Diplomarbeit

Applicability of Social Media Analysis for Spatial Planning

ausgeführt zum Zwecke der Erlangung des akademischen Grades
eines Diplom-Ingenieurs
unter der Leitung

von

Projektass. Dipl.-Ing. Dr.techn.
Julia Forster

E280 - Institut für Raumplanung

und

Ao.Univ.Prof. Dipl.-Ing. Dr.techn.
Andreas Voigt

E280 - Institut für Raumplanung

eingereicht an der Technischen Universität Wien

Fakultät für Architektur und Raumplanung

von

Balázs Cserpes

01226235

Wien, am 28. 11. 2018

Ich erkläre an Eides statt, dass die vorliegende Arbeit nach den anerkannten Grundsätzen für wissenschaftliche Abhandlungen von mir selbständig erstellt wurde. Alle verwendeten Hilfsmittel, insbesondere die zugrunde gelegte Literatur, sind in dieser Arbeit genannt und aufgelistet. Die aus den Quellen wörtlich entnommenen Stellen, sind als solche kenntlich gemacht.

Das Thema dieser Arbeit wurde von mir bisher weder im In- noch Ausland einer Beurteilerin/Beurteiler zur Begutachtung in irgendeiner Form als Prüfungsarbeit vorgelegt. Diese Arbeit stimmt mit der von den Begutachterinnen/Begutachtern beurteilten Arbeit überein.

Wien, am 28. 11. 2018

Balázs Cserpes

Kurzfassung

Die fortschreitende technologische Entwicklung übt einen starken Einfluss auf die Wahrnehmung und Planung von Städten aus. Einen ausschlaggebenden Faktor bildet die reichliche Verfügbarkeit von Daten. Die Menge an Informationen, die jeden Tag generiert werden, wächst exponentiell und beinhaltet zahlreiche Aspekte des Alltags. Für viele ForscherInnen stellt das Phänomen Big Data eine potenzielle Bereicherung für die Wissenschaft dar. Durch die Anwendung neuer Technologien können potenziell auch PlanerInnen tiefe und genaue Einblicke in das komplexe System der Städte erhalten.

Allerdings stellen solche Technologien auch einen starken Eingriff in die Privatsphäre der Menschen dar. Viele KritikerInnen deuten auf missbräuchliche Verwendungsmöglichkeiten solcher Daten hin. Außerdem kann die Überschätzung der Potenziale von Big Data zur Ziehung von falschen Schlussfolgerungen führen.

Die Potenziale sind trotzdem vielversprechend, daher lohnt es sich, neue Datenquellen zu untersuchen. Eine solche neue Quelle stellen Social-Media-Portale dar, die von Millionen von Menschen weltweit genutzt werden. Die NutzerInnen solcher Seiten generieren Daten, die für die Raumplanung eine bedeutende Bereicherung darstellen können. Social Media Geographic Information (SMGI) beinhaltet nicht nur Informationen darüber, wo sich Menschen wann aufhalten, sondern gewährt auch Einblicke in Diskussionen und subjektiven Wahrnehmungen von bestimmten Phänomenen. Dadurch kann sie helfen, die Dynamiken des komplexen Systems der Städte zu verstehen.

Allerdings ist die Verarbeitung von SMGI aufgrund ihrer fehlenden Strukturiertheit, ihrem Mangel an Reliabilität und Validität, sowie der daraus resultierenden Anfälligkeit für Fehler, erschwert. Jedoch stehen wertvolle Potenziale demgegenüber, die aus der räumlichen, zeitlichen und thematischen Flexibilität dieser Daten stammen.

Ein wichtiges Ziel der vorliegenden Arbeit ist, die Methoden zur Aufbereitung von Big Data darzustellen. Insbesondere wird auf Natural Language Processing (NLP) eingegangen, mit Schwerpunkt auf Topic Modelling (zur Erkennung der Inhalte der Texte) und Sentiment Analysis (zur Erkennung des subjektiven Empfindens).

Das Hauptaugenmerk dieser Arbeit liegt bei der Definition und Bewertung der konkreten Potenziale von Social-Media-Analysis für RaumplanerInnen. Das Projekt beschäftigt sich mit der Frage, wie man derartige Informationen in der Raumplanung in einer sinnvollen, anwendungsorientierten und ethisch korrekten Weise nutzen kann.

Dazu wurden im Rahmen einer Fallstudie zwischen Mai und Oktober 2018 über die Streaming API von Twitter 8,3 Millionen Tweets erfasst. Diese wurden im Anschluss mithilfe der oben angeführten NLP-Methoden auf ihren Inhalt, sowie mithilfe von GIS auf ihrer räumlichen Verteilung analysiert. Dabei stand die Qualität und Nützlichkeit für die Raumplanung im Vordergrund.

Trotz der beträchtlichen Größe des Datensatzes, war die Menge an wertvoller Information sehr niedrig. Die Tweets konnten in einigen Fällen der Realität korrekt widerspiegelten, beispielsweise ist eine Anhäufung von Nachrichten in der Nähe zu Schulen und Universitäten zu erkennen, die sich mit dem Thema Bildung beschäftigten. Allerdings standen andere (für Raumplanung relevante) Indikatoren nicht mit validierten statistischen Daten im Einklang.

Aus diesem Grund schätzt die vorliegende Arbeit die Nützlichkeit von Twitter-Daten in der Raumplanung als niedrig ein. Obwohl die Daten ein großes Themenspektrum abdecken und eine hohe Flexibilität aufweisen, sind sie in der Regel nicht valide. Des Weiteren kann die Überbewertung von Big Data auch zu einem Wiederaufkommen des „Gott-Vater-Modells“ von Planung führen. Daher ist es bedeutend, als PlanerIn sich mit dem Phänomen Big Data kritisch auseinanderzusetzen, um die Gefahren und Potenziale realistisch abwägen zu können.

Abstract

The ongoing technological advancement has a huge effect on the ways how we view and plan our cities. A key factor in these developments is the abundant availability of data. The amount of information we generate every day grows exponentially and covers more and more aspects of our lives. Various researchers agree that Big Data has the potential to become a rich and fruitful asset for spatial planning. Planners might become able to monitor multiple aspects of our cities in real-time, making it possible to set measures or interventions immediately.

On the other hand, these developments also propose a threat to privacy of people's lives. Risks of misusing the recorded data have been addressed my many critical thinkers. Besides, even when having good intentions, the excitement surrounding Big Data may easily lead to drawing false conclusions.

Still, new technologies and Big Data have some enormous potentials which can help us to secure and enhance many aspects of quality of life in our cities. A novel data source are social media sites, used by millions of people worldwide. Users of these portals also generate data having the potential of becoming highly valuable for spatial planning. Social Media Geographic Information (SMGI) contains not only information about the location of people at certain times, but enables to localise discussions and sentiments of people towards specific topics. Therefore it might help us to understand the dynamics of cities and unveil some knowledge we had no access to before.

By its nature SMGI also often conveys many challenging characteristics of Big Data. It is vaguely structured, very diverging in quality, and cannot be assessed by traditional methods. Furthermore it is heavily biased, lacks validity, and representativeness. Still, when applying the correct pre-processing and analysis steps, its spatial and temporal flexibility, together with the broad range of its content promise to become a highly valuable source of information for spatial planning.

A main aim of the project was to show how data processing works and to present Natural Language Processing methods currently mostly unknown in the domain of spatial planning, such as topic modelling and sentiment analysis.

The master's thesis tries to define and assess the concrete value of social media analysis for urban planners. It deals with the questions of how to make use of such data in a clear, correct, ethical and useful way.

To answer these questions, a case study was conducted by recording all tweets from May to October 2018 in London through Twitter's Streaming API. The 8.3 million captured tweets were analysed on their spatial and temporal distribution, their content and their sentiment measures, and the combinations of these factors. Then, the results were assessed on their quality and the usefulness for spatial planning.

In conclusion, the amount of valuable, valid and useful information that could be extracted from these 8 million tweets was very little. In many cases, Twitter data did reflect real world phenomena, for example by showing that there is a higher activity of tweets dealing with the topic education around schools and universities. Still, topics and themes being potentially more useful for spatial planning, such as crime, social controversies or user's sentiment scores, didn't reflect real-world indicators.

For this reason the following assesses work the value of Twitter data for spatial planning to be relatively low. Although the downloaded data displays an enormous flexibility regarding space, time and content, it is generally not reliable enough. Furthermore leads the overassessment of new technologies to the reappearance of the God-Father-Model of planning. Therefore it becomes crucial for planners to evaluate the qualities, dangers, and potentials of Big Data critically.

Table of Contents

1. Introduction 1

2. Scope 5

2.1. “The End of Theory“ 5

2.2. What is Big Data? 5

2.3. Big Data and Society 6

2.4. Big Data and Social Research 6

2.5. Big Data and Spatial Planning 8

2.6. Volunteered Geographic Information 9

2.7. Social Media Geographic Information 10

2.8. Twitter 11

2.9. Relevant Projects 14

2.10. Further Projects 16

2.11. Critique 16

2.12. Ethical Questions 20

3. Methods 23

3.1. Accessing the Data 23

3.2. Anatomy of a Tweet 27

3.3. Data Preparation 27

3.4. Getting the Message 33

3.5. Allocating the Message 40

3.6. Implementation in Python 42

4. Case Study 45

4.1. Companion Website 45

4.2. Scale 46

4.3. Measures/Filters 47

4.4. Spatial and Temporal Distribution 49

4.5. Content 52

4.6. Internal Quality 56

4.7. External Quality / Usability 58

5. Conclusions 69
5.1. Lessons Learnt 70
5.2. Further Steps 72
5.3. Recommendations 74
References 79
Literature 80
Programming Components 89
Appendix 91
Queried Topics 91
Queried Amenities 105

1. Introduction

2.5 quintillion bytes. Or 25 million terabytes. That is the estimated amount of data humanity created every day in 2018. Probably unsurprisingly, but this number is growing exponentially. 90 per cent of all data available today was created just in the last two years (Marr 2018).

It's needless to say that technology shapes the way we look at the world and the way we behave. The number of mobile phone subscribers hit in 2017 the 5 billion mark (GSMA 2017). Countless more numbers could be presented to show to which extent the technical revolution transformed our everyday lives.

Numerous characteristic terms have emerged in the recent years, all trying to explain different aspects (or sometimes the whole phenomenon) of technological advancement. Big Data, Petabyte Age, Network Society. These buzzwords all refer to a world where information can be generated and shared in a volume and speed that has been unthinkable even a few decades ago.

Technological advancement has transformed the ways we create and disseminate data completely, especially when talking about geographic information. For a long time mapping our world was a complex task, mostly carried out by cartographers, who had the knowledge and access to the equipment needed to survey our environments. Nowadays almost every mobile phone is equipped with a GPS receiver enabling the localisation our position within a few seconds.

Together with the emergence of Web 2.0 this has led to a decentralisation of geographic data creation. Goodchild coined the term "Volunteered Geographic Information" (VGI) in 2007 where he described the phenomenon that nowadays maps can be created by people who have no (or very limited) cartographic knowledge. Probably the most prominent example is OpenStreetMap, a mapping service operating solely on basis of data provided by volunteers (OpenStreetMap Contributors n.d.).

Yet, geographic data is also created to some extent unintentionally. By using social media sites, our messages also often contain some kind of locational information. Social Media Geographic Information (SMGI) does not only reveal our position, but is equipped with further information (Campagna 2016).

SMGI promises to provide a clear picture about a number of aspects of our everyday life. Local governments and planners may use it to access information about the dynamics that characterise the complex systems of our cities. In this context SMGI may provide valuable knowledge as a basis for planning decisions (Campagna 2016). Several research projects tried to extract information from social media posts and identified spatial and temporal distribution patterns of people, languages, topics, sentiments, and various other subjects.

At the same time, Big Data and consecutively SMGI, are not easy to handle. Such data is large in size and contains a broad range of information, but this information is available in a much unstructured form. It is often hard to find and in many cases lacks valuable metadata. SMGI is furthermore characterised by a lack of representativeness, is heavily biased, and lacks validity (Shelton 2017).

Still, when applying the correct pre-processing and analysis steps, its spatial and temporal flexibility, together with the broad range of its content promise to become a highly valuable source of information for spatial planning. Therefore we can observe an eager discussion about the value of SMGI. By some it is regarded as a rich source of knowledge, others feel that its drawbacks make it unusable in serious applications.

This discussion about the usability of SMGI is the main motivation behind the following master's thesis which is based around the following question:

Are we able to capture reliable and (in terms of spatial planning) useful information through social media analysis?

Of course this question needs to be broken down into several parts. Main aspects of the research question can be assorted into three categories by which the following work is structured.

The third chapter (Scope) starts with defining the main topics of this project. It addresses the questions how technological advancement shapes our cities, our society, and research. Furthermore it presents the concept of VGI and SMGI. Potential areas of application of such information are sketched by reviewing some relevant research projects. The chapter continues with describing the most important and relevant shortcomings of SMGI. Finally it addresses the most important legal and ethical questions that come up when dealing with (personal) social media information. The main research questions for this chapter can be defined as follows:

- What is Big Data, how can it be defined and what specific characteristics does it carry?
- How do the discussions surrounding Big Data shape our society, our cities, and the ways we research social phenomena?
- In which ways has technological advancement shaped the ways we create and disseminate (geographic) data?
- Which phenomena are we potentially able to approach via social media analysis?
- What is the current state-of-the art in research?
- Which pitfalls and critical points of Big Data- (and SMGI-) analysis have been identified in literature?
- As SMGI contains highly personal data, how can we ensure the full privacy of the authors of the analysed content and what legal framework needs to be considered?

As SMGI (and Big Data) carry some specific characteristics, they do not allow extracting information through traditional ways. Therefore the third chapter (Methods) deals with methodological questions. It presents some ways of data analysis being (currently) generally unconventional in spatial planning. The chapter begins with describing how to access Twitter's servers through the portals APIs (Application Programming Interfaces) and presents the information its dataset may contain. Then, the basic terms, approaches, and definitions of computational linguistics and Natural Language Processing (NLP) are defined. The chapter continues with presenting ways and methods how to analyse the content of texts, focusing on semantics and sentiments.

As this project deals mainly with the geographic distribution of Twitter content, this chapter also explains the ways how the portal handles locational attributes and how such information should be prepared in the most accurate and correct way. Finally, the chapter also describes the software architecture and basic structure of the project. Research questions of the third chapter are the following:

- How can we access Twitter data?
- What kind of attributes does a tweet contain?
- What are the basics of computational linguistics and NLP?
- Which methods underlie information extraction in NLP?
- How does topic modelling and sentiment analysis work?
- How does Twitter deal with geographic information?
- What challenges/pitfalls need to be addressed?
- How can we aggregate and visualise this locational data correctly?

The case study constitutes the third chapter. From May to October 2018, more than 8 million tweets were captured in the city of London. Aim of the case study was to identify the most relevant contents of social media regarding urban planning. Furthermore, an important objective was to evaluate the quality and value of Twitter content as a source of information for spatial planning. This was done by assessing the (general) internal quality through some predefined indicators as defined by Devillers and Jeansoulin (2006a). The external quality (usability) was tested on the basis of three theses, namely inspecting the downloaded data as a source of information about the spatial and temporal distribution of people, topics, sentiments and the combinations of these. Concrete research questions are the following:

- What are the basic characteristics of the downloaded content?
- Which topics does the applied topic model identify?
- Is there a clear spatial and/or temporal distribution of tweets and their content identifiable?
- Are these patterns reliable/plausible? Do they reflect real phenomena and/or validated demographic indicators?
- Does Twitter data possess a sufficient quality to be applied as a source of information in a spatial planning context?

The final chapter (Conclusions) recapitulates the project and its results in general. It shifts the focus back to the questions surrounding new technologies and Big Data and reflects the critical discussions with focus on results of the research project. It summarises the most important findings, their relevance for planning and gives an answer to the main research question.

As the results are very heterogeneous, both because of the large number of topics, the different calculation, and filtering methods, it would have been impossible to present all the combinations of these factors in a static form. Therefore an interactive companion website was set up which contains all the maps and figures created during the course of the case study. It allows to combine all the different contents and factors that were assessed during the project. The website is accessible via the following link:

www.londontweets.eu

2. Scope

2.1. “The End of Theory“

It is maybe Chris Anderson’s provocatively titled article “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete” that serves as a good example when trying to describe the enormous cultural and societal expectations towards Big Data. The text was published in the Wired in 2008, the magazine he was editor in chief at that time. In Anderson’s opinion, because of the abundance of data and information, there is no more need for a theoretical scientific foundation of knowledge (Anderson 2008).

The sheer amount of data enables us to develop models that allow us to describe and prognosticate our world accurately. He cites Google’s advertisement engine as an example, which essentially consists only of applied mathematics and has no understanding about the culture or conventions of advertising. Still, the system works very well. In his opinion, “[w]ith enough data, the numbers speak for themselves” (Anderson 2008).

The example for advertising is just a mere indication of something much larger. In future, data can become the sole basis of science. The current hypothesis driven approach can be replaced by a purely data-driven procedure. This marks for Anderson “The End of Theory”, where “[c]orrelation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all” (Anderson 2008).

Although it doesn’t appear in Anderson’s text, the term Big Data has shaped the discussions around the potentials, chances and also dangers of this large amount of information. It became a commonplace to refer to the enormous importance that is ascribed to information. The quote by the British mathematician Clive Humby, namely “Data is the new oil” is a frequently used aphorism to describe its relevance in our everyday lives.

Catchwords as information age, network society, petabyte age, and a lot more try to describe how technological innovations will shape our lives in the next decades. Such concepts have of course also arrived in urban and spatial planning, maybe most importantly in connection with the umbrella term “Smart Cities”.

Still, before inspecting the impact of Big Data on our society, it is important to define what the term means, as despite its frequent usage in discussions, its definition still often remains unclear (De Mauro, Greco, and Grimaldi 2016, p. 122).

2.2. What is Big Data?

In general, various characteristic aspects of Big Data can be used to describe its meaning. Big Data, is by definition, characterised by a great volume. It is also generated very quickly and often needs to be processed in real time. Furthermore, Big Data is unstructured, containing very heterogeneous information, both in its quality and its content. These three aspects – namely volume, velocity, and variety – are the most commonly used definitions of Big Data when illustrating it based on its attributes and main characteristics (De Mauro et al. 2016, p. 128ff).

From a functional point of view, one can define Big Data by its requirements neccessary for its analysis and processing. Big Data cannot be handled with an ordinary (even high-end) home computer. The required computing capacity lies at a much higher level and needs often multiple computers to work at the same task at the same time (De Mauro et al. 2016, p. 130f).

Big Data can also be described by its form, which characterises also the general approach how it should be prepared, processed and saved. Due to its size and structure, it often cannot be stored in traditional (relational databases) and processed with traditional approaches. The amount and the unstructuredness of Big Data needs and induces the development of new approaches and fuels technological innovation (De Mauro et al. 2016, p. 130f).

Moreover, Big Data can be defined based on the social impact it creates. Big Data is a "cultural, technological, and scholarly phenomenon that lies on the interplay of technology, analysis and mythology" (boyd and Crawford 2012 as cited in De Mauro et al. 2016, p. 130). Big data does not only shape technology but also impacts the whole society in general, and the way we look at the world.

Concluding the points above, De Mauro et al. (2016, p. 131) define the term as follows:

> *"Big Data is the Information asset characterised by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value".*

Although this definition covers the most important technological aspects, it unfortunately doesn't include the societal aspects of Big Data. From a spatial planner's point of view, the latter might be the most important. Even if planners work with lots of information, they usually won't pre-process Big Data, therefore the technological aspects are not neccessarily the highest priority. Planning decisions are not only shaped by the cultural and mythological attributes of Big Data, buzzwords as „Smart Cities" also contribute to the enhancement of the imortance of new technologies.

