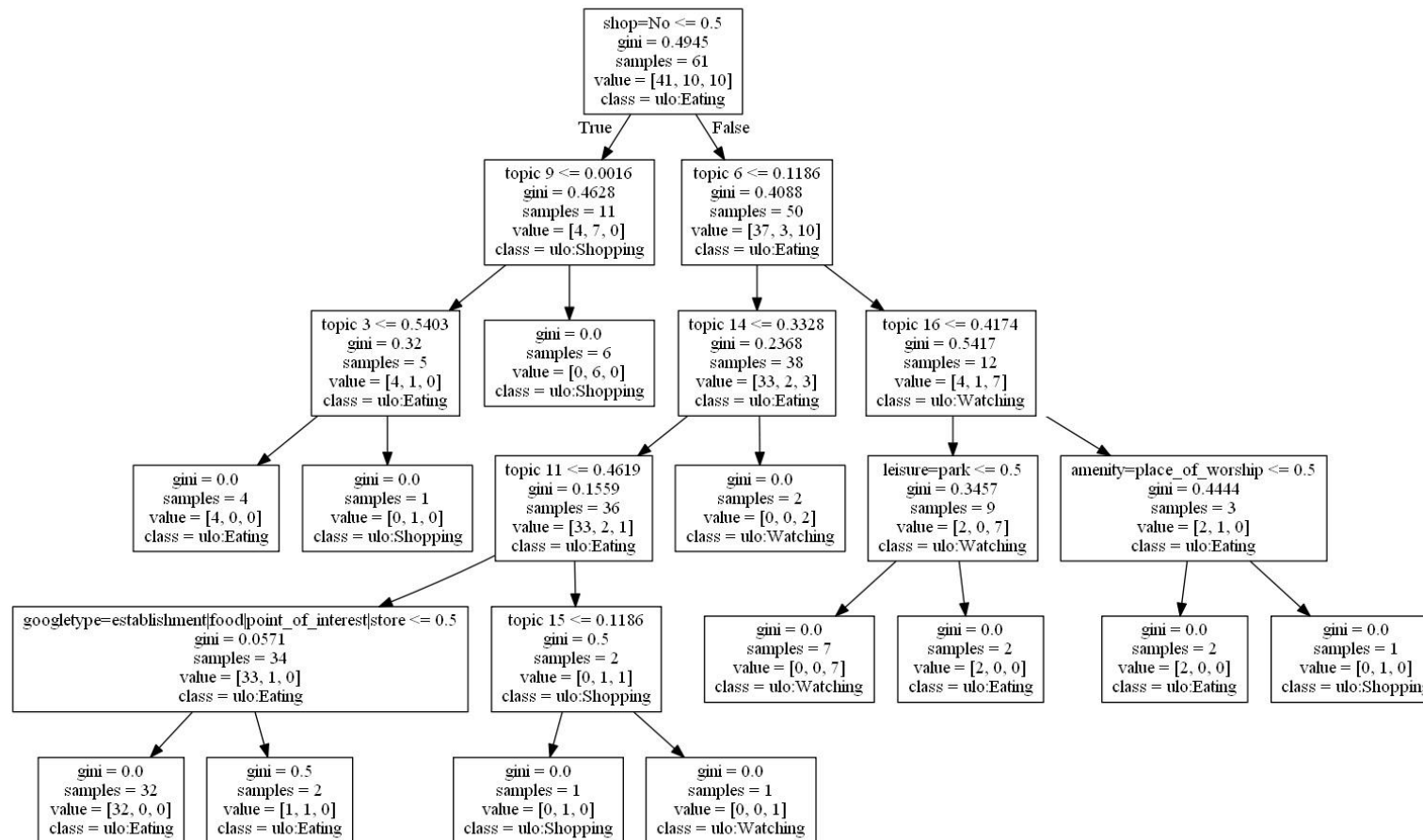
Overview of cross-validated results of all ten machine learning classifiers

	<b>Logistic Regression</b>	<b>Nearest Neighbours</b>	<b>Linear SVM</b>	<b>RBF SVM</b>	<b>Gaussian Process</b>
<b>Accuracy</b>	0.717	0.767	0.671	0.671	0.671
<b>Standard deviation</b>	+/- 0.167	+/- 0.133	+/- 0.014	+/- 0.014	+/- 0.014
<b>Weighted precision</b>	0.679	0.699	0.451	0.451	0.451
<b>Weighted recall</b>	0.717	0.766	0.671	0.671	0.671
<b>F-measure</b>	0.537	0.593	0.268	0.268	0.268
<b>Improvement compared to naïve classifier?</b>	Yes	Yes	No	No	No
	<b>Decision Tree</b>	<b>Random Forest</b>	<b>Neural Net</b>	<b>AdaBoost</b>	<b>Naïve Bayes</b>
<b>Accuracy</b>	0.638	0.688	0.783	0.652	0.624
<b>Standard deviation</b>	+/- 0.586	+/- 0.050	+/- 0.167	+/- 0.120	+/- 0.221
<b>Weighted precision</b>	0.586	0.451	0.66	0.553	0.640
<b>Weighted recall</b>	0.655	0.688	0.767	0.652	0.624
<b>F-measure</b>	0.655	0.300	0.560	0.397	0.496
<b>Improvement compared to naïve classifier?</b>	No	Yes	Yes	No	No

# Appendix VI - Decision Tree LDA model 1



## Appendix VII - Decision Tree LDA model 2



## Appendix VIII - L-LDA Topics

A production of labelled topics showing the top 20 word stems for each activity class.