2.3. Big Data and Society

The cultural and societal impact baceame apparent in Chris Anderson's text too. He didn't only describe a possible change Big Data can induce on our society but also provoked a discussion about the role of the large amount of information will have on the way we look at the world. Up until today, his thesis is frequently referred in texts as an illustration for the enormous expectations many towards Big Data have (for example in Crampton, Graham, Poorthuis, Shelton, et al. 2013; Kitchin 2016; Shelton, Poorthuis, and Zook 2015).

The provocative position of his text led to some very critical debates with many researchers pointing out that although Big Data is (and is becoming) a valuable source of knowledge, it still shouldn't be regarded the sole basis of knowledge.

It is of course possible to gain understanding through the observation of certain patterns in large amounts of data. This approach is used in many cases, for example in bioinformatics. As Mazzocchi (2015, p. 1251f) notes, explorative analysis have helped to gain knowledge in extremely complex systems such as genetics or in exploring molecules.

At the same time, it is still important to emphasize two main fundamentals of scientific work. Both are highly unlikely to change, even if the way we look at things will become much different in future than today. The first fundamental is that correlation does not always mean a real causation. The second one is that it is important to understand the causes of a phenomenon before being able to apply it in practice (Mazzocchi 2015, p. 1252).

It lies in the nature of research that data cannot be atheoretical. Experiments working with a huge amount of data (for example at the LHC) also bear a theoretical foundation. Without theory, it would be impossible to answer the question of what to look at in the first place. As Mazzochi puts it: "Pre-existing assumptions create expectations on how the world should function, and it is these assumptions and expectations that allow us to detect the odd things." (2015, p. 1254)

2.4. Big Data and Social Research

Nevertheless, Big Data analysis still has the potential to assist us in discovering novel and unforeseen phenomena. It is just crucial to start an analysis with realistic assumptions and expectations.

Members of the Parisian research centre "Sciences Po médialab" started working with digital with enormous expectations. Although they didn't manage to gain the amount of knowledge they expected from the amount of data they had, they discovered some ways digital data can transform the traditional means and methods of sociological research. They have recognised "how the digital transformed our relation to the data, the methods and the theory of social research" (Venturini, Jacomy, Meunier, and Latour 2017, p. 2).

Basis of their projects was the collection and analysis of digital traces, which are defined as "inscriptions as *originally* [BC: cursive in original] produced by digital devices" and transformed into "data" to become "useful knowledge objects" (Venturini et al. 2017, p. 3).

They projected the discovered changes onto three dimensions, namely the "Continuity in data", the "Continuity in methods" and the "Continuity in theory". These three dimensions are described briefly below:

2.4.1 Continuity in data

Typical for Big Data is its feature to be completely unstructured and its need for an extensive pre-processing phase before being able to be analysed. At the same time, the quality of the data cannot be derived just from its quantity. Lots of traces don't automatically mean lots of information. In many cases, digital data isn't able to provide the same amount of inforamation, as traditional would (Venturini et al. 2017, p. 2ff).

What's really novel is the range flexibility of Big Data regarding time, space and content. Its strength lies in its "breadth and depth [BC: cursive in original]", as it is "more diverse and more evenly distributed across the span of collective existence" (Venturini et al. 2017, p. 4).

2.4.2 Continuity in methods

A common premise in (social) science is the clear distinction between quantitative and qualitative methods. Digital data overcomes this distinction as it incorporates both quantitative and qualitative approaches and allows us to observe an analysed phenomenon in both ways (Venturini et al. 2017, p. 4).

A new feature of digital data is also its flexible situation and aggregation. In case of a study it is now much easier to enable the traceability of the research process. Research results don't need to be published just in an aggregated form, it is possible to show all the steps tracing back to the sampling of the data (Venturini et al. 2017, p. 5).

An adequate possibility to publish the results in a more transparent and reversible way is setting up an interactive "companion web-site". This website allows readers to navigate through the data interactively and makes it much easier to understand how the findings were discovered (Venturini et al. 2017, p. 5).

2.4.3 Continuity in theory

Societal research doesn't only distinguish between qualitative and quantitative methods, but the levels of observation are also clearly categorised onto three levels, namely the micro, meso, and macro scale.

Through the continuity in both data and methods also the distinction between these levels becomes less significant. The nature of digital data enables us to look at almost any level or scale of a process, for example from "the international funding decisions [BC: of climate change regulations] to the adaptation of projects carried out in specific districts of Bangladesh" (Venturini et al. 2017, p. 7f).

2.5. Big Data and Spatial Planning

Of course these expectations and approaches can also be transferred onto the domain of spatial planning. Cities can be interpreted as exceptionally complex systems, consisting of lots of actors, interactions, and subsystems. Using innovative computer-based technological solutions to help to understand how cities work and what kind of planning interventions are needed, dates back to as early as the 1970s.

The Chilean Project Cybersyn was implemented between 1971 and 1973 to track and monitor the country's economy in real-time (Beckett 2003), but also New York tried to develop a computer-based model for managing the city's fire fighting system (Townsend 2014, p. 80).

The first implementations of such approaches didn't led to achieve the desired results, in New York the computer-driven system even led to a disastrous level of inefficiency, the first excitements slowly faded. Ongoing technological developments still found their ways into city planning, especially after the millennium when a data-based holistic approach to monitor and plan our cities (re-)emerged (Townsend 2014, p. 81f).

"I have a rule of thumb: if you can't measure it, you can't manage it" is how Michael Bloomberg (2014, p. v) described his basic approach as a mayor of New York. He praises the large amount of available urban data as an extremely valuable asset for mayors, planners, and governments. It can bring "more transparency, accountability, and efficiency to government" and help to induce long-lasting changes that will improve the quality of life in cities efficiently (Bloomberg 2014, p. vi).

Today "Smart City" has become an important umbrella term to describe all actions where a city can make use of technological advancement (Cocchia 2014). Almost every city today has implemented different methods to monitor the urban system in real-time, with measures ranging from recording routes of passengers of public transport, energy consumption, and lots of more aspects.

There are several Smart City projects all around the world, including recently built (or proposed) urban projects such as Songdo in South Korea based on technologies of Cisco (Townsend 2014, p. 22ff) or Google's Waterfront in Toronto (Scola 2018), but such technologies have also found their way into existing cities on a larger scale. A good example is Rio de Janeiro, with its "Sala de Controle" (Control Room) where the real-time recordings of 400 cameras throughout the city are displayed on a single wall. In this way, the city government expects to capture the city as a whole at one single place and take action immediately when needed (Townsend 2014, p. 66f).

Although these technological measures do often enable to gain information in a very short time and possibly help to understand a city in its complete complexity, they also need to be viewed critically.

In the last years, lots of works have been published that deal with these controversies, for example Adam Greenfield's "Against the Smart City" (2013) describing how a few big corporations shape the image of the cities of tomorrow. Or Anthony Townsend with "Smart Cities" (2014), a book describing alternative ways new technologies can strengthen the power of citizens shaping their lived environments.

From a more theoretical point of view on spatial planning, Kitchin (2016, p. 3) identifies a phenomenon where "data-informed urbanism is increasingly being complemented and replaced by data-driven urbanism". Nowadays it became extremely easy to obtain urban data in such a scale and flexibility that was unimaginable even a few years ago.

From a philosophical point of view, as Kitchin points out, current urban science is shaped by two epistemological positions. The first is a form of "inductive empiricism in which it is argued that through data analytics urban big data can speak for themselves free of theory or human bias or framing". This form can also be compared to Anderson's article and is the less prevalent of the two positions (2016, p. 4).

The second attitude describes an approach that is still based on a scientific basis but identifies hypothesis from the data itself and tests the validity in the next step (Kitchin 2016, p. 4).

In general, these data-driven approaches lead to some far-reaching ethical questions, among others regarding the privacy of a city's citizens, but (as a central aspect for spatial planning) also data determinism. Applied to urban planning, this would mean that the analysed Big Data is not only expected to be used as a source for information about the processes that shape our cities, but also as a basis for concrete and fundamental decisions (Kitchin 2016, p. 10f).

Therefore, it is crucial for planners to be able to identify the value of information derived from Big Data and not assume uncritically that a big amount of data leads to a big amount of reliable and useful knowledge.

As a conclusion, Kitchin (2016, p. 11f) points out this dismissing the potentials of new technologies is neither the right approcach. However, it is important not to assume that technology and urban science can provide an "objective, neutral, God's eye views of the city". A city is a complex system and the data its inhabitants create is also just the outcome of other "ideas, instruments, practices, contexts, knowledge and systems used to generate and process them". Therefore, "data are never raw, but always already cooked".

2.6. Volunteered Geographic Information

Shifting the focus away from the (formal and informal) urban planning systems, technological advancement of course always shaped the way we interact with each other in our everyday lives. In close concordance with technological developments, the British-American cartographer Michael F. Goodchild recognises a change in how geographic information is being collected, produced and used. Nowadays not only professional cartographers generate data by making use of traditional methods. There are increasingly non-professional ones who are also able to make a substantial contribution to geography (Goodchild 2007, p. 211).

This phenomenon he calls the upcoming of "Volunteered Geographic Information" (VGI) and defines it as shown below:

> *"...the widespread engagement of large numbers of private citizens, often with little in the way of formal qualifications, in the creation of geographic information, a function that for centuries has been reserved to official agencies. They are largely untrained and their actions are almost always voluntary, and the results may or may not be accurate. But collectively, they represent a dramatic innovation that will certainly have profound impacts on geographic information systems (GIS) and more generally on the discipline of geography and its relationship to the general public. I term this volunteered geographic information (VGI), a special case of the more general Web phenomenon of user-generated content [...]"*
>
> Goodchild 2007, p. 212

In his paper, Goodchild discusses several significant developments that made the use of individually collected information possible. The introduction of "web 2.0" enabled citizens to contribute information on the internet more easily. Georeferencing of data can be now done without prior cartographical knowledge. Methods, systems and appliances like GPS, mapping programmes (e.g. Google Earth) or geotagging tools (e.g. Wikimapia[1]) can easily

[1] a service that is used to insert a standardised code for the location of the topic of a specific Wikipedia article

transform the contributed information into a standardized format. Likewise, improved and enhanced visualisation techniques as well as access to broadband communication channels can be considered essential contributing factors that play a significant role in creating and sharing VGI (Goodchild 2007, p. 214ff).

In addition to this, profound changes regarding organisational and institutional features of cartography have increased the importance of VGI. Since the beginning of the 1990s several countries abandoned or transformed their official cartographic agencies that used to be responsible for mapping the ubiquitous territory of their state. Nowadays data is collected from different sources and cartographic organisations are usually responsible for providing only services as standards and protocols. It's this what makes the interchangeability and comparability of spatial data feasible thus enabling the collection of information from different sources (Goodchild 2007, p. 217).

Crowdsourcing of geographic information is being encouraged by private initiatives such as "Christmas Bird Watch" or OpenStreetMap. But also mapping companies rely on individually collected data, for example by tracking car movements to gather information on traffic-related issues (Goodchild 2007, p. 217f).

It is important to emphasize that Goodchild's article was written in 2007, – the year iPhone was introduced (Ritchie 2018). Shortly before, the launch of Google's Android had taken place in 2006 (Ziegler 2012) and OpenStreetMap had gone online in 2004 (Leonardo, Mooney, and Minghini 2017, p. 38). These major changes were followed by ongoing technological developments and a rising importance of technical devices in our everyday lives that made the role of VGI grow rapidly in prominence.

2.7. Social Media Geographic Information

A relevant subset of VGI that is increasingly of importance is called Social Media Geographic Information (SMGI).

It has – as its name suggests – its sources in social media channels like Facebook, Twitter, Instagram or other services. Social media sites have access to a huge amount of data provided by their users, including geographical and temporal references.

The term SMGI can be defined as

> *"any piece or collection of multimedia data or information with explicit (i.e. coordinates) or implicit (i.e. place names or toponyms) geographic reference collected through the social networking web or mobile applications."*
>
> Campagna 2016, p. 48

SMGI may include "texts, images, videos or audios" enhanced by metadata and it may become a valuable additional asset to the (traditional) "Authoritative Geographic Information" (AGI) (Campagna 2016, p. 48).

SMGI becomes probably most valuable in connection to the term "Geodesign" which refers "to a process able to inform design by geography in its broader holistic sense, including its physical, biological, social, cultural facets" (Campagna 2014, p. 598). It is commonly seen as a method making use of digital geographic tools and as a data-based approach carried out by multidisciplinary teams with a strong emphasis on citizen involvement (Campagna Michele, Steinitz Carl, Di Cesare Elisabetta Anna, Chiara Cocco, et al. 2016).

Campagna compares AGI to SMGI as follows:

AGI consists of the geographic coordinates of an instance with one or more attributes that are normally stored in a relational database model (Campagna 2016, p. 49). An example of this could be a GIS layer of a city's boroughs, where the extent, form and location of its territory are defined in a shapefile and the attributes (population, age structure, area, etc.) are linked via the attribute table.

By contrast, a piece of SMGI is made up of the coordinates and a timestamp, the content (which can be text, video, picture, audio clip or a combination of any of these attributes), the user and the online community's various virtual responses (likes, favourites, count of retweets, etc.) (Campagna 2016, p. 49).

This distinction is also comparable to the findings of the Sciences Po médialab (Venturini et al. 2017) – see *2.4. Big Data and Social Research – p. 6*. Social media data should be used differently than "traditional" data, as its value is revealed more easily by analysing networks, connections and combinations. For example, a Twitter user can be described by attributes like their language or location. In certain cases, information about interactions between users can be however more relevant.

Campagna describes on the basis of his theoretical framework, knowledge, and experience gained from his practical projects, several approaches to different ways of acquiring information by means of a GIS-based analysis of SMGI (2016, p. 51):

- "Spatial analysis of user interests" can help to gain insight on the importance of certain places by looking at the density of social media activities of a specific community.
- "Temporal analysis of user interests" lays the main focus on the question of how and by whom are different spaces used at different times.
- "Spatial statistics of user preferences" can help to discover the needs and perceptions of users and different user groups like young or elder, families, or tourists, travelling in groups or alone.
- "Multimedia content analysis on texts, images, video or audio" is often the basis of the above types of work but its validity and significance are still constrained by the state of the art of available technologies (for example text analysis modules).
- "User behavioural analysis" is about looking at the types of people using spaces, where they come from, where they go and how much time they spend at certain places.

Of course a "combination of two or more of the previous" approaches can also be applied to get a more sophisticated and reliable image of our society looking through the lens of social media and to get access to completely new kinds of information. The potentials of working with SMGI may be summarised as gaining access to insights about user "perceptions or needs, opinions on places, daily routine events, so helping to get better insights on local identities" (Campagna 2014 as cited in Campagna, Floris, and Massa 2015, p. 43).

With ongoing technological advancements and an increasing importance of citizen participation, SMGI may become an increasingly important part of our urban and regional planning systems.

For example Huang and Gartner (2016) draw up potentials of using SMGI for capturing people's affective responses to different environments. This would enable to get insight into factors like safety or attractiveness of certain places. A huge promise of SMGI lies in the fact that it enables us to get information without expensive and time consuming laboratory or empirical field experiments (2016, p. 386).

2.8. Twitter

Twitter is a microblogging platform and (as of October 2018) with 325 million users the 7th biggest social media site worldwide (Kemp 2018).

The content (created solely by the portal's users) on the platform covers a broad range of topics. Twitter has the reputation to be a source for quickly spreading news with lots of users posting messages, pictures or videos when a major event occurs. These posts are also often embedded and incorporated in news articles (Ahmed, Bath, and Demartini 2017, p. 82f).

Setting up a profile is very simple, only a username, a password and an active e-mail-address (or telephone number) are needed (Twitter Inc. n.d.-a).

Users can enhance their profiles by adding further data, they can compose a short biotext about themselves, select a location (which can be a city, a country, or any real or fictious geographical place, like "The Death Star"), their birthday and a link to their website. Furthermore it is also possible to upload a profile and/or a header picture (Twitter Inc. n.d.-b).

After registration, any user can create a tweet, which is a short text message that shows up on the user's timeline. This timeline is by default settings public (thus visible to everyone, including not registered users) but the user can set the profile to be visible only to their followers.

A tweet can be enhanced by adding one or multiple pictures, a video, a hyperlink or a small survey (Twitter Inc. n.d.-c). The length of a tweet was limited initially to be shorter than 140 characters. Since 2017 also longer messages (up to 280 characters) are allowed (A. Rosen and Ihara 2017).

The content of a tweet is not restricted or monitored by the portal, except in cases where a violation of the User Agreement is reported. These cases include "copyright or trademark violations, impersonation, unlawful conduct, or harassment." (Twitter Inc. 2018a)

Users can engage with each other by liking, "retweeting" (the tweet of the other user shows up on the retweeter's timeline) or commenting on each other's tweets. To get the newest updates automatically, users can follow each other. In this case tweets created by the followed user show up on the main page of the follower. Furthermore, users can also send each other direct messages. These are not visible for the public (Ahmed et al. 2017, p. 83).

Two characters play a distinct role in the portals content. An "@" refers to a specific "@user" while a "#" (hashtag) to a specific "#topic". Latter has gained importance at signalising specific trends (Ahmed et al. 2017, p. 83)., for example tweets using the hashtag "#PeoplesVoteMarch" refer to a London demonstration for a second Brexit-Referendum.

2.8.1 Twitter Demographics

As of 2018, there are 12.6 million Twitter users in the United Kingdom, which accounts for 47 per cent of the country's online adult population (Battisby 2018).

A major challenge in using Twitter as a source for a research project is the lack of metadata about the users' demographic attributes. Besides from a few verified accounts, the platform doesn't request any kind of information about the users, therefore there is no reliable information about their age, gender or location, making conducting a representative study almost impossible.

Data about the users is available either via "traditional" survey methods, like telephone interviews, or demographic indicators can also be estimated through proxies – using "estimates of the user's characteristics based on a set of rules" (Sloan 2017, p. 1).

The quality of these proxies was tested by Sloan (2017) by comparing the automatically detected demographic indicators to results of the 2015 British Social Attitudes (BSA) report, where representative indicators were presented about the British Population, including their use of Twitter. It was compared to two studies, one estimating the gender (Sloan, Morgan, Housley, Williams, et al. 2013) derived from the username and the other inferring the age and a socio-economic classification based on profile the users' profile descriptions (Sloan, Morgan, Burnap, and Williams 2015).

Results have shown that female users are overrepresented through these indicators. The BSA report states that of the British Twitter 57 per cent of Twitter‘s British users are male and 43 female (Sloan 2017, p. 3). Estimations through proxies indicated a higher (51.2%) rate of females (Sloan et al. 2013 as cited in Sloan 2017, p. 4). This phenomenon may stem from the high number cases where gender identification was not possible (86.3%) but also from different behaviour patterns of females and males in online media (Sloan 2017, p. 4).

Regarding age, it is important to note that while Twitter allows users to register above the age of 13 (although this is not checked by the portal upon registration), the BSA report only included the population above the age of 18 (Sloan 2017, p. 4). Still, estimates show an over-counting of younger users. In this case the derivation may be explained through the different usage patterns of different demographic groups. Younger users may use the portal for "social interactions with peers" while older users may "prefer to use Twitter as a news source" (Sloan 2017, p. 4). Another reason might be "identity play", defined as "pretending to be someone they are not or presenting what they perceive to be a more desirable virtual self" (Sloan 2017, p. 5).

1	Higher managerial, administrative, and professional occupations
2	Lower managerial, administrative, and professional occupations
3	Intermediate occupations
4	Small employers and own account workers
5	Lower supervisory and technical occupations
6	Semi-routine occupations
7	Routine occupations
8	Never worked and long-term unemployed

Table 1: NS-SEC Analytic classes (Office for National Statistics 2010, p. 3)

The socio-economic classification is based on occupation groups (as listed in Table 1). It has been shown that these classes can be derived well by analysing the user profiles' metadata. On Twitter, there is an overrepresentation of classes 1 and 2, while classes 4, 5 and 7 are underrepresented (Sloan 2017, p. 5ff). In this case, it is important to note that user profiles may state hobbies (for example "photographer") that might be recognised as occupations through the proxy (Sloan 2017, p. 7).

Concluding, the above study has shown that social research based on Twitter data is aggravated by the factors of Twitter being not representative of the population and that it is not possible to estimate demographic indicators by the users' data. Furthermore the causes of these discrepancies can also only be guessed and not identified precisely (Sloan 2017, p. 9).

In addtion, the above study didn't discuss an important factor, namely the difference in activity patterns among different user groups. It is a crucial factor, as it has been shown that a large number of users only use the portal passively. Only about half of UK Twitter users check the network on a daily basis (We are Flint 2018) and a large share of the posted content stems from only a very small number of users (see *4.3. Measures/Filters – p. 47* and *2.11. Critique – p. 16*).

Another challenge is the very uneven distribution of the portal‘s users. The number of Twitter profiles varies not only between countries, but also within states. Generally speaking, bigger cities have a higher density of Twitter profiles than rural regions (Shelton 2017, p. 727).

In addition, social media sites often display large dynamics in terms of user behaviour, the numbers presented may be outdated in 2018.

For spatial research it is important to underline that only a very small fraction of the tweets is georeferenced (see *4.3. Measures/Filters – p.*

47), thus in case of a spatial analysis, only a very small subgroup of a nevertheless small and biased group of internet users is captured.

Yet, as also the examples below show, Twitter is still a popular source for different scientific studies. As Shelton (2017, p. 728) puts it, despite the bias and the lack of validity, georeferenced tweets "provide substantial advantages in urban analysis precisely because they aren't constrained by conventional areal units".

2.9. Relevant Projects

The portal is characterised by a significant openness at granting access to its data for external users and organisations (see *3.1. Accessing the Data – p. 23*). Furthermore also the general attitude of its users to post with a more public intention makes it for researchers safer to not overstep privacy boundaries (see *2.12. Ethical Questions – p. 20*). Therefore, there is a keen research interest in Twitter as a source of geographic information and several studies based on the platform's data have been published in the recent years.

García-Palomares, Salas-Olmedo, Moya-Gómez, Condeço-Melhorado, and Gutiérrez (2018) conducted a study about the relationship between different types of land uses and Twitter activity in the city of Madrid. The researchers downloaded 6.8 billion tweets from January 2012 to December 2013 which have been linked to the cities transport zones which are generally characterised by homogenous land use profiles (e.g. retail, culture, leisure, education, etc.).

Temporal user activities were recorded by dividing the tweets into 15-minute time slots throughout the day. Based on multiple regression models, it was possible to indicate and predict land use activities in different areas at different times in the city. Moreover the model was able to prognosticate distribution patterns of people and activites for a proposed urban development project, by projecting current land use activities on the new area (García-Palomares et al. 2018, p. 317f).

Mocanu, Baronchelli, Perra, Gonçalves, et al. (2013) gathered linguistic patterns on a global scale to indicate the distribution of languages. Not surprisingly, the most used language globally is English, followed by Spanish, then by several Asian languages (Indonesian, Malay, Japanese, and Korean), followed by Portuguese, Dutch, Thai and Turkish in the top 10 languages. The authors found a positive correlation (R^2 = 0.56) between Twitter penetration levels of countries with their corresponding GDPs per capita (2013, p. 5).

On a local scale (by identifying different language groups in a city based on Twitter activity) results have shown that "Twitter trends mirror census data quite accurately". The accuracy still depends on the adoption rate of the microblogging platform and English as the most commonly used language may distort the results. This effect is shown in Montréal, where the study found a 2.4 times higher number of English speaking Twitter users than French. Despite that, census data indicates the opposite language distribution. Another case study, in New York, provided language pattern results that were in line with official data of the distribution of ethnic communities and their corresponding languages (Mocanu et al. 2013, p. 5f).

With the help of Twitter data it was also possible to identify possible tourist groups. Their presence can be derived from analysing seasonal changes in language, although as the authors point out, the overrepresentation of certain languages/countries (they list Dutch as an example) has to be considered in the analysis (Mocanu et al. 2013, p. 6f).

A similar study (Lamanna, Lenormand, Salas-Olmedo, Romanillos, et al. 2018) analysed immigrant community integration levels in 53 world cities, based on language detection. The analysed cities were divided uniformly into

60 km² large grids, consisting of 500*500 m² cells. Derived from the comparison of language patterns inside these cells and the overall language distribution of the cities, an integration measure was calculated. The authors clustered the cities into three categories based on their integration measures. London showed both the highest overall number of languages detected and the most even distribution of communities among all cities.

According to the authors, Twitter data may become a useful measure tool of community integration scales. It can assist in monitoring community developments and compare spatial integration phenomena both on a geographic scale between different cities and on a historical scale between different time frames. As stated in the previous study, user bias need to be considered in the analysis, it is not always possible to detect all communities (for example the large Chinese speaking group of Barcelona didn't appear in the data). Furthermore is the overrepresentation of specific demographic groups (e.g. young people) throughout the users of the social media platform an important issue (Lamanna et al. 2018, p. 15).

In addition to the number of tweets and languages it is also possible to get qualitative information about the sentiment of users and their subjective perception of certain places (Mitchell, Frank, Harris, Dodds, and Danforth 2013). This study linked the structure of Twitter messages (word use and message length) and its sentiments to happiness levels (measured by different indicators, e.g. different well-being-surveys and composite health and peace indicators) and urban character indicators, such as obesity rates or education levels.

The results indicated significant correlation patterns with different well-being indicators, meaning that the sentiment of tweets can be used as indicators for measuring factors such as happiness. Furthermore, specific words may suggest a connection to specific demographic indicators. For example the word "cafe" shows a negative (r=0.509), whereas "mcdonalds" a positive (r=0.246) correlation with obesity levels (Mitchell et al. 2013, p. 10).

It is important to remark that this study used a really simple sentiment analysis approach. Furthermore, it did not filter out foreign language tweets (which led to more negative scores for texts containing the Spanish word "sin") and remove neutral stop words (for example "the", "at", "is"). The overall score of a text was calculated by taking the mean score of all its distinct words. Other, more sophisticated approaches are able to filter out stop words more effectively and are capable of taking the text's context into account to some degree (for example by identifying the effect of exaggerations, punctuations and further factors, see *3.4.1 Sentiment Analysis – p. 33*).

Lansley and Longley (2016) detected the content of Twitter messages and recognised distribution patterns of certain topics in London. They recorded all georeferenced tweets in central districts of the city from the 1st of January to the 31st of December 2013. The authors applied a probabilistic topic modelling approach (Latent Dirichlet Allocation - see *3.4.3 Topic Model Analysis – p. 37*) and localised the topics based on the corresponding tweets.

Similarly to the Madrid study (García-Palomares et al. 2018), the project took also different land use categories and amenities into account and the distribution of the topics of the messages was in line with the assumptions. For example the topics "fashion" and "shopping" seemed to appear closer to the Regent's Street (Lansley and Longley 2016, p. 93).

Besides usual spatial planning topics, a GIS-based analysis of Twitter data also enables to gain insight into further social issues. Yang and Mu (2015) selected all tweets in the United States containing the word "depress" (or its variations) between 5th of September 2013 and 5th of March

2014. The context of these tweets was defined by "non-negative matrix factorization". NMF is a data mining technique with the aim of identifying different contexts a word may appear in. For example, when taking the word "depress", the word can appear together with terms as "music", "songs", "listening", but also with some like "feel", "hate", "me". Based on these appearances, it is possible to detect the context of the single messages (Yang and Mu 2015, p. 220).

Users with depression were selected in case they tweeted about a depression topic at least five times in a two weeks period, potential users were also validated manually. The validity of results was checked by correlating the number of MDD-diagnosed users with county level census data (e.g. household income, ethnicity). According to the authors, their findings echo "some of the findings in the literature" (Yang and Mu 2015, p. 221) but don't elaborate further, to which accuracy.

2.10. Further Projects

In conclusion, the above studies signalise that there is a keen interest in working with social media data in spatial planning and geographic research. The projects gained some very interesting conclusions and covered a broad range of topics on different geographic scales.

Analysing, visualising and mapping of social media data is also a popular online content. Websites like OneMillionTweetMap, OmniSci Tweetmap, or Social Bearing provide interactive and appealing visualisations about georeferenced tweets, enhanced by further various analysis methods. Of course, these visualisations are in many cases just a demonstration of these companies' analytical skills, they provide further deeper and broader social media analytic services for businesses and organisations.

Mainstream media has also taken up such interactive social media visualisations. The Guardian's Datablog published maps that show how different events and topics spread around the world. These maps show various topics like football games (Rogers 2014a) , political elections (Rogers 2014b) but Twitter data can also be used to identify the distribution of ethnic groups based on usernames (Rogers 2012) .

The maps listed above were drafted and created by the data scientist Simon Rogers, his visualisations were also presented in other media, in the Atlantic's CityLab (Capps 2014) or the Time Magazine (Stampler 2013) for example.

Taking a look at the map (Rogers 2014c)[2] displaying every tweet referring to the death of Michael Brown on the day after his death[3] shows an increasing amount of popping up bubbles as time lapses. Twitter activity grows especially in larger cities in the US and in some European cities (especially London and Istanbul).

Another map (Rogers 2013)[4], working with the same principle but dealing with a completely different topic, the release of Beyoncés eponymous album, tells a very similar story. The bubbles are popping up at very similar places, indicating an increased Twitter activity in those cities.

2.11. Critique

Such mappings are actively discussed in the cartographic community. Twitter's reliability as a data source is debated, as well as the ways to visualise social media data correctly.

Kenneth Field, a British cartographer is an eager critique of these visualisations and has concluded his views on the above presented maps as being nothing more than an "eye candy". As he puts it, Twitter maps are strongly biased, the points account only for representing

[2] https://srogers.carto.com/viz/4a5eb582-23ed-11e4-bd6b-0e230854a1cb/

[3] Michael Brown was an unarmed black man who has been shot by the police on the street in Ferguson (Missouri) during daytime. The incident led to large (and in some cases violent) protests and to a discussion about racism and discrimination in the US police forces (Lopez 2016).

[4] https://srogers.carto.com/viz/337d9194-6458-11e3-85b5-e5e70547d141/

a minor part of the population, and simply expressed, the display of an absolute number of tweets visualises only an increased Twitter activity of users who have enabled sharing their position on their devices (Field 2013).

Limitations of geographic social media research have of course also been actively discussed in academic papers. Although the upcoming prevalence of VGI promises huge potentials, this new kind of data needs to be processed with care and caution.

A recurring caveat in literature regarding VGI is "digital divide" (see for example Crampton et al. 2013, p. 132; Shelton et al. 2015, p. 201). The term was coined in the 1990s and describes the unequal access to new forms of information technology across different demographic and social groups. Although, as van Dijk (2006) puts it, the term itself can lead to some misunderstandings. Using the word divide can suggest a sharp and absolute, "difficult to bridge" gap between two groups. In reality, the transition between the two extremes is of course much more fluent. Still it is important to bear in mind that social media research will most probably capture just a subset of the whole population and can hardly account for representative results.

Shelton et al. (2015) also identify a notion of digital divide in Louisville. African-American neighbourhoods of the city show a lower tweeting activity compared to predominantly white areas.

In contrast to some other social media portals (as Facebook), Twitter also lacks demographic information. Only secondary information is available about users, either based on internet usage surveys or automatic tagging of users (for example by identifying the gender of users based on their usernames).

Furthermore (as detected in the case study – see *4.3. Measures/Filters – p. 47*), a very little number of users account for a very large number of tweets. This phenomenon is also called the "1% rule" or "90-9-1 principle" and has already been observed in the 1990s during the upcoming of the first online chat rooms. This observation describes the fact that 90 per cent of the users (so-called "lurkers") of an online community do not contribute content to the forum or portal at all. 9 per cent ("contributors") contribute little content and only one per cent ("superusers") account for the most content available on the website (Van Mierlo 2014).

Another challenge in working with social media data is the identification and filtering of automated content created often for marketing purposes. These bots are in many cases capable of identifying trends and responding with relevant messages. A study (Crampton et al. 2013, p. 137) has identified "trend spam" tweets that incorporate a currently trending hashtag (in their case #LexingtonPoliceScanner) and post a promotion link to webshops selling phones or even sandals.

Georeferenced posts, that are obviously relevant for geographic research, account for only a really small fraction of tweets, studies estimate their amount to be around 1 to 3 per cent of all tweets posted (Crampton et al. 2013; Leetaru, Wang, Cao, Padmanabhan, and Shook 2013; Morstatter, Pfeffer, Liu, and Carley 2013). The accuracy of these coordinates is in addition often not correct (see *4.6.3 Positional Accuracy – p. 57)*.

In contrast to "traditional" survey methods, the data downloaded from social media channels frequently lacks validity. Twitter doesn't confirm the correctness of provided (meta-)data, for example, a user can provide their location as a place name (which itself may also be fictitious) (Crampton et al. 2013, p. 133). Of course, it is also possible to tweet from a user location set in Vienna while in London, or to publish a tweet from the Buckingham palace, while being virtually anywhere in the world.

Even with correct and cautious data preparation, many Twitter maps lead to false conclusions by making some (general) cartographic visualisation mistakes. As the number of users are unevenly distributed, the display of absolute number of tweets may distort the message completely. As stated above, Twitter trends have in general very similar spatial concentration patterns, with lots of activity in large cities (especially in the United States and in South-East-Asia). At the same time, in these regions with increased activity, a lot of information can get lost simply because points placed on top of other points may conceal information (Field 2013).

Although these issues are relatively obvious, lots of maps display such mistakes. An interesting example is pictured in Figure 1. The two maps are both displaying the same map excerpt of Manhattan. Blue points display tweets made by locals, red points are made by tourists. By changing the drawing order of the points, the maps convey different messages and lead to contradictory conclusions. (Field 2013)

Concluding the above mentioned critique points, it can be said that geolocated tweets represent only a small fraction of tweets posted only by a small fraction of users on a social media site that is used only by a fraction of internet users who also only account for a part of the population. For this reason, some researchers discourage the use of tweets as simple latitude/longitude pairs without proper pre-processing and deriving any results by focusing solely on the absolute number of Twitter messages.

Legal Framework

Working with Big Data might lead to some serious privacy concerns. This is especially true in case of social media analysis, as it handles potentially private information, not necessarily provided for research purposes.

Following the Cambridge Analytica scandal beginning in May 2017 (Cadwalladr 2017) but gaining worldwide attention in 2018 (Cadwalladr and Graham-Harrison 2018) concerns about misuse of personal data have been thermalized in media frequently. Over the course of the year, the portal Facebook was in addition affected by several more issues. In September it became public that due to a security breach, hackers gained access to personal data of 50 Million users worldwide (G. Rosen 2018).

In contrast to Facebook, Twitter has a more open and clear policy in sharing data with third parties. The portal's Terms of Service allow accessing, downloading and processing of the

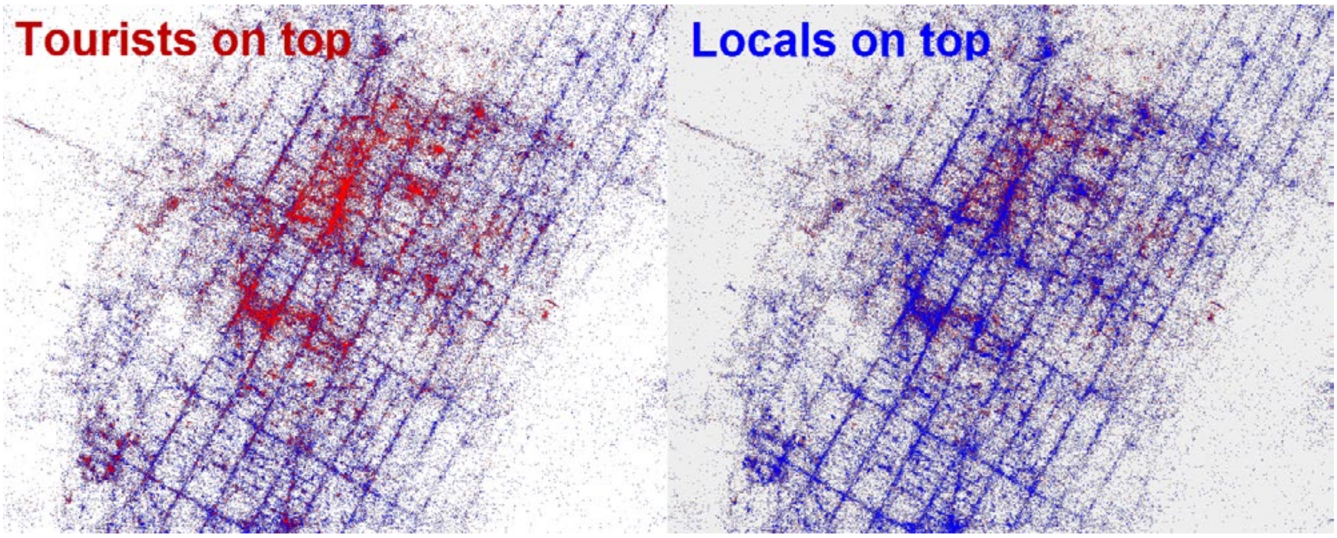


Figure 1: Effects of changing the drawing order of points on a map (Field 2013)

user's data. The conditions are set in the Developer Policy and the Developer Agreement. Latter guarantees

> *"a non-exclusive, royalty free, non-transferable, non-sublicensable, revocable license solely to:*
>
> 1. *Use the Twitter API to integrate Twitter Content into your Services or conduct analysis of such Twitter Content;*
> 2. *Copy a reasonable amount of and display the Twitter Content on and through your Services to End Users, as permitted by this Agreement;*
> 3. *Modify Twitter Content only to format it for display on your Services; and*
> 4. *Use and display Twitter Marks, solely to attribute Twitter's offerings as the source of the Twitter Content, as set forth herein."*
>
> Twitter Inc. 2018b

The terms of using geographic data is stated in paragraph II.C. of Twitter's Developer Agreement. Twitter only allows the usage of geographic data only in conjunction with the Twitter Content (tweet ID, user ID, content of the tweet, etc.) and not on the standalone basis (Twitter Inc. 2018b). In this work, geographic data is always linked to additional information and follows the preparation and visualisation approaches of the reference projects (especially as done in Crampton et al. 2013).

Furthermore, the agreement prohibits the transfer of downloaded data and derived information to any entities dealing with surveillance and tracking, or may violate human rights with access to the information (Twitter Inc. 2018b).

The Developer Policy describes the ways how Twitter content may be used and displayed. The most important points are that Tweets may not be modified and need to be displayed completely and correctly. Furthermore, it needs to be ensured that any kind of Twitter Content may not be associated with any person, household or other identifier (Twitter Inc. 2017).

It is important to note that only the official API may be used for getting information. The Terms of Service prohibit any kind of scraping or downloading the data with any other means apart from those provided by the portal (Twitter Inc. 2018a).

The Terms of Service and additional documents of the portal covers only the relation between Twitter and its users as a contract on a private law basis. For social media research also further legal documents need to be followed, most importantly the Regulation (EU) 2016/679, commonly known as "General Data Protection Regulation" or "GDPR".

Tweets contain information that belong to the definition of "personal data" of the regulation which is defined as

> *"any information relating to an identified or identifiable natural person […] who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier […]"*
>
> European Union 2016

Responsible for ensuring the compliance with principles of the GDPR is the data controller, as defined in Article 5.2 of the Regulation. In the case of Twitter, controller is the social media portal itself (Twitter Inc. 2018c).

It is also the responsibility of the controller to provide "sufficient guarantees to implement appropriate technical and organisational measures in such a manner that processing will meet the requirements of this Regulation and ensure the protection of the rights of the data subject" (Article 28.1) in case data processing is carried out on behalf of the controller.

This is also the case in this work (and generally in social media research) and although the data controller holds the most responsibility of dealing with personal data, data processors need also to meet certain requirements. It is the data processors responsibility to guarantee the safety of the data (in case of a data breach the controller has to be notified immediately), they may only use the data in accordance to the data controller's principles (Article 28.3).