Activity class	Word stems
<i>Canoeing</i>	Order, kinderfeestjes, zoon, kleintjes, counter, ongelimiteerd, change, faciliteiten, omgeschreven, visa, verschonen, kiosk, intoetsen, heerde, verjaardagsfeestje, master, gebruikmaken, easy, toiletten, gerestyled
<i>Climbing</i>	Zwolle, zaal, stadsmuur, toren, Cuypers, Sassenpoort, Peperbus, poort, wwwleuketipnl, beiaard, e, pierre, stadsgids, beklimmen, zwager, verdieping, verbetering, bevindt, hanze, buiten
<i>Cooking</i>	Workshop, we, personen, bonbonnerie, borrel, zaterdag, deelname, chocoladeworkshop, reservering, groepsdeelname, tijdsduur, chocolade, persoon, per, maand, kosten, verplicht, slag, gaan, mee
<i>Dancing</i>	Bommel, jack, club, underground, evenementen, reactie, bloopers, onvoorstraat, parkeerplaatsprijsklassegestald, dansennederlands, spreken, bruut, terrasseizoen, stappen, uurtjes, feestbar, no, speciaalbiertje, gemarkeerd, soundcloud
<i>Drinking</i>	Zwolle, t,s, we, koffie, onze, caf, bier, wij, plek, barista, kunt, beoordelingen, lunch, wijn, alle, informatie, zondag, waar
<i>Eating</i>	Zwolle, eten, sfeer, service, kwaliteit, prijs, drinken, schaal, we, onze, restaurant, augustus, wij, s, gegeten, heerlijk, lekker, italiaanse, echt, waar
<i>Listening</i>	Dienst, live, zwolle, podia, uur, thor, a, verkocht, bv, by, gesteld, diverse, s, show, scheepswerf, ijmuiden, schip, tijdens, hasselt, terug
<i>Meditating</i>	Patipada, gerjan, webshop, cart, empty, your, naast, meditatiecentrum, patipadanl, schoemaker, wishlist, meditatie, yoga, emptyyour, weg, artikelen, achter, mee, onze, wij
<i>Playing</i>	Stoof, uur, zaterdagavond, dj, prijzen, elke, aanvang, mail, night, sup, Friday, studentenavond, gregis, dartteams, donderdagavond, wave, reggae, hits, rock, cookies
<i>Relaxing</i>	Zwolle, park, wikipedia, massage, stadsstrand, wezenlanden, potgietersingel, informatie, ter, spa, b, pelkwijkpark, beauty, wwwleuketipnl, ligt, zie, we, schoonheidsinstituut, cadeaubon, Chinese
<i>Sailing</i>	Zwolle, rondvaart, wwwleuketipnl, stadsgids, verbetering, vanaf, water, gratis, ca, bekijk, doorgeven, plek, bezienswaardigheden, ontdek, vele, website, hanzestad, partyschip, bbq, ak
<i>Shopping</i>	Zwolle, onze, wij, winkel, augustus, we, cookies, nederland, website, s, waar, graag, informatie, pistache, samen, uur, wwwleuketipnl, zwolse, nieuwe, koffie
<i>Sightseeing</i>	Zwolle, km, beoordelingen, weet, nee, zeker, sassensstraat, afstand, monument, beoordeeld, raadhuis, wikipedia, informatie, nederland, aanbevelen, activiteit, mei, tripadvisor, plek, zie
<i>Sitting</i>	Zwolle, augustus, kerk, mizu, zoeken, indebuurt, l, zwolse, zon, close, terrasje, isra, synagoge, waalse, tussen, page, pakken, joodse, werk, plekken
<i>Sporting</i>	Order, kinderfeestjes, zoon, kleintjes, counter, ongelimiteerd, change, faciliteiten, omgeschreven, visa, verschonen, kiosk, intoetsen, heerde, verjaardagsfeestje, master, gebruikmaken, easy, toiletten, gerestyled
<i>Strolling</i>	Monument, herdenkingsmonument, park, drinkfontein, zuilen, geplaatst, augustus, diabas, indi, oudste, ramaekers, guinea, ernstig, kopie, wwwparkeekhoutnl, namens, duitse, stadsparken, urn, honderden
<i>Tasting</i>	Winkelwagen, uur, librije, tijdens, atelier, workshop, kok, slag, gaat, bereiden, culinair, chef, heerlijke, opgelet, ongehutste, pot, leiding, tips, browser, prachtige
<i>Watching</i>	Zwolle, langhuis, stichting, jaar, path, kunst, pand, nieuwe, kunstenaars, galerie, witte, alle, commissie, exposities, culturele, museum, activiteiten, brouwerij, unlimited, samenwerking