The above points covered the legal basis of data processing but during the work with Twitter data further ethical questions may come up, which are discussed and reflected critically in the next sections.

2.12. Ethical Questions

The disregard of important ethical questions may result in serious consequences as shown for example in a case in 2016 where researchers made the profiles of 70.000 users of the dating site OkCupid publicly available on a data sharing website. The data was scraped through a script which scanned and downloaded personal information automatically. As the usernames were not excluded from the uploaded information and it contained highly personal details about the subjects' views on sensitive topics like politics or sexuality, it proposed a basic breach of research ethics and the privacy of users (Resnick 2016).

Justifying their actions, the researchers argued that the information was accessible to anyone with a profile nevertheless. Still the question remains whether in such a case it is possible to assume that the research subjects would have given their consent. OkCupid sees this case to be a "clear violation of [their] terms of service" and many researchers responded with a harsh critique to the authors of the study (Resnick 2016).

In this case, it becomes relatively obvious that the researchers clearly infringed the privacy of the users of the portal, as firstly information one would publish on a dating profile can be assumed to be private and secondly because the researchers had to actively circumvent the portals terms of service.

In the case of Twitter, these questions become harder to answer. The portal is used for posting public statements and messages and the platorm itself makes the data of their profiles publicly available for everyone without the need for using debatable data acquisition methods like website scraping.

However, even the implied assumptions of the question above are uncertain. As a general ethics guideline published by the university of Sheffield points out, there is a "lack of clarity whether an online space is public or private" (Ahmed et al. 2017, p. 86).

Though Twitter is generally considered to be a more open platform with users being usually aware that their information is publicly available (the default setting of a Twitter profile is to be public), it is of course disputable whether this assumption is true in all individual profiles and cases (Ahmed et al. 2017, p. 86).

Getting informed consent (as it is usual with studies dealing with personal data) from social media users is an extremely labour intensive process (Ahmed et al. 2017, p. 87). It is still important to verify the question whether informed consent of participants is needed.

In case of Twitter, it is plausible that it is a decision of the user whether to post privately or publicly (the Twitter Streaming API only catches tweets that are visible by the public) or even to post at all (Beninger, Fry, Jago, Lepps, et al. 2014 as cited in Ahmed et al. 2017).

In a survey, where users of various social media platforms were asked about their opinions about informed consent in social media rese-

Question	Answer	Explanation
Can the collected data be considered as public?	Yes	› Only data is collected that is publicly available (even for those without a Twitter profile) › Twitter can be considered as more open (in contrast to Facebook e.g.) (Ahmed et al. 2017, p. 85)
Is there a need for getting informed consent of the participants?	No	› The data can be considered to be public (see above) › The nature of the research doesn't deal with controversial topics, such as private opinions of specific users (although it is very likely to capture also controversial Tweets through the research approach, therefore it is important to guarantee data safety – see below) › More sensitive content types, such as photos or videos are not collected
Is anonymity of the research subjects guaranteed?	Yes	› To ensure non-traceability of the results, there will no tweets or user profiles be published. Only aggregated data will be shown. › During the analysis in some cases it has been shown that some users' tweets show up a spatial concentration, which can indicate the location of their homes or workplaces precisely. Therefore it will be abstained from visualising tweets as points (although many projects and websites exercise this approach). › Tweets for demonstration purposes were created by the author.
Is the downloaded data stored securely?	Yes	› The streaming code runs on a trustworthy (and password-protected) server by the British company "PythonAnywhere", the data is stored in a password-protected database. Therefore, accessing the data requires two passwords. › The demonstration website works with pre-processed data, it doesn't download any content directly from the database. In all of these files no user or tweet ids are stored. The website is hosted on the servers of the same company where the database runs. › Some processing of the data was done locally on a notebook. Any non-aggregated data will be deleted after the work is complete › The data stored in the database will be kept for one year after completion of this work.
Are the results of the study being used to generate profit?	No	› The data and the results of this work are only used for acquiring an academic degree.

Table 2: Ethical principles (own table)

Downloaded Data	Principles of processing/publishing
Tweet ID	Not shown/published in any case but used as key for linking further data/research results
Tweet text	Never published in full, only in some cases in a non-traceable, tokenized form
Tweet coordinates	Never published, not even on the map as coordinate pairs, only in an aggregated form
Tweet creation timestamp	Never published, used only to recognise temporal trends of Twitter usage
Tweet language	Never published, only in an aggregated form
User ID	As Tweet ID, only used for linking data but won't be published
Username*	Never published, no further processing
User description*	As above
User creation timestamp*	As above
Number of user's followers*	As above
User language*	As above
User location*	As above

Table 3: Data handling (own table)

arch the following nine cases/examples were listed where they found it to be necessary:

- *"It is morally and legally required*
- *To promote trust between the researcher and the participants*
- *When researchers are quoting a username alongside a social media post*
- *When a post is no longer recent, it was noted that it would be important to ascertain whether the participant still holds the same view*
- *When researchers seek to publish photographs or other images*
- *If a social media post is considered particularly sensitive and/or personal*
- *In order to ascertain whether a user intended to post publicly*
- *If the social media post would be used to generate a profit*
- *In order for users to determine both the quality and purpose of the research"*

Beninger et al. 2014 as cited
in Ahmed et al. 2017

The authors identified four factors that shape the attitude of the participants to this issue. These are the sensitivity, the mode and the content of a post, the privacy/openness of the platform, its users' expectations and the nature of the research, the affiliation of the researcher and their organisation and the purpose of the research (Ahmed et al. 2017, p. 90).

Another important issue is guaranteeing the security of the collected data. The collected information should only be accessible for the researcher and it needs to be ensured that published results don't contain private data, or access to private data. Even when only publishing the text of the tweet without any additional metadata about the user, it might be possible to identify the user's profile by searching the text in a search engine (Ahmed et al. 2017, p. 92).

Here, it is also important to note the contradiction with Twitter's Terms of Service which demands the Tweets to be published completely and with all information it originally contained. This, in contrast, might also lead to the publication of private information and would contradict the ethical guidelines presented above.

This project oriented closely to Ahmed et al.'s (2017) recommendations, with the approaches to dealing with the most important ethical questions presented Table 2.

Table 3 displays the principles of data handling throughout the project. Fields marked with an * were downloaded initially but as the containing information was not processed further on, these were removed from the database.

3. Methods

The previous chapter sketched some substantial discussions that emerged together with the increasing amount and volume of data we create. Big Data promises to grant insight into many new phenomena in a unique scale and depth. Still, a lot of data doesn't necessarily mean lot of information, let alone valuable information. As the computer scientist Mitch Kapor concluded, "Getting information off the internet is like taking a drink from a firehose".

The following section deals with the methodological questions in data access and preparation. As described in the chapters before, extracting knowledge from social media data is a complicated process. It lays in its size, form, and nature (see *2.2. What is Big Data? – p. 5*) that it cannot be processed with traditional methods. Critiques have pointed out that it is not enough to simply load the coordinates of a tweet into a GIS-application. The results of such an approach may easily become misleading, or completely wrong.

In this chapter, the steps of getting data and extracting information are presented. It begins with the ways Twitter grants access to its data collection, how data can be downloaded, prepared and how the most important pitfalls in social media analysis can be avoided or tackled.

3.1. Accessing the Data

As stated before, Twitter is known for granting a relatively extensive access to its servers. Users can deploy various tasks, such as posting and downloading content, updating profile information, interacting with other users, etc.

The access is enabled through the portal's "Application Programming Interface", or "API". An API can be defined as a "set of commands, functions, protocols, and objects that programmers can use to create software or interact with an external system" (Christensson 2016). It can be explained as a communication channel, where a programme can send a code with a specific query or request to an external system (for example another programme, or another computer). The external system responds with sending back data or performing an action (Ahmed et al. 2017, p. 83f). In the case of Twitter, it is possible for users to interact with the website through their code and perform actions (like reading or publishing tweets or interacting with other users) that can also be done through the portal's user interface (for example typing in a tweet into the field and clicking on publish).

Of course, for ordinary users the API does not provide much benefits, but it is a highly valuable asset for companies and organisations. Through the Twitter API they can interact with a large number of users at the same time, can track specific keywords (for example every tweet where a company's name is mentioned), control multiple accounts simultaneously, send automated messages, and carry out lots of further tasks.

For different tasks, Twitter offers six different APIs, these are presented in the list below (Twitter Inc. n.d.-d):

- Search API – This API provides the user all tweets (of the last one week for free accounts) matching a specific set of criteria. For example, it is possible to query all tweets sent in London that contain the word "urban".
- Account Activity API – Through this API, it is possible to control multiple accounts (up to 15 in the free version) with the possibility of posting tweets, retweeting statuses, sending and receiving direct messages.
- Twitter for Websites – This is probably the simplest API, it allows users to embed their Twitter timeline on their website.
- Direct Messaging API – As the name suggests, via this API it is possible to send and

receive direct messages from and to other users.

- Ads API – For managing and running advertisement campaigns on Twitter, it is possible to use the Ads API. It also provides insight into how many users could be reached with the campaign and how their reacted.
- Streaming API – This API allows users to capture and filter tweets in real time meeting a specific set of criteria, such as containing a word, a hashtag, created by a specific user, using a specific language, published in a given geographical region, or a combination of the above.

The data is not obtained directly through the user's account but through an application created on Twitter's application management portal[1]. For a long time setting up a developer's account and creating applications was possible for any user anytime. However, this policy was changed in July 2018, requiring an approval by the Twitter staff before updating the account (Roth and Johnson 2018).

For an approval, the applicants need to "provide detailed information about how they use or intend to use Twitter's APIs" in order to ensure compliance. Furthermore, changes in an application's API usage may be checked additionally (Roth and Johnson 2018). There were some further restrictions introduced, but they mostly concern active actions (such as posting tweets) or rate limits that this project's application doesn't concern.

Setting up an application is relatively simple, it consists of a name, a short description, a website and some more information (for example Privacy Policy) that may be required in case of interacting with users[2].

For accessing the APIs, the app generates four keys/tokens. These are needed to authenticate the user and the application before establishing a connection to Twitter's servers.

3.1.1 Streaming API

The Streaming API of the website downloads a sample of all public tweets in real time that meet a specified set of criteria. A possible case of application might be the tracking of all tweets that mention the word "London". Of course, it is possible to set up filtering conditions based on further features of a tweet, like username, language, location (and a combination of those). A list of Tweet attributes can be found on the Twitter's Developer's website[3] and in chapter *3.2. Anatomy of a Tweet – p. 27*.

To capture the tweets in a specific area, it is possible to define a bounding box over the study area (for example a city) to capture all tweets that fall within boundaries. The bounding box is defined by its southwest and northeast corners by their longitude/latitude pairs.

Whether a tweet falls into this bounding box is inferred from its coordinates if available. If not, and a place name is provided, it will be tested whether this place falls into the bounding box. In case the user didn't provide any of these information, it is not possible to check whether the tweet falls in the bounding box, thus it won't be returned (Twitter Inc. n.d.-e)[4].

It is important to note that a large number of tweets provide some kind of location information. By standard the place attribute of a tweet is populated, making the probability high to capture a high number of tweets that were submitted in the area covered by the bounding box.

A more significant limitation of the free version of the Streaming API lies in the fact

[1] https://apps.twitter.com/

[2] Due to changes in Twitter's development policy (as described above), the Application Management website will be closed and the app management transferred to the Developer Portal. Therefore, the information provided in the text may already be outdated.

[3] https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object

[4] Twitter's handling of geographic data is somewhat complicated, sometimes contradictory. The main issues to consider are described in *3.5.1 Location/Place/Coordinates – p. 40*

that it returns just a sample of the requested tweets. Twitter doesn't provide an exact size of the sample (or further information how the sampling is conducted) but most studies estimate its size to be around 1-3 per cent (García-Palomares et al. 2018, p. 311; Lansley and Longley 2016, p. 86; Morstatter et al. 2013, p. 400) of the tweets.

For sampling 10 per cent (using the "Decahose API") or 100 per cent of the tweets ("PowerTrack API" with access to the "Firehose"-stream), a Business or an Enterprise subscription is necessary, respectively (Twitter Inc. n.d.-f). Twitter doesn't disclose the price of these two accounts, they are set for each case individually.

For a comparison, the premium package of the search API pricing ranges from 99 $ per month (for 100 search requests) to 1 899 $ per month (for 2 500 search requests). The free version enables 50 requests per month and delivers only results for the last 7 days (Twitter Inc. n.d.-g).

When planning/conducting a research project, it is an important question to ask whether it is necessary to pay for a premium access to the complete dataset or does the complimentary access still provide data with a quality high enough to be used in a project. Morstatter, Pfeffer, Liu and Carley (2013) compared the sampled data obtained through the Streaming API with the full data of the Firehose stream.

In the time period from December 14th, 2011 to January 10th, 2012 they downloaded 1 280 344 tweets from the Firehose and 528 592 from the Streaming API. The data was analysed on their content (evaluated through the comparison of hashtags and through topic modelling), their network measures (by comparing connections between users and users and users and hashtags) and their geographic distribution (through the coordinates of georeferenced tweets) (Morstatter et al. 2013).

In general, the results have shown that the coverage of the Streaming API depends strongly on the concrete parameters of the query sent to the API. The more tweets match the given parameters, the less data will be sent to the user (therefore it is useful to specify exact and accurate filters that match the research question very well) (Morstatter et al. 2013, p. 406).

The comparison of the two data sets indicate some bias "in the way that the Streaming API provides data to the user". This means, the sampled data of the Streaming API is likely to be not completely random but filtered by Twitter itself using undisclosed criteria (Morstatter et al. 2013, p. 407).

Interestingly, and most importantly for this work, in case a boundary box is used for sampling, the Streaming API "almost returns the complete set of the geotagged tweets despite sampling" (Morstatter et al. 2013, p. 407).

Unfortunately, there is no more recent study available that could provide updated insight into the representativeness of the Streaming API data. A 2014 study with a similar question (Joseph, Landwehr, and Carley 2014) found similar results, namely that data provided via the Streaming API is probably not random but filtered by Twitter using probably some specific criteria. As in the other study, the researchers here also couldn't come up with any understanding about the functioning of the sampling method.

Nevertheless, in most of recent practical studies (for example García-Palomares et al. 2018; Lansley and Longley 2016; Rzeszewski and Beluch 2017; Sloan 2017) the free Streaming API is used, only one (Leetaru et al. 2013) works with data obtained through Twitter Decahose (which samples, as mentioned above, only 10 per cent of the requested data).

Because the Streaming API sends the requests in real time (and not in batches), a persistent in-

Field	Format	Description	Example
Tweet ID	Int64 (because of its length, it can also be requested as a string)	A unique identifier of the tweet	1005528230433050000
User ID	Int64 (because of its length, it can also be requested as a string)	A unique identifier of the user that posted the tweet	974297894223319000
Text	String	The message of the tweet	Hello World!
Creation date	String	Creation date of the tweet	6/9/2018 19:13
Language	String	ISO 639-2 code of the language. It is detected by Twitter automatically. The value is “und” if it’s undetected.	de

Coordinates	geoJSON point	A geoJSON object with a latitude/longitude coordinate pair of the location where the tweet was created	[..]“coordinates“:
			{
			„coordinates“:
			[
			-18.14310264,
			16.05701649
			],
			„type“:“Point“
			}

Place	geoJSON polygon	A geoJSON object containing the bounding box object of the place (for example a city) provided by the user	Too large to display but includes: bounding box coordinates › id › name › name of city/country › place type (e.g. “city”) › ... › URL to the JSON file stored on the server

Table 4: Tweet object (own table based on Twitter Inc. n.d.-c)

Field	Format	Description	Example
User ID	Int64 (because of its length, it can also be requested as a string)	A unique identifier of the user	974297894223319000
Username	String	Displayed name of the user	Balázs Cserpes
Screen Name	String	Name identifier of the user (the part that comes after the @)	Bal_Cse
Followers Count	Int	Number of followers	781133
User language	String	ISO 639-1 code (with an optional ISO 3166 sub-code[1]) of the language used by the user on the Twitter interface (not necessarily the same language used in the tweets)	en_gb
User description	String	A short self-description of the user	Currently working on a master’s thesis about the usage of Twitter data in spatial planning
User location	String	A location that can be selected from a list of cities, regions and countries but also provided freely (also fictious names are allowed)	Vienna, Austria Airstrip One, Oceania

Table 5: User object (own table based on Twitter Inc. n.d.-b)

ternet connection of the recipient's computer is necessary. If the connection disrupts, the query needs to be restarted by the user. The connection will also break in case the recipient is not able to process the incoming tweets with the speed they are provided by the portal (Twitter Inc. n.d.-h).

Therefore, it is useful to specify exact filter parameters beforehand to exclude unnecessary data. In addition, further processing (for example sentiment analysis) should be done after the dataset was saved in a database.

3.2. Anatomy of a Tweet

Tweets returned by the API are encoded in JavaScript Object Notation (JSON) format consisting of up to 150 key-value pairs. A tweet object always contains a Tweet ID, a User ID, a message, a timestamp of the creation date, and a language that has been detected automatically by Twitter. A user object consists of a User ID, a username, and a number of followers (Twitter Inc. n.d.-i).

The most important attributes are presented in Table 4 for the tweet objects and in Table 5 for the user objects.

3.3. Data Preparation

Opening up a tweet, humans can normally identify in seconds what the text is about and also guess the sentiment of the content, i.e. if it carries a more positive or a more negative opinion about a subject. We recognise patterns based on our expectations, knowledge and experiences. We also attribute certain features to certain keywords or to certain people. A message sent from a friend might have a completely different meaning for us compared to the same text if it would appear in a book or on a billboard. We can read and follow books hundreds of pages long, still we would be certainly overwhelmed in case we would have to go through the hundreds, thousands or millions of tweets that are generated in a city.

The language we use has evolved through the thousands of years. The way we speak is under constant change, some words appear, others disappear, or their change in their meaning. But not only the words, also the structure and grammar of languages is not static. The rules a language had a hundred years ago, might be outdated as of today. Languages that have evolved naturally are described in linguistics as "natural languages" (Bird, Klein, and Loper 2009, p. ix).

In contrast, computers use languages with a predefined set of rules and conditions. It doesn't make a difference if a code input is given on one computer or on another, if both systems understand the language, the result will be same. This group of languages is called "artificial languages" (Bird et al. 2009, p. ix).

The translation between these languages constitute an enormous challenge. As a compiled (thus quickly readable by computers) programme code is hardly understandable by humans, for machines it constitutes a complicated task to understand the natural language we use, with its lax and sometimes contradictory rules, and the sometimes hardly comprehensible meanings behind its elements.

Methods to overcome this barrier in order to perform language processing more quickly and efficiently with computers appeared surprisingly early in our technological history. Computational linguistics date back to the 1940s when the interest came up to translate languages by machines (Hays 1967a).

Of course, the first tasks performed by computational linguistics were quite simple. They consisted of structuring and ordering inputs in computers and storing them efficiently. Still, complex problems soon became themes computer linguists worked on.

In a 1967 book, Hays listed three degrees of computer participation in linguistic research:

- *"The lowest level uses computer merely as a compiler of data"*
- *"The middle degree is computer testing of information gained in other ways"*
- *"The highest degree is reached when the computer program actually embodies the linguist's analytic ideas"*

Hays 1967b, p. 180

Reaching the last level would mean the possibility of computer programs learning an unknown language completely without supervision and participating in computer research themselves (Hays 1967b, p. 180). It might be interesting to refer to Anderson's (2008) article. The highest degree of computer participation is closely in line with his views of letting computers do research with just some basic human input.

Today, tasks dealing with linguistic tasks on computers are summarised under the term "Natural Language Processing", or "NLP". By definition NLP:

> *"is a subfield of linguistics and artificial intelligence (AI). It studies the problems inherent to the processing and manipulation of natural language (NL). The ultimate goal of NLP is to make computers 'understand' statements written in human languages."*

Linckels and Meinel 2011, p. 61

Without diving too much into detail, the next section defines and describes the most important terms that come up when working with NLP.

3.3.1 Corpus

The Oxford Handbook of Computational Linguistics defines the corpus[5] as a "body of linguistic data, usually naturally occurring data in machine readable form, especially one that has been gathered according so some principled sampling method" (Mitkov 2004, p. 734).

It forms the basis for corpus linguistics, where linguistic knowledge is extracted through the computational analysis of large amounts of texts. Corpus linguistics tries to identify patterns and find rules in a large amount of textual data, collected and stored in different text corpora (Hunston 2006).

The first systematic collection of English language texts was published in 1964 and contains 1 014 312 words of texts issued in 1961 (Francis and Kucera 1979). The corpus called officially "Standard Sample of Present-Day American English" but is most commonly referred to as "Brown corpus", was compiled by W. Nelson Francis and Henry Kucera at the Brown University. The contents were collected from 500 different texts and classified in 15 different categories (for example Press Reportages, Religion, Science Fiction, Humour ...) (Bird et al. 2009, p. 42f; Kholkovskaia 2017).

Today's corpora are much larger, the amount of texts and data they contain exceeds the Brown corpus multiple times. For example Google created a corpus from all the texts of Google Books, containing 155 billion words in American English, 34 billion in British English, and 45 billion in Spanish (Davies n.d.).

These astonishing sizes allow us to study the use of language in many contexts and capture how language was changing throughout the last decades and centuries. An interesting implementation of this corpora is Google's N-Gram Viewer where the appearances of specific words and word sequences can be tracked from 1800 to 2008[6].

The single elements of a corpus are called documents, these are for example the text passages or books that constitute the collection. In case

[5] lat. body

[6] https://books.google.com/ngrams

of this project, the single tweets are the documents that form the tweet collection as the corpus.

3.3.2 Text Segmentation/Tokenisation

While humans recognise the structure of a text and the logic between single elements (what a word means or what "?" signalises) of a text quickly, for a computer, the same text is just a sequence of characters. Machines need to break down the documents into single linguistic units (words, punctuation, numbers …). In NLP the process dealing with this task is called "tokenisation" (Mikheev 2004, p. 201).

In many cases, the identification of tokens is an easy task. In segmented languages (like English or German), the words are split by a blank space as a word divider, so tokens can be recognised as the segments between the spaces. However, the process becomes quickly complex when dealing with punctuations, abbreviations or other exceptions. As the end of the sentence, a full stop is a single token itself. In case of an abbreviation, it is part of the token and cannot be treated distinctively (Mikheev 2004, p. 203f). Twitter texts are especially difficult to tokenise, as the limitation in message length leads to an increased use of abbreviations and the general language used on the portal incorporates characteristic elements such as internet slang (for example "LOL") or emoticons (for example ":D").

For such cases the tokenisation algorithm works with comparing the text with entries in predefined lexica. Some algorithms are capable of looking ahead and looking back, as well as applying different sets of rules at the same time and choosing the best matching algorithm at the end (Mikheev 2004, p. 204f).

3.3.3 Morphology

For a machine, a word is just a sequence of characters, thus every derivation from this basic sequence may constitute a new word. "Run" and "Runs" have the same meaning and differ in just one letter. Still, expressed in a formal language, run != runs. In cases such as topic modelling, where the results are derived from the frequencies of words, such inaccuracies may lead to imprecise and distorted results.

Through the evolution of natural languages, these different formations have developed through a systematic way, with different forms of a word carrying different meanings or playing a different functions. Linguistic morphology deals with words and with the processes behind their formation (Trost 2004, p. 25f).

Figure 2 displays the morphology of the word "decompose". The basic building blocks of a word, called "morphemes" are displayed in the upper row. These morphemes express either semantic concepts ("compose" in this case), these are called roots, others signalise abstract features, like the plural form, or the tense (Trost 2004, p. 26).

Through the combination and adjustment of these morphemes (as parts of the words they are called morphs) a word is created. The part that carries the meaning is the base form, the variations (including the base) of this form are called paradigms (Trost 2004, p. 26).

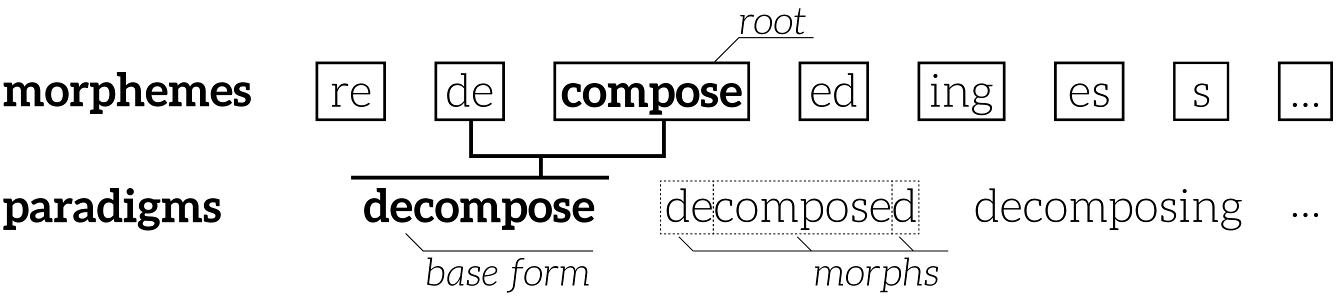


Figure 2: Text morphology (own figure based on Trost 2004, p. 26)

To find the base of a word, NLP uses two techniques, stemming or lemmatisation.

Stemming means the automatic transformation of words into their stems. It is done by an algorithm that identifies a stem of a word, which is not necessarily its base, nor a form that could be found in a lexicon (Trost 2004, p. 37f). For example the stemmed version of "knowledge" is "knowledg", of "only" "onli" and of "considering" "consid". The advantages of stemming lie in its robustness and its speed. Furthermore, when creating an algorithm, it requires less training data and is easier applicable for languages that don't have corpora in that quantities and quality as bigger languages (Manning, Raghavan, and Schütze 2009, p. 32ff).

Lemmatisation is a more sophisticated approach, it tries to find a "corresponding dictionary form for a given input word" (Trost 2004, p. 38). However, this process needs more computing resources than stemming and its availability is limited to some larger languages.

For a comparison, Table 6 displays a stemmed[7] and a lemmatised[8] output of a given sentence.

Though the table above shows that the stemmed versions of the sentences yield results that are closer to the words' roots, sometimes the algorithm relates "forms that are not morphologically related", causing overstemming. In contrast, when the stemmer fails to identify common roots, is called understemming (Tzoukermann, Klavans, and Strzalkowski 2004, p. 532).

To avoid such cases and use a more reliable approach in the case study, the word roots were defined through lemmatisation, applying the "WordNet Lemmatizer" of the NLTK Python-library.

3.3.4 Part-of-Speech Tagging

To understand a message behind a text it is necessary to understand the functions of the single elements in a document. The different roles an item (in this case a word) may have were recognised as early as 100 BC when Dionysius Thrax categorised words the first time into eight categories[9] (Voutilainen 2004, p. 220).

In NLP the identification of these syntactical functions happens through Part-of-Speech (POS) Tagging. POS tagging dates back to the 70s where first sets of rules were applied on texts, with identifying and correcting errors manually. The challenge of POS lies at recognising the correct tag in cases where a word has multiple meanings and can play different roles (Voutilainen 2004, p. 226). As an example, tagger needs to identify whether the word "play" is used as a verb or as a noun.

Rule based approaches apply certain rules that have been provided beforehand, or identified during the model's training. The possible function of an element may for instance be derived

[7] Using the "SnowBallStemmer" algorithm of the NLTK Python-library

[8] Using the "WordNet Lemmatizer" algorithm of the NLTK Python-library

[9] nouns, verbs, participles, articles, pronouns, prepositions, adverbs, and conjunctions

Input	Stemmed version	Lemmatised version
London is the best city I've ever been to	london is the best citi i'v ever been to	London is the best city I've ever been to
Sunset at the docklands #streetphotography	sunset at the dockland streetphotographi	Sunset at the docklands streetphotography
Finally, the weekend can begin boys and girls!	final the weekend can begin boy and girl	Finally the weekend can begin boy and girl
My cat eats 2 kilos of food every week. I think, they sold me a tiger	my cat eat 2 kilo of food everi week i think they sold me a tiger	My cat eats 2 kilo of food every week I think they sold me a tiger

Table 6: Stemming and lemmatising of texts (own table)

from its position. For example, there is no case where a verb follows an article (in case of "the play" it is clear that the second word is a noun). Although these models can work with a very high accuracy, modern POS taggers make use of more sophisticated machine learning methods and apply a more probabilistic approach (Voutilainen 2004, p. 227).

Briefly, a probabilistic approach derives the tags from a probabilistic sequence of random variables, which are, in this case, the tags. This means, if the tagger identifies a sequence of for example "article – noun – verb – preposition" it should guess what type the next tag may have.

The most commonly applied approach is the application of the Hidden Markov Model. This model helps to predict the probability of an outcome of a POS-tag based on the previous word. In simple terms a Markov Model tries to predict the n-th outcome of a variable based on its current state (Samuelsson 2004, p. 364ff).

Figure 3 demonstrates the basic functioning of POS-tagging based on the Hidden Markov Model. The sentence "Children want to play" is the observable part of the structure. For the first three words, the tags are clearly identifiable. However, the last word ("play") could function both as a verb, as well as a noun in the sentence. Here, the model tries to infer the tag of the word based the hidden structure (Markov Chain), in this case the probabilities of a given word sequence (Samuelsson 2004, p. 366). In this case the model would look at the probability of the sequences Noun-Verb-Preposition-Verb and Noun-Verb-Preposition-Noun. Because the first sequence is assumed to appear with a higher probability, the POS tagger would in this case identify "play" as a verb.

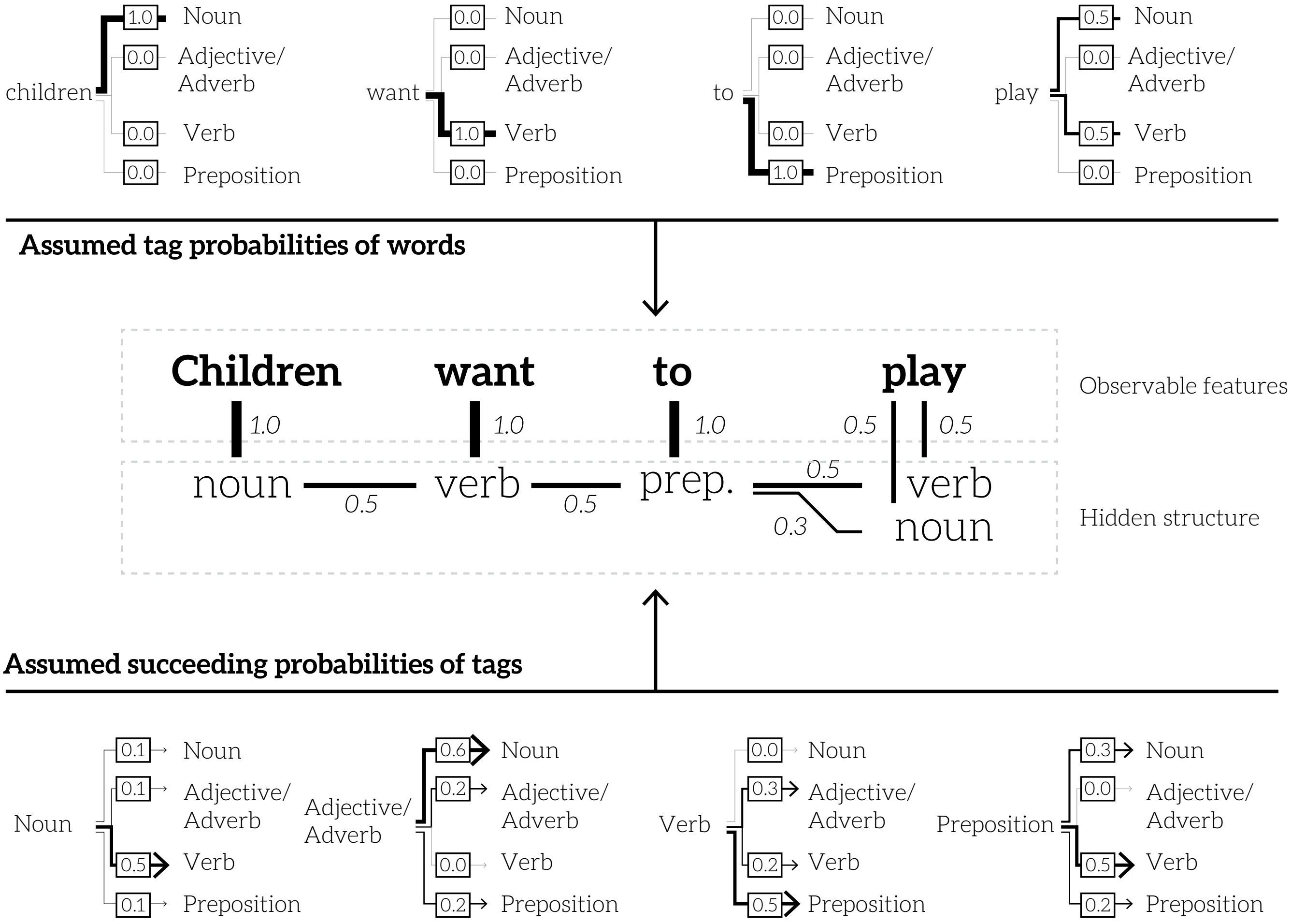


Figure 3: Hidden Markov Chain (own figure)

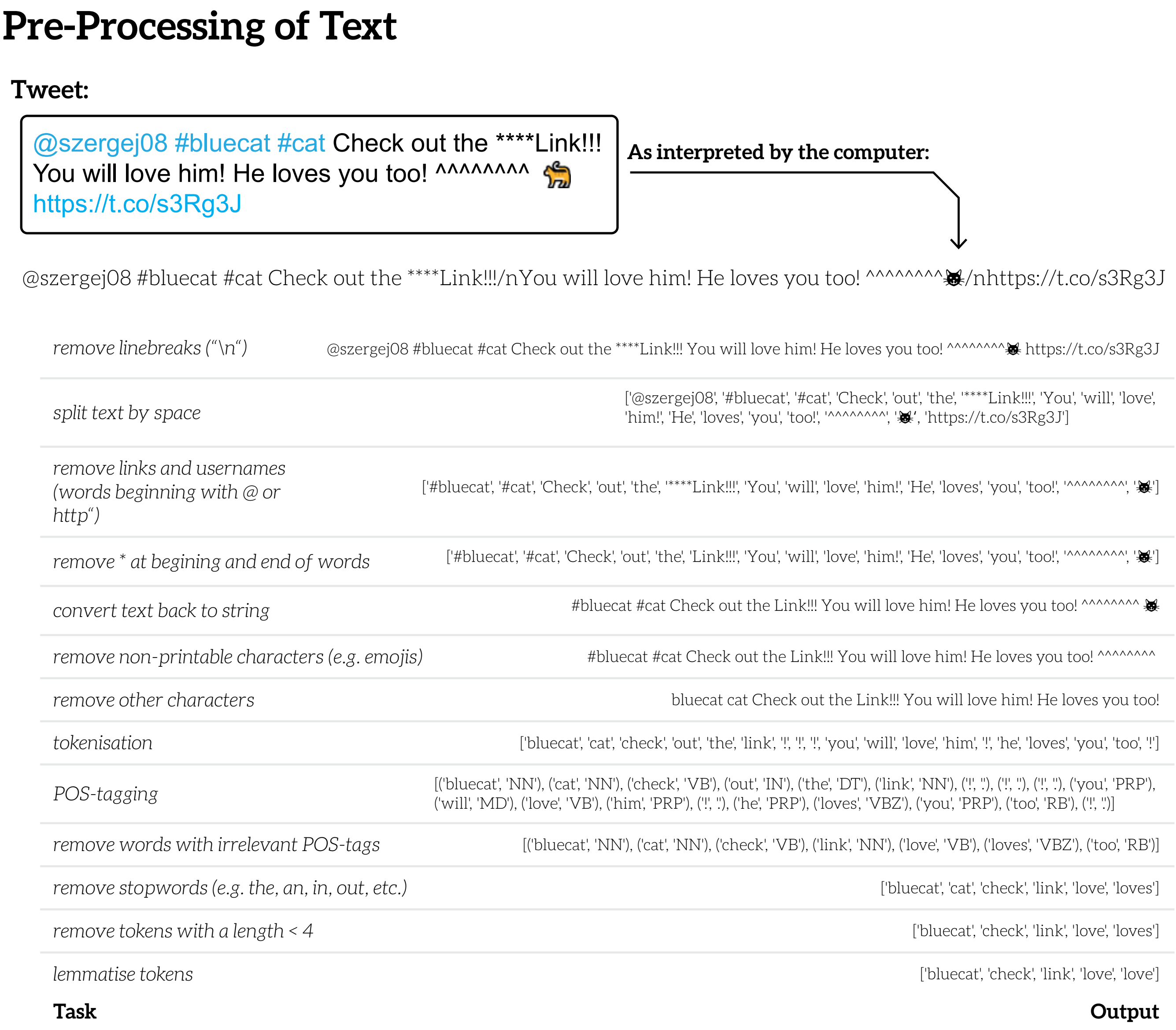
Of course, in many cases the underlying model is a multiple time more complex, containing a much higher number of tags, much less obvious assumptions, and cases where the words themselves don't have any tag probabilities which ale solely inferred based on their position and the tags of the previous and next words.

3.3.5 Further Steps

The pre-processing of social media texts also involves the removal of stopwords. "Stopwords are frequently occurring and insignificant words in a language that help construct sentences but do not represent any content of the documents". Mostly these are articles, conjunctions, prepositions, and similar elements, for example "a, on, the, this, etc." (Bing 2011, p. 227).

The easiest way to identify and remove stopwords is through looking them up in a predefined list. This approach can also be utilised via the Python library NLTK which also includes a list of stopwords.

As Twitter is characterised by a very informal language, a number of messages carry unconventional sequences of characters. For example some tweets begin with multiple "*"-s. Such characters need to be removed because the tokeniser identifies them as parts of the word (for



Pre-Processing of Text

Tweet:

@szergej08 #bluecat #cat Check out the ****Link!!!
You will love him! He loves you too! ^^^^^^^^ 🐈
https://t.co/s3Rg3J

As interpreted by the computer:

@szergej08 #bluecat #cat Check out the ****Link!!!/nYou will love him! He loves you too! ^^^^^^^^🐱/nhttps://t.co/s3Rg3J

Task	**Output**
remove linebreaks ("\n")	@szergej08 #bluecat #cat Check out the ****Link!!! You will love him! He loves you too! ^^^^^^^^🐱 https://t.co/s3Rg3J
split text by space	['@szergej08', '#bluecat', '#cat', 'Check', 'out', 'the', '****Link!!!', 'You', 'will', 'love', 'him!', 'He', 'loves', 'you', 'too!', '^^^^^^^^', '🐱', 'https://t.co/s3Rg3J']
remove links and usernames (words beginning with @ or http")	['#bluecat', '#cat', 'Check', 'out', 'the', '****Link!!!', 'You', 'will', 'love', 'him!', 'He', 'loves', 'you', 'too!', '^^^^^^^^', '🐱']
*remove * at begining and end of words*	['#bluecat', '#cat', 'Check', 'out', 'the', 'Link!!!', 'You', 'will', 'love', 'him!', 'He', 'loves', 'you', 'too!', '^^^^^^^^', '🐱']
convert text back to string	#bluecat #cat Check out the Link!!! You will love him! He loves you too! ^^^^^^^^ 🐱
remove non-printable characters (e.g. emojis)	#bluecat #cat Check out the Link!!! You will love him! He loves you too! ^^^^^^^^
remove other characters	bluecat cat Check out the Link!!! You will love him! He loves you too!
tokenisation	['bluecat', 'cat', 'check', 'out', 'the', 'link', '!', '!', '!', 'you', 'will', 'love', 'him', '!', 'he', 'loves', 'you', 'too', '!']
POS-tagging	[('bluecat', 'NN'), ('cat', 'NN'), ('check', 'VB'), ('out', 'IN'), ('the', 'DT'), ('link', 'NN'), ('!', '.'), ('!', '.'), ('!', '.'), ('you', 'PRP'), ('will', 'MD'), ('love', 'VB'), ('him', 'PRP'), ('!', '.'), ('he', 'PRP'), ('loves', 'VBZ'), ('you', 'PRP'), ('too', 'RB'), ('!', '.')]
remove words with irrelevant POS-tags	[('bluecat', 'NN'), ('cat', 'NN'), ('check', 'VB'), ('link', 'NN'), ('love', 'VB'), ('loves', 'VBZ'), ('too', 'RB')]
remove stopwords (e.g. the, an, in, out, etc.)	['bluecat', 'cat', 'check', 'link', 'love', 'loves']
remove tokens with a length < 4	['bluecat', 'check', 'link', 'love', 'loves']
lemmatise tokens	['bluecat', 'check', 'link', 'love', 'love']

Figure 4: Document pre-processing (own figure)

example “**London**” is treated in this form instead of converting it to “london”).

As mentioned above, Twitter uses two characters identifying users and trending topics. For text processing, usernames should be removed, as they can be misinterpreted as nouns. The easiest way to do this is by identifying and deleting all words that begin with a “@”.

The same identification and removal procedure is to be carried out with hyperlinks (beginning with “http”).

Hashtags carry by definition an important message and may form a substantial part of a tweet. Therefore, during pre-processing only the “#” should be removed and the remaining word kept.

Another common element of tweets are emojis. These can be removed through comparing them with a predefined list of possible characters. The string module of python has a set of printable characters that includes digits, ASCII letters, punctuation and whitespace (Python Software Foundation 2018b).

A line break in a Tweet is decoded as the character sequence “\n” and needs to be removed because it may also be interpreted as a token by the computer.

As a last step, all characters need to be set to lowercase.

The complete procedure of tokenisation and pre-processing is displayed in Figure 4.

3.4. Getting the Message

The steps described above had the aim to make the texts machine-readable, and thus interpretable for tasks which involve the identification of specific patterns, such as topic modelling. On the next pages, the core concepts of two semantic text analysis approaches (sentiment and topic identification) are sketched. The descriptions include a brief presentation of the algorithms involved and describe the methods best suitable when dealing with short Twitter texts.

3.4.1 Sentiment Analysis

One of the most valuable assets of social media data is its underlying information about the opinions, views and sentiment of people towards specific topics. Sentiment analysis (also called opinion mining) helps to get to the core of this information and promises to give answers to such questions. In spatial planning, sentiment analysis can help to identify fear inducing public spaces, and also provide insight into the perception of certain locations during different times.

By definition, the aim of sentiment analysis is “to define automatic tools able to extract subjective information in order to create structured and actionable knowledge.” (Pozzi, Fersini, Messina, and Liu 2016, p. 1)

The synonymous term “opinion mining“ may lead to confusions about the focus of sentiment analysis as it implies a different subject of analysis (Pozzi et al. 2016, p. 1). The (simplified) distinction between opinion and sentiment lies in the fact that the first one describes an attitude determined by mind whereas the other is determined by feeling.

Furthermore, an opinion does not necessarily imply a (subjective) value judgement about the statement (as for example in the sentence “I think the governing party will be defeated in the upcoming elections”). In contrast, sentiment (“I’d be happy if the government would be defeated”) is an attitude towards a subject and can be most easily classified as “positive”, “negative” or “neutral” (Pozzi et al. 2016, p. 1).

Formally, an opinion can be defined as a quintuple of a specific entity (ei), an aspect of this entity (aij), the sentiment of this aspect (sijkl) by an opinion holder (hk) expressed at a certain time (tl) (Liu 2012, as cited in Pozzi et al. 2016, p. 1).

In case of a spatial social media analysis, this quintuple can also be enhanced by the aspect of location (li).

The sentiment aspect can be described by its positivity, negativity or neutrality, or with "different strength/intensity levels", such as awarding points, stars or other scores (Pozzi et al. 2016, p. 2).

Regarding levels of analysis, Pozzi et al. identify three levels of granulation/scope. Message level works with sentiment of a document as a whole (for example a product review). Sentence level analysis deals single sentences, while entity and aspect level analysis deals with extracting sentiments directed to a single entity (for example in case of a product review the battery of a phone) (Pozzi et al. 2016, p. 6f).

In most cases, sentiment scores are derived from opinion words, the elements of a given text that indicate a certain polarity score of the sequence. In case of "London is an amazing city", amazing would be the opinion word (Luo, Chen, Xu, and Zhou 2013, p. 53).

The identification and score definition of such opinion words can be carried out through different basic approaches. The simplest but most labour intensive process is the manual collection of opinion words. In practice, the most common methods used are either using and expanding a seed list, or deriving opinion words and scores from the "syntactic or co-occurrence patterns in large text corpora." (Luo et al. 2013, p. 55)

A significant advantage of the lexicon-based approach is its robustness. Once a collection of word-sentiment pairs is set[10], the scores can simply be accessed and connected to the words of the analysed text. The word-sentiment pairs are enhanced by finding synonyms and antonyms and adding the respective sentiment scores. Then, the process is repeated with the new words until all opinion words have been found and defined. For correction, the scores can be inspected manually (Luo et al. 2013, p. 55f).

The above approach is simple and can lead efficiently to good results. However, a shortcoming comes from the fact that it isn't able to identify the context-related sentiment of a word. For example, the word "long" signalises two completely different sentiments in the sentences "we had long and exhausting bus ride here" and "I'm really glad that I am able to spent a long time abroad" (Luo et al. 2013, p. 56).

Incorporating context-related sentiment scores can be done through a corpus-based word generation. This approach starts, as above, with predefined word-sentiment pairs. Then, through POS tagging it is possible to find conjoined adjectives and adverbs. Their scores are defined through the respective conjunction words[11]. In the end, through clustering, the sets of opinion words (for example positive and negative) can be produced (Luo et al. 2013, p. 56f).

An even more sophisticated approach is applied through identifying the respective sentiment scores to individual aspects of an entity[12]. This approach yields the most accurate results as a document often includes different sentiments to different entities. Therefore, depending on the specific cases, a document-level score definition might lead to too generic results (Luo et al. 2013, p. 57).

The identification of the overall sentiment of a document can be done simply by just counting opinion words (with their respective scores), although this approach may be not accurate enough, as it fails for example at identifying negation ("not beautiful" is counted as positive if "not" is not identified) words. To overcome this problem, a set of rules can be defined, for ex-

[10] In most cases, it is possible to revert to already existing word-sentiment lexical resources, for example SentiWordNet or LIWC

[11] For example in "ugly and simple" the second adjective carries probably a more negative sentiment than in "ugly but simple".

[12] "object attributes" in the referred text

ample by finding these negation words (Luo et al. 2013, p. 57f).

Of course, the effectiveness of manually finding such patterns is limited. For more robust results, the identification of these patterns can be done by supervised machine learning, where known examples are used to find patterns in new documents of the same domain (Luo et al. 2013, p. 59).

3.4.2 Sentiment analysis of social media texts

Sentiment analysis of social media content includes some specific challenges and complexities. Besides the general difficulties (see *2.11. Critique – p. 16*), as lack of validity, unstructuredness, etc. further particularities exist that make sentiment analysis of this content more complex.

First, the language used on social media sites changes rapidly. Texts on these portals use a specific, informal, and personal language that is evolving constantly and being adapted by its users. Therefore, sentiment analysis needs to apply to these changes and recognise the correct and up-to-date sentiment of a given message (Pozzi et al. 2016, p. 5).

Furthermore, the relational aspect of a social network may become relevant when conducting sentiment analysis. The creators of social media content are dynamically connected to each other with a significant amount of information lying in these networks. This has also impacts on the ways how to inspect, summarise, and visualise these opinions (Pozzi et al. 2016, p. 5).

Social media texts are also characterised by the use of a number of specific features that are normally not listed in current word-sentiment-lexica ("WTF" is used to signalise a negative sentiment) or emoticons (:D). Furthermore, the strength of a sentiment is also expressed by intentional deviations from a grammatically correct language. For example both CAPITALISATION and multiple use of punctuation marks (!!!!!) signalise a stronger emphasis of the message (Balahur 2013).

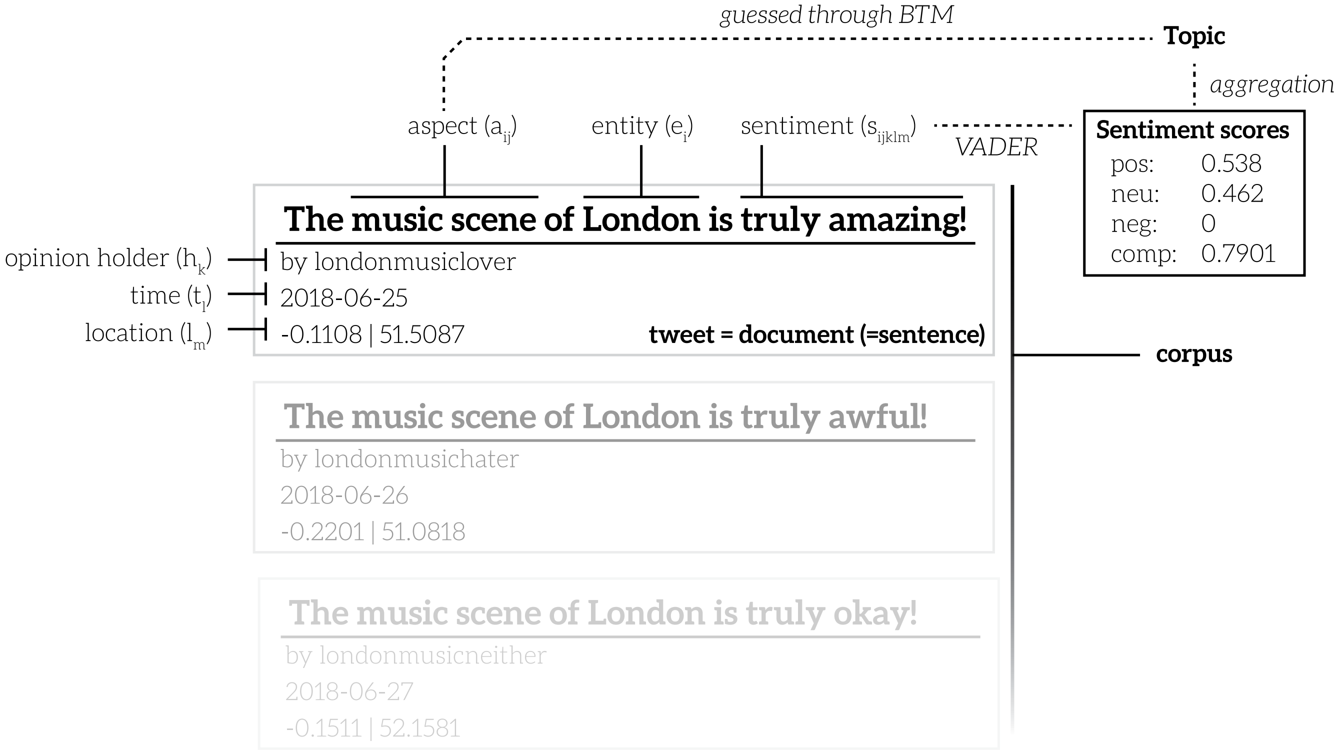


Figure 5: Sentiment analysis (own figure)

Another challenge in polarity classification of social media texts lies in its lack of context. The limited length of a tweet makes it very hard to contextualise the message. Furthermore, communication on the internet is often characterised by the use of sarcasm or irony, which both are hard to detect by machines (Farias and Rosso 2016, p. 114).

In case of Twitter (and therefore in this work) the levels of analysis flow into each other and are consequently hard to identify. All collected messages constitute the corpus, which is an aggregation of documents, in this case, the tweets. These are in most cases not longer than one or two sentences, making the document and sentence levels mostly the same. With the assumption that every tweet deals with one topic, the entity level analysis was done through identifying these topics and bringing together the single document-level sentiment results.

Further focus levels were placed on the aspects place and time, as well the combination of these factors. The sentiment was expressed by the positivity, negativity and neutrality score, as well as a general compound indicator, incorporating these three assets. The scheme of the analysis is shown in Figure 5, the methodology of gathering the results is described in the following section.

3.4.2.1. VADER

To deal with the above described specific characteristics and constraints of social media text, Hutto and Gilbert (2014) created a text classifier specifically for social media texts, called VADER (Valence[13] Aware Dictionary and sEntiment Reasoner).

The aim of VADER was the creation of a rule-based sentiment model, being capable of taking the context of a message into account, understands special features of social media content, is fast enough to use it in real time, requires no training data and still "does not severely suffer from a speed-performance tradeoff" (Hutto and Gilbert 2014, p. 4).

The creation of the model was done through a human assisted supervised machine learning approach. The developers started with extracting already existing scores from different sentiment lexica. The features were supplemented with further elements that are characteristic for social media texts. Through the process, the individual steps were regularly validated by human examiners. The quality of the model was compared to gold-standard[14] sentiment lexica dealing with different domains (social media text, product reviews, movie reviews, NY Times editorials), and could outperform all other lexica when working with social media text (Hutto and Gilbert 2014, p. 7f).

When analysing a given input, VADER infers the sentiment using two types of indicators. It defines multidimensional ratios of proportions (adding up to 1) of the text to fall into one of the categories positive, negative, or neutral (Hutto 2018). This means in practice, the sentence "I love London" has the scores 0.808 – 0.192 – 0.

[13] "The number and kinds of words and phrases a word can combine with in regular patterns" (Mitkov 2004, p. 760)

[14] Validated by humans

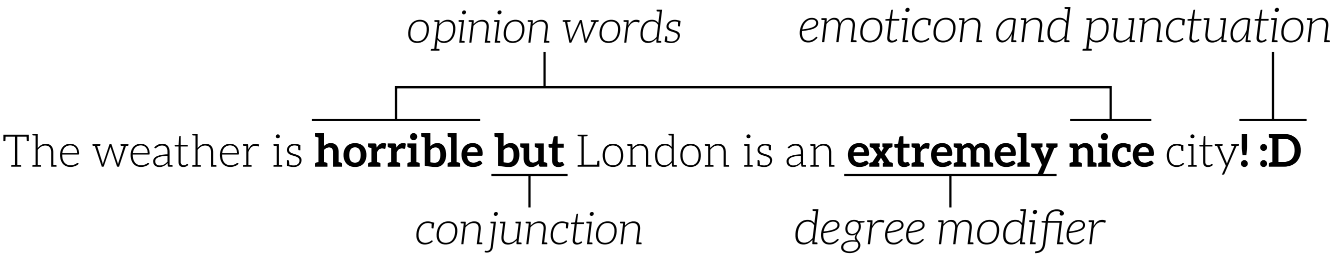

opinion words
emoticon and punctuation
The weather is **horrible but** London is an **extremely nice** city**! :D**
conjunction
degree modifier

Figure 6: Factors considered by VADER (own figure based on Hutto and Gilbert 2014, p. 6f)

This means, the sentence carries to over 80 per cent a positive, and to 19 per cent a neutral sentiment.

The unidimensional compound score is defined as a “normalized, weighted composite score”. This means, the sentiment is calculated by summing the valence scores of each word while adjusting their scores to the predefined rules of the algorithm (for example by giving a capitalised word more weight). Then, the score is normalised to be between -1 and +1, with -1 indicating a completely negative and +1 a completely positive sentiment (Hutto 2018).

The accuracy of VADER and its suitability for social media texts was confirmed in further tests (Ribeiro, Araújo, Gonçalves, Gonçalves, and Benevenuto 2016) indicating its suitability for analysing sentiment scores of Twitter messages. Because of this, and its easy applicability, the VADER model (as implemented in the corresponding Python Library[15]) was used to access the sentiment scores of the Twitter messages in the case study. The model can be tested on the companion website.

3.4.3 Topic Model Analysis

A considerably more complex task in NLP lies in the identification of the content of a given document. While sentiment scores can usually be captured on a two (positive-negative) or on a three (positive-negative-neutral) dimensional scale, topic modelling deals with much more dimensions and corresponding probabilities. The general methodology of topic assignment is also different. While sentiment scores are based on word-sentiment pairs (that can be obtained externally), topic modelling usually uses only the corpus as a basis and doesn’t derive information from external sources.

In general, topic modelling can be described as applying probabilistic models that try to infer semantic clusters (topics) in document collections (Wiedemann and Niekler 2016, p. 78). Most topic models treat documents as a mixture of topics, “where a topic is a probability distribution over words” (Yan, Guo, Lan, and Cheng 2013, p. 1445).

Usually topic models work with assigning topic proportions to each word of a document and imply the document allocation from the word distribution. The topics can be defined as different collections of words that appear together frequently. Therefore the term “bag-of-words” is also used when describing this approach. The topic is only defined by the words that its bag contains, regardless of their order. As an example, we can assume that if the words “Brexit”, “Johnson”, “Corbyn”, “Farage“, ”May“ (and further corresponding words) appear in a text, the topic will be probably politics (the bag that contains these, and lots of more similar words).

Obviously, in reality the structure and the underlying topic distribution is much more complex. It is seldom that a document only carries one topic and words can also fit in multiple bags, “May” might also indicate a text about referring to a month or using the word as a verb.

One of the most commonly used topic modelling approach is Latent Dirichlet Allocation (LDA), which assumes that documents have a specific topic distribution, which is derived from the word specific topic distributions (Blei 2012, p. 78).

LDA tries to allocate the words to the topics through inferring latent variables that constitute the document and topic structure. These hidden variables are “the topics, per-document topic distributions, and the per-document per-word topic assignments” (Blei 2012, p. 79).

The biggest challenge in LDA lies in the estimation of this hidden structure, which is done in most cases through sampling of the observed variables (the documents and the words in these documents). This process is also called model

[15] https://github.com/cjhutto/vaderSentiment

training/learning. One possibility is using Gibbs sampling, which approximates a Markov chain that can be used to compute the conditional distribution of latent variables (Blei 2012, p. 81).

In short, Gibbs sampling starts with a random assignment of topic proportions for each word in the document. The number of topics is set beforehand as a model input. The sampling goes through each word and topic individually and calculates the proportions of words in the document that are assigned to the topic t and the "proportion of assignments to topic t over all documents that come from this word" (Chen 2011).

Then, a new topic assignment for the word is given, based on the proportions calculated in the step before. This means, the assignments of all words stay the same, only of the single word's in question are changed. Through iterating over all words and changing the values one by one, it is possible to develop a model that depicts the topic proportions for the word and the overall proportion of the topics accurately (Chen 2011).

With the outputs of the sampling, LDA is able to run through all words in each document (or any given text, even if they were not in the training set). The model assigns the topic proportions for each word, the overall topic proportion of a document is calculated by summarising the single topic distributions of its containing words (Blei 2012, p. 80).

LDA has been used in lots of projects dealing with Twitter content, for example in the study about the spatial distribution of topics in London (Lansley and Longley 2016). Still, the specific characteristics of Twitter content make its application difficult and often inaccurate.

3.4.3.1. Biterm Topic Model

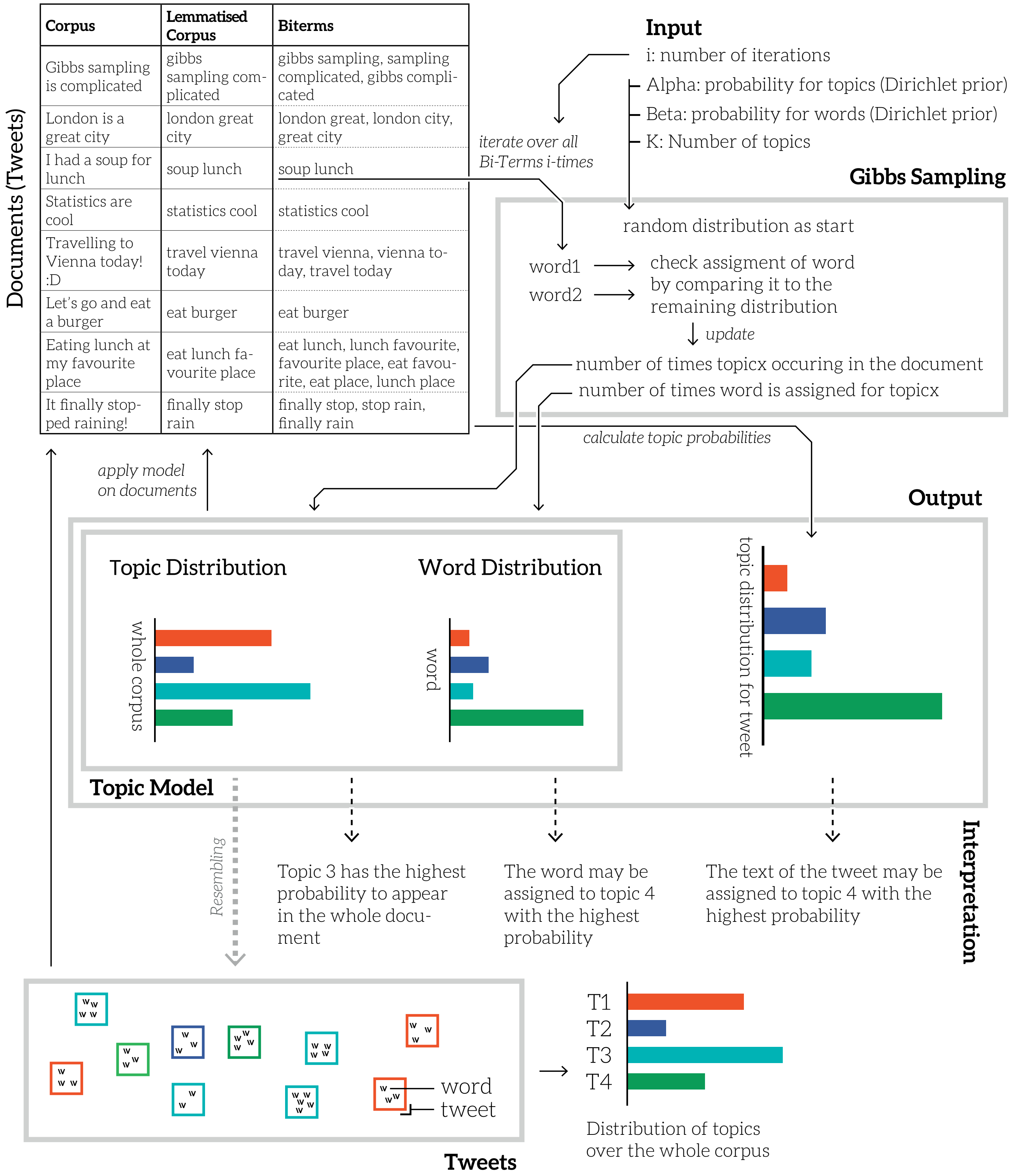
As described above, if the corpus for social media text analysis is the collection of tweets, the single messages become consequently its documents. But as LDA calculates and infers topic proportions based on documents, model training with Twitter text becomes difficult because of the short lengths.

In this case, it is possible to treat the whole collection as one document, which is inaccurate because the single tweets are in general individually created contents by different users and are dealing with different themes. In some cases, it might be possible to aggregate the messages to their respective authors. However, many users create only very little content and it cannot be assumed that one user won't deal with different topics in different messages. Therefore this approach also won't necessarily provide better results (Yan et al. 2013, p. 1445).

To overcome this problem, Yan et al. developed a novel approach, called Biterm Topic Model (BTM) which, in contrast to LDA, doesn't model the document generation process through topic distributions of single words but of that of the whole corpus using word-pairs (biterms) (Yan et al. 2013, p. 1446).

As shown in Figure 7, the pre-processed tweets are turned into biterms. Every possible word pair is derived from the tweet, and added to the set of biterms, which constitutes the basis for model learning. As an input, the number of topics is to be specified, together with the two Dirichlet priorts Alpha and Beta (Yan et al. 2013, p. 1448).

Topic learning is done through Gibbs sampling. The algorithm goes through all biterms in the biterm set and infers the topic and the word distributions. Outputs of the topic model are the topic distributions for the whole corpus in general (P(z)), as well as the distributions for each single word (P(w|z)).



Corpus	Lemmatised Corpus	Biterms
Gibbs sampling is complicated	gibbs sampling complicated	gibbs sampling, sampling complicated, gibbs complicated
London is a great city	london great city	london great, london city, great city
I had a soup for lunch	soup lunch	soup lunch
Statistics are cool	statistics cool	statistics cool
Travelling to Vienna today! :D	travel vienna today	travel vienna, vienna today, travel today
Let's go and eat a burger	eat burger	eat burger
Eating lunch at my favourite place	eat lunch favourite place	eat lunch, lunch favourite, favourite place, eat favourite, eat place, lunch place
It finally stopped raining!	finally stop rain	finally stop, stop rain, finally rain

Documents (Tweets)
Input
i: number of iterations
Alpha: probability for topics (Dirichlet prior)
Beta: probability for words (Dirichlet prior)
K: Number of topics
iterate over all Bi-Terms i-times
Gibbs Sampling
random distribution as start
word1
word2
check assigment of word by comparing it to the remaining distribution
update
number of times topicx occuring in the document
number of times word is assigned for topicx
calculate topic probabilities
apply model on documents
Output
Topic Distribution
whole corpus
Word Distribution
word
topic distribution for tweet
Topic Model
Resembling
Topic 3 has the highest probability to appear in the whole document
The word may be assigned to topic 4 with the highest probability
The text of the tweet may be assigned to topic 4 with the highest probability
Interpretation
word
tweet
Tweets
T1
T2
T3
T4
Distribution of topics over the whole corpus

Figure 7: Functioning of the Biterm Topic Model (own figure based on Yan et al. 2013)

The application of the model is relatively simple, the topic distribution of a given tweet is calculated using the following formula:

$$P(z|d) = \sum_b P(z|b)P(b|d)$$

with:

- P(z|d) being the proportion of topic z for the document (=tweet) d
- P(z|b) being the proportion of topic z for the biterm b
- P(b|d) being the proportion of biterm b in the document b

Formula 1: BTM application (Yan et al. 2013, p. 1448)

Assuming the biterm b consists of the word w1 and w2, P(z|b) can be calculated through Bayes' theorem:

$$P(z|b) = \frac{P(z)P(w1|z)P(w2|z)}{\sum_z P(z)P(w1|z)P(w2|z)}$$

Formula 2: Topic proportions for biterm b (Yan et al. 2013, p. 1448)

The proportion of biterms is obtained through the empirical distributions of biterms in the document:

$$P(b|d) = \frac{n_d(b)}{\sum_b n_d(b)}$$

Formula 3: Proprortion of biterm b in document d (Yan et al. 2013, p. 1448)

Here, nd(b) is "the frequency of the biterm b in the document d". As the authors underline, P(b|d) is in most cases a nearly uniform distribution. Nevertheless, the estimation still obtains good results (Yan et al. 2013, p. 1448).

Tests and comparisons to other topic modelling approaches were done by the model authors themselves (Yan et al. 2013, p. 1449ff), as well as by external researchers (Jónsson and Stolee 2017). The results of the evaluations have shown that BTM outperforms other approaches when working with short social media texts. Still, until now there were no projects dealing with topic modelling and geography that used BTM.

The good evaluation score of BTM and the lack of its application in the context of geographic research led to the decision to use this approach for the case study of this work. As the sentiment analyser VADER, the Biterm Topic Model can also be tried out on the companion website.

3.5. Allocating the Message

3.5.1 Location/Place/Coordinates

Twitter's way of handling location data might be in some cases complicated or maybe even contradictory. Although the terms location, place and coordinates may be used in many cases interchangeably, the portal uses them to describe different phenomena.

Location describes a place a user's profile is assigned to. By default, it is the name of the city where the user created their profile. This information can be derived by the IP-address without the need for further information (such as GPS-coordinates).

The location can be changed anytime by typing in a new location in the corresponding field in profile settings. It can be chosen from a list containing city, region, and country names. The location field does not change if the user travels or moves to another city, it needs to be updated manually. Alternatively, the user can also define their place themselves by typing in any location that is not included in the list of toponyms. Therefore, some user's location doesn't refer to a real geographic place, but to a fictious one, like "Mordor", "Nowhere" or "at work".

The place attribute is attached to a tweet and the user can chose it from the same list of toponyms as at the location field, but the chosen place is saved as a geoJSON object on the server. By default, this attribute is enabled and is their profile's location (if it is a real place), or the last place the user tweeted from. The place can be changed anytime and allows to be set freely (although only real places from the list may be chosen). Meaning, a user can create a tweet in

London, shortly thereafter one in Hanoi, while the whole time being in Vienna.

The most accurate position of a tweet is found in the attribute field coordinates. If the user enables sharing the GPS-coordinates of their device with Twitter[16], the exact location data will be shared with the portal as a latitude/longitude pair, saved as a geoJSON file. This input cannot be changed, Twitter saves the exact coordinates provided by the device without the possibility of further modifications.

In some cases, where users link a post from a third party site (for example Instagram or Foursquare), the coordinates are provided from the other site. These sites have a different method of collecting location information. Users don't necessarily need to share their GPS coordinates. Alternatively, they can also select a place or an amenity they have visited (for example, a park, a bar, or a museum) and their post will be georeferenced to be in the centre of the bounding box of the object they have chosen.

For further analysis it is important to discuss the question how to treat such cases. These tweets have exact coordinates but the correctness of these coordinates is not guaranteed. If a user selects a bar as their location manually, the error may be just a few meters but if they chose a city, the deviation may make up many kilometres. As Twitter doesn't provide any metadata whether the coordinates were provided by the device directly or they were set by the third party portals, there is no possibility to check the validity of the coordinate information.

Although this phenomenon can account for a large difference (in the case study 9 per cent of the tweets have a populated coordinate field but only 7 per cent of those are really accurate), only a few research projects did distinguish between the two forms of coordinate definitions.

To distinguish between place and coordinates is also crucial. As stated above, Twitter provides all tweets that were created within the queried bounding box. A tweet will be queried if either the coordinates are inside the bounding box or the centre of the bounding box of the given place (for example a city). This explains why the most tweets don't have their coordinates provided (even though they were sampled from a geographic bounding box) and also why there were also tweets recorded that have exact coordinates that fall outside the bounding box. In this case the users provided the location of their tweets to be in London (thus within the bounding box) but tweeted from somewhere else.

3.5.2 Aggregation of Tweets

As pointed out before, the aggregation and visualisation of data should be done with care, as these techniques intensively influence the way we perceive the depicted phenomenon. The common display of tweets as points on a map may be questioned, both in a technical and in an ethical way (see *2.11. Critique – p. 16* and *2.12. Ethical Questions – p. 20*).

To solve this problem, it is possible to aggregate the tweets on given administrative entities, such as boroughs, cities or regions (as done for example in García-Palomares et al. 2018; Yang and Mu 2015).

This way is most suitable if Twitter data is blended with other information that is available on the level of these entities (for example census data). However this approach disregards a relevant capability of Twitter data, namely its flexible possibilities of aggregation (Venturini et al. 2017, p. 4).

Therefore, in many projects Tweets are aggregated into evenly distributed rectangular grids in the study area (as done for example in Lamanna et al. 2018; Lansley and Longley 2016). Although this approach mostly delivers a better picture of the described phenomenon, it is

[16] By default this feature is turned off.

harder scalable than using hexagons (Shelton, Poorthuis, Graham, and Zook 2014, p. 171).

Furthermore, hexagons are less distracting for map readers (Carr, Olsen, and White 1992 as cited in Shelton et al. 2014, p. 171), have a "higher representational accuracy" (Scott 1985 as cited in Shelton et al. 2014, p. 171) and each cell having six neighbours instead of four further smoothens the visualisation.

As it is a basic paradigm in cartography to abstain from displaying absolute values on choropleth maps (unless the single features have the same basic attributes – as population), it is possible to display the number of the tweets with a certain phenomenon in a specific cell in relation to the total number of tweets in that cell. Although this approach shows the local distributions accurately, it doesn't provide information about the overall concentration of the searched phenomenon.

Therefore, Poorthuis et al. (2014, p. 8) favour the Odds Ratio, a method borrowed from spatial economics where it is called location quotient[17].

The Odds Ratio can be calculated as follows:

$$OR = \frac{p_i/p}{r_i/r}$$

with:

- pi being the number of tweets in area i related to a specific phenomenon
- p being the total number of tweets related to the phenomenon
- ri being the total number of tweets in area i
- r being the total number of tweets in the whole set

Formula 4: Odds Ratio (Poorthuis et al. 2014, p. 8)

When using this formula, a result of 1 indicates that in the inspected cell the distribution of the phenomenon equals the overall distribution in all the tweets. 0.5 indicate the distribution to be the half, 2 the double.

This formula helps to create more correct assumptions about Twitter activities, with a drawback being the susceptibility for more extreme results in cases with a low number of tweets.

Hence, a weighting is necessary to get finer results. Poorthuis et al. (2014, p. 8) argue for using the lower bounds of the confidence interval of the OR (in this project called Weighted Odds Ratio – WOR), using the following formula:

$$OR_{lower} = e^{\ln(ORi) - 1{,}96 * \sqrt{\frac{1}{p_i} + \frac{1}{p} + \frac{1}{r_i} + \frac{1}{r}}}$$

Formula 5: Weighted Odds Ratio (Poorthuis et al. 2014, p. 8)

The formula above indicates a value being over 1 also being significant with 95% of confidence as well (Poorthuis et al. 2014, p. 8).

Shelton (2017) uses this formula to show the different messages two maps can carry when displaying the same phenomenon. He compares Simon Roger's Ferguson map (see chapter other projects) using the absolute number of tweets with his approach using the Weighted Odds Ratio.

Shelton's map shows a strong concentration of tweets related to the phenomenon in and close to the St. Louis area (where the killing happened), whereas Roger's visualisation suggested an explosion of the topic throughout the world (Shelton 2017).

These calculation and visualisation approaches are also shown on the companion website to present the different messages a map can carry when using different methods.

3.6. Implementation in Python

The implementation of the steps sketched above requires a flexible and a specifically fitted approach. Although some software solutions exist that can conduct natural language processing to a certain grade, their limitations appear soon. The needed flexibility and applicability is guaranteed when conducting the analysis using

[17] Some projects (for example Lansley and Longley 2016) use this term for the same calculation method.

a self-written code, in this case drawing on the programming language Python.

Python was designed in the 1990s by Guido van Rossum and has been steadily maintained and developed by an active community since then (Pérez, Granger, and Hunter 2011, p. 15). This work uses the version 3.6, released in 2016[18].

Python is a language commonly used by scientists and researchers with various focuses, as it uses a syntax that is very easy to learn and apply. Furthermore, there are lots of libraries and modules available. These are code collections that can be imported and adjusted for a certain task (Pérez et al. 2011, p. 14).

For example, when conducting sentiment analysis, it is not necessary to write an algorithm or create a word-sentiment lexicon. The developers of VADER provided a Python library that can be downloaded and applied by anyone without any profound prior knowledge in text processing.

A full Python tutorial would go beyond the constraints of this thesis, there are several resources online[19] providing good guides for learning the language, therefore in the next section the used modules, libraries and the general structure of the implementation are sketched.

3.6.1 Software Architecture

Figure 8 shows the basic computing scheme of this project with the most important software components.

The connection to the Twitter server via the official API was established through the Tweepy library. Tweepy enables to send any kinds of inquiries to the API, it can be used for posting and reading messages, updating the profile, or as in this case, streaming data.

As the Streaming API needs a steady and uninterrupted connection, the query was transferred to a server by the British provider Pythonanywhere[20]. The downloaded data was stored in a MySQL database, which was accessed via the library PyMySQL.

The streaming data is saved in three database tables. The first, “tweets”, contains all information of the specific tweets and is the basis for most of further analysis steps. User information is separated from the tweet information, and is only updated in case changes occur. Therefore duplicate information can be eliminated, leading to less required disk space and easier data analysis. The third table, errors, contain all information about problems that occurred during code execution. The most common problem was caused by database locks, connection problems, and exceeding Twitter’s number of reconnection attempts (Error 104).

Data analysis tasks were done partially on the server (especially long running tasks, such as tokenisation, sentiment analysis and tasks involving a lot of communication with the database), partially on a notebook (mostly when writing and testing new codes or conducting smaller tasks).

For sentiment analysis, the library VADER was used, which is distributed by the authors of the algorithm and available for everyone.

The pre-processing of texts for topic modelling was done with help of the NLTK-library (Natural Language ToolKit), which is a broad framework for several NL processing tasks. The tasks carried through NLTK included tokenisation, POS-tagging, and lemmatisation.

[18] Python 3.7 was released in June 2018 (Python Software Foundation 2018a) but for ensuring the compatibility with all modules and the server, the previous release was used.

[19] For example **https://www.w3schools.com/python**/, **https://docs.python.org/3.6/tutorial**/, and also the documentations of the modules and libraries provide often profound explanations.

[20] **https://www.pythonanywhere.com**/

There is currently no Python framework for BTM available, but the authors of the model provided a shell script that carried out model building. Applying the topic model on the tweets was also done by a self-written Python code.

The localisation and of tweets and their assignments to polygons was also done externally, using the GIS-programme QGIS. Further geographic tasks (for example calculating OR) were carried out in Python.

For visualisation, the framework Plot.ly was used, with a website setup through Dash. The companion website is hosted on the premises of the same company that also hosts the computing server and the database and is implemented via the web framework Flask.

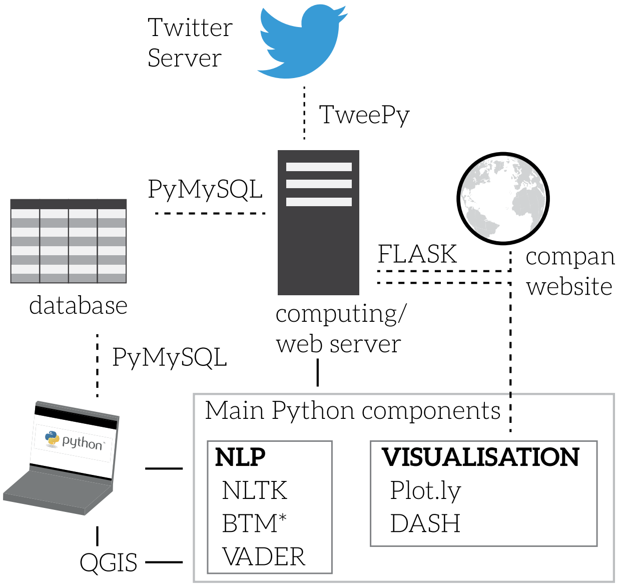


Figure 8: Software architecture (own figure)

4. Case Study

As presented in the previous chapters, the usage of Twitter data is eagerly discussed by researchers and planners. Although numerous projects try to capture different phenomena of a city through the lens of social media analysis, many scholars argue that such data is too heavily biased to become useful in everyday life. In addition, most of the projects presented above conducted very basic descriptive research about their social media usage in the cities, but had only little relevance for concrete tasks of spatial planning.

In addition, it is important to emphasise the specific challenges that come with Big Data analysis. Data needs to be pre-processed, prepared, and analysed in order to extract information. Still, this information also isn't necessarily knowledge. As HP's former CEO Carly Fiorina once put it, the goal of data analysis is to "transform data into information and information into insight" (Fiorina 2004).

The overall motivation behind this project was not only to find traces of information in a dataset of enormous volume, but also to evaluate the application fields of such information for the domain of spatial planning. To do so, a case study was conducted, based on 8.3 million Tweets captured from May to October 2018 in the city of London. This case study applied the computational linguistic methodologies presented in the third chapter, identified topics, captured the sentiments of Twitter messages and tried to discover spatial and/or temporal distribution patterns of different aspects of the analysis.

As many critical researchers pointed out, the large amount of data might easily lead to drawing false or misleading conclusions. Therefore, an important part of the case study was also the evaluation of the quality of the captured data.

For this project, Twitter data is expected to be used as source of information regarding themes and topics that couldn't be measured easily by traditional means, or where the possibility of real-time surveying and processing is expected to become a valuable asset. There were no concrete assumptions about the topics to capture at the beginning of the analysis. These were discovered through the application of the Biterm Topic Model, identifying 100 topics in the corpus.

London was chosen as a study area because of its high number of Twitter users (as compared to other European cities) and the widespread use of the English language. NLP techniques work with this language in the most sophisticated form and will presumably provide the most accurate results.

Although, as described before, SMGI differs in many cases from AGI (to which many data quality indicators were developed), its assessment can be carried out based on standards and common approaches of traditional data quality assessment (Fonte, Antoniou, Bastin, Estima, et al. 2017).

First, the overall quality of data is evaluated, focusing on its accuracy, consistency, and completeness, commonly conjoined under the term "internal quality" (Devillers and Jeansoulin 2006a, p. 37). In contrast, the concrete usability of the information is described as "external quality" (Devillers and Jeansoulin 2006a, p. 38), assessed in the case study by comparing the downloaded data to various reliable indicators and the validation of three thesis referring to the usage of Twitter data in urban planning.

4.1. Companion Website

The nature of the case study by working with a very large number of tweets and topics, and applying different aggregation and calculation methods, as well as the combination of different aspects and approaches led to very heterogeneous results. Because all of the combinations cannot be presented in this paper, a companion website was developed, containing figures and maps presented in the following chapter.

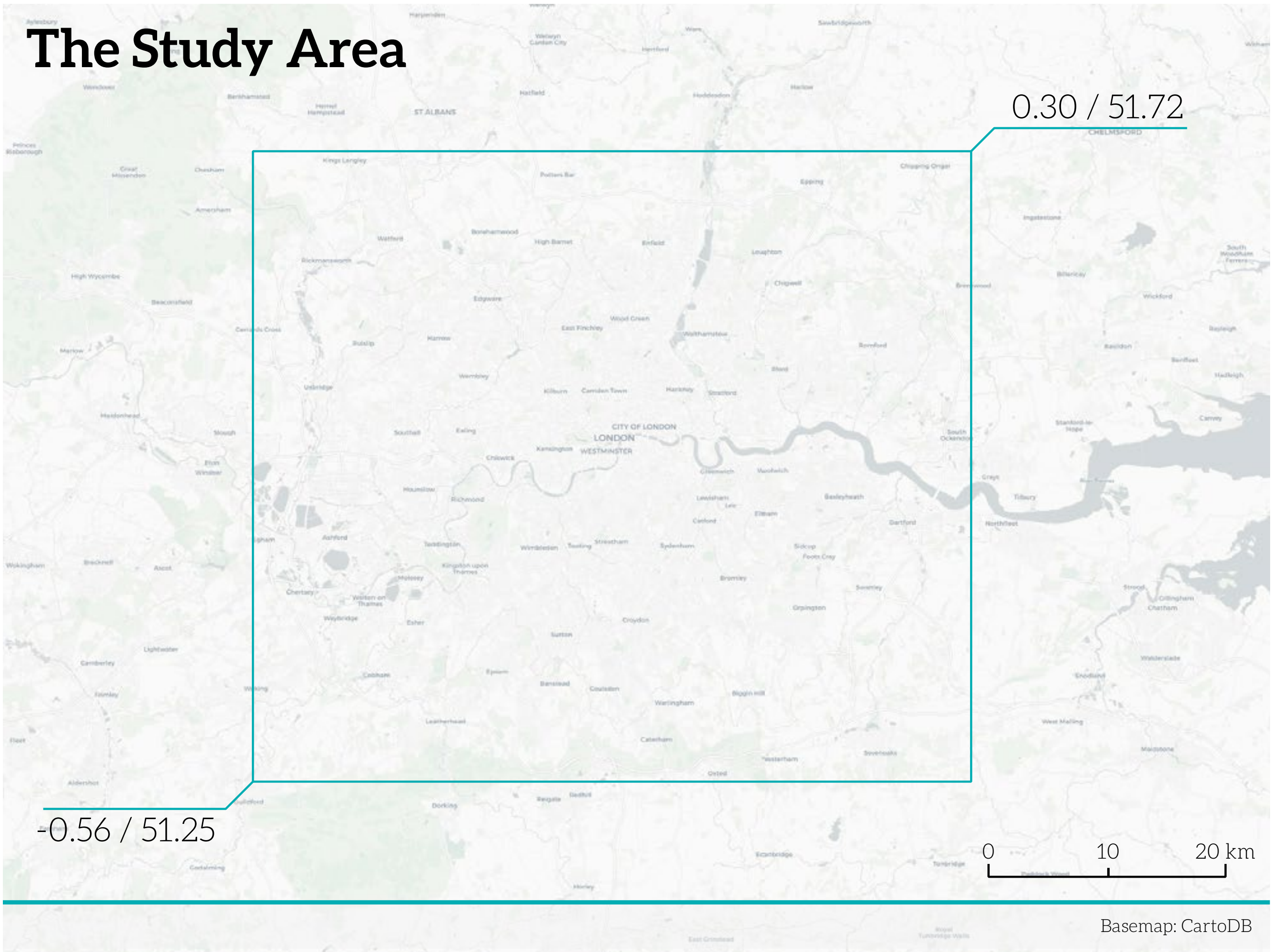
The figures on the following pages depict the most relevant and interesting combinations of themes and filter/calculation methods. Still, it is often advisable to open the interactive graphics on the website and apply different combinations and compare the differences of the outcomes. An important aim of this project is also to show how knowledge might be derived from the (almost) raw dataset and which impact only small changes in the calculation methods on the outcome may have.

The website can be accessed via the following link: **www.londontweets.eu**

4.2. Scale

During the course of the case study (from May 17th to October 23rd 2018), 8 350 771 tweets were recorded in the city of London. The dataset was downloaded via Twitter's Streaming API, by defining a bounding box around the Motorway M25. The study area is shown in Figure 9 below.

756 241 of all recorded tweets were georeferenced, accounting for 9.05 per cent of the whole dataset. This number is much higher than the numbers shown in other studies (which commonly report a share to be around 1 to 4 per cent), although lots of these tweets do not have correct coordinates. The number of tweets with correct coordinates makes up only 53 825, or 0.64 per cent, which lies below the numbers of other projects.



Basemap: CartoDB

Figure 9: Study area (own map)

Looking at the users of the portal, an indication of the 90-9-1 rule (Van Mierlo 2014) is observable. Lots of users created a little number of tweets, while a small fraction accounted for a large share of the messages. Figure 10 shows the accumulated numbers of tweets users created. In the case study 3 per cent of users created 54 per cent of all tweets in the dataset.

The whole dataset had altogether 400 502 users. Focusing at georeferenced tweets, the number of captured users was 108 120. Concentrating only on users that created correctly georeferenced tweets, the number is 18 865. As a comparison, these users (including bots, tourists, and corporate accounts) would make up 0.2 per cent of London's population.

As there is no demographic data available on Twitter, it is important to emphasize that it is impossible to draw representative and valid conclusions about the city as a whole. Therefore, the aim of this case study is also not to describe London and aspects of the city's life in general, but some certain topics and characteristics that may be captured through social media analysis.

4.3. Measures/Filters

Of course the 8.3 million tweets captured need to be pre-processed, filtered, aggregated and transformed into a form that makes the data suitable for further evaluation. The first step is to define the subset of tweets that will constitute the basis for the analysis.

Because the content analysis was applied to English language tweets, messages composed in other languages were not included in the assessment. The assignment of a language was based on the automatic language tags Twitter assigns to its content.

A large portion of Twitter messages make up content created by bots and/or for advertisement purposes. Scientific studies estimate the share of bots being around 9-15 per cent of active users, (Varol, Ferrara, Davis, Menczer, and Flammini 2017, p. 280), while Twitter's own evaluations predicate this number to be around 8.5 per cent (Seward 2014).

Filtering out bots is a challenging task. In general, non-human users might be identified by their activity patterns (for example posting a very high number of tweets in a short time frame, in some cases from different places too), or also based on the content of their messages. In

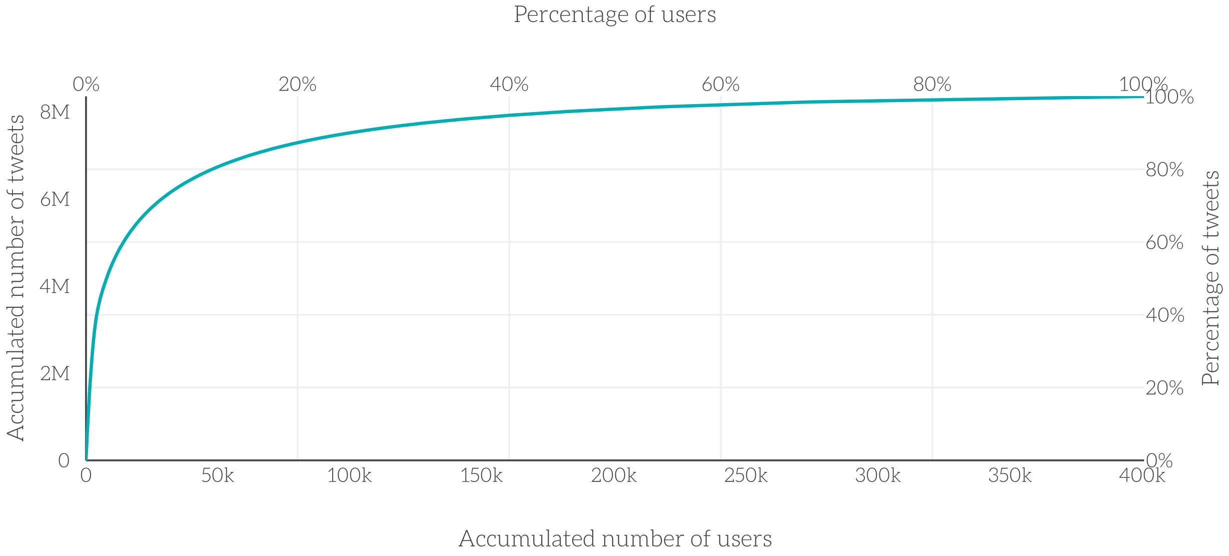


Figure 10: Aggregated numbers of users and Twitter messages (all recorded tweets - own figure)

this case study, the second approach was chosen. As the Biterm Topic Model was applied on all Twitter messages, the results indicated that the model successfully identified such messages and assigned them to certain topics. The categorisation was evaluated manually and has proven to yield satisfactory results. Therefore, for further analysis, messages assigned to these topics were disregarded (they are shown only if they wouldn't affect the outcomes of calculations).

An important aim of the companion website is to present the often very differing results that come up when applying different calculation and/or filtering methods on the data. As there is a very strong concentration of tweets in some areas of the city, displaying the absolute number of tweets yields in almost all cases the same results. This difficulty was sketched in *3.5.2 Aggregation of Tweets (p. 41)* where the two measuring concepts, Odds Ratio and the lower bound of the confidence interval of the Odds Ratio (Weighted Odds Ratio) were presented. These calculation methods can be chosen in the figures on the companion website, along with the absolute and the relative (e.g. ratio of a certain topic in all tweets of a polygon) numbers of tweets.

For enabling an easier comparability, the figures adopted the denotations of Shelton et al. (2015) with p referring to the total number of tweets related to the presented phenomenon and r to the total number of tweets the analysis was based upon.

Another main uncertainty results from the unreliable correctness of coordinates. As also sketched in the third chapter (*3.5.1 Location/Place/Coordinates – p. 40*), in case coordinates are provided by an external service, they don't necessarily refer to the accurate location of a tweet, but to the centre of the bounding box where the tweet was submitted (or the location tagged in the post). When conducting a study on a regional level (and where the location of tweets within a city is irrelevant) such errors will possibly not affect the outcomes considerably, but the case study required a higher level of accuracy.

As there is no direct information in the metadata of the tweets about the source of the coordinates, the approach was to count the number of Twitter posts for each coordinate and assuming those to be correct that are referred to in only one message. Positions are provided by an accuracy of 9 decimals, meaning that a deviation in one decimal means a distance of only one meter. Therefore it is highly unlikely that two tweets contain the same exactly coordinates.

This assumption was also tested by inspecting the content of some Twitter messages manually and looking for references to other social media sites.

Because of the two locational attributes Twitter assigns to its messages (place and coordinates), there were also many tweets captured outside the defined bounding box. As Figure 11 shows, these tweets are distributed not only near the edges of the bounding box but they appear virtually all around the world. These have their coordinates field populated, but the place attribute set to be in London.

In spatial analysis only the tweets with coordinates within the bounding box were included, but the general evaluation was applied to all tweets set to be in London. Therefore it is important to underline that the general evaluation also included a number of tweets that were not composed in London. Because of the lack of metadata it is impossible to estimate the numbers of such tweets, let alone being able to filter them out.

The above described large differences in user's activity patterns also resulted in varying spatial distribution patterns. A very small number of users created a large amount of tweets, therefore in many cases the activity of distinct users in some polygons led to distorted results. For getting a clearer picture, their messages were aggregated to a form that they appear only once in the figures (for example if a user sends 20 messages containing the topic education in one polygon, only one of these is counted).

Of course, these approaches are very generic and might possibly lead to losing some valuable data. For example at cases when only a very small subset of the tweets is relevant, applying too strict filter methods might lead to not having any relevant tweets to analyse. Therefore the case study applied the strictest possible (but still suitable) filter methods. The figures presented in this chapter depict the most valuable and significant results. Still, all the maps and diagrams are also available on the companion website, with the possibility of testing and evaluating the different combinations of these settings.

4.4. Spatial and Temporal Distribution

As a basis for spatial analysis, the tweets were aggregated into 4 types of polygons. The smallest aggregation level constitutes a 620 * 620 m rectangular grid consisting of 8613 cells. This level might provide insight into the captured phenomena on a very fine scale.

The next level constitute 4408 regular hexagons with an outer radius of 540 m each. In most cases, they were the most suitable way to show the phenomena, as the size of the grid cells was often too small and didn't capture enough tweets to draw reliable conclusions.

Furthermore, the tweets were also aggregated into two administrative divisions, the 33 boroughs and 625 wards of London. This enabled the direct comparison of the recorded data with

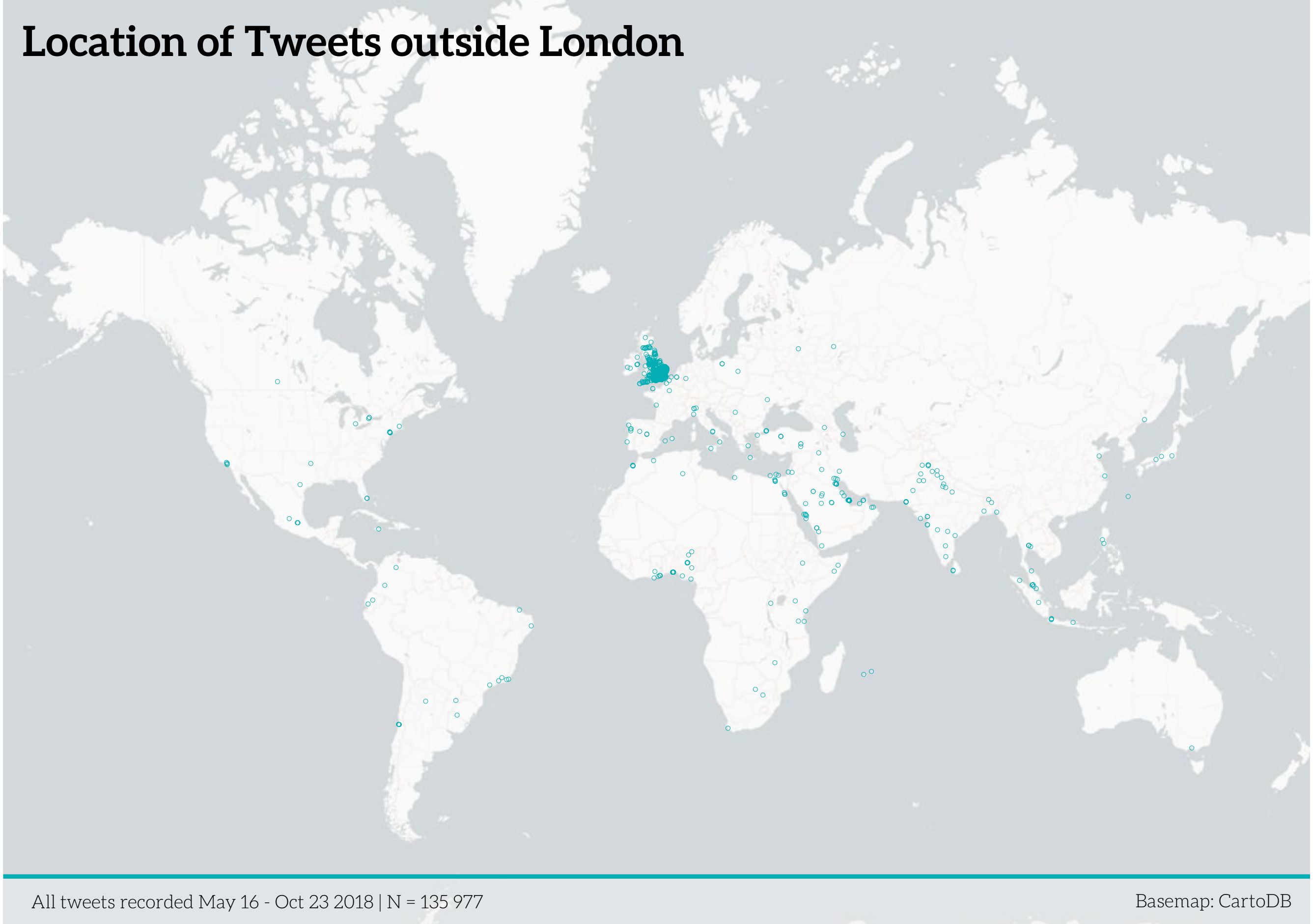

Location of Tweets outside London
All tweets recorded May 16 - Oct 23 2018 | N = 135 977
Basemap: CartoDB

Figure 11: Location of tweets outside London (own map)

administrative (and validated) information and helped to check the underlying phenomena.

Unsurprisingly, the highest Twitter activity is observable in the central districts of the city, near to (touristic) attractions as the Hyde Park, the Tower or SoHo (Figure 12).

Some enhanced activity appears near transportation hubs, at the Airport Heathrow, or near to some outer centres and attractions, as the Docklands.

Figure 13 shows the temporal distributions of the tweets during the course of the case study. It is important to note the two gaps in the charts. As mentioned before, when downloading tweets using the Streaming API, Twitter needs a stable connection, else it stops sending data. Then, the code needs to be restarted manually. In the beginning, the code connecting to the API was not robust enough and disconnections happened unfortunately frequently. Therefore, in the first two weeks of data capture, there was a smaller amount of tweets downloaded. Because of the very small number of valuable tweets, the information captured during these days were not disregarded in further analysis and processing.

The second gap appears between the 21st and 26th of August results from too frequent reconnection attempts which remained unfortunately undetected until the data was downloaded.

Apart from these two issues, the number of captured tweets ranged between 55 and 60 thousand messages per day, with three very active days (3rd, 7th, and 11th of July). These were the evenings when England played matches during the Football World Cup, leading to an increased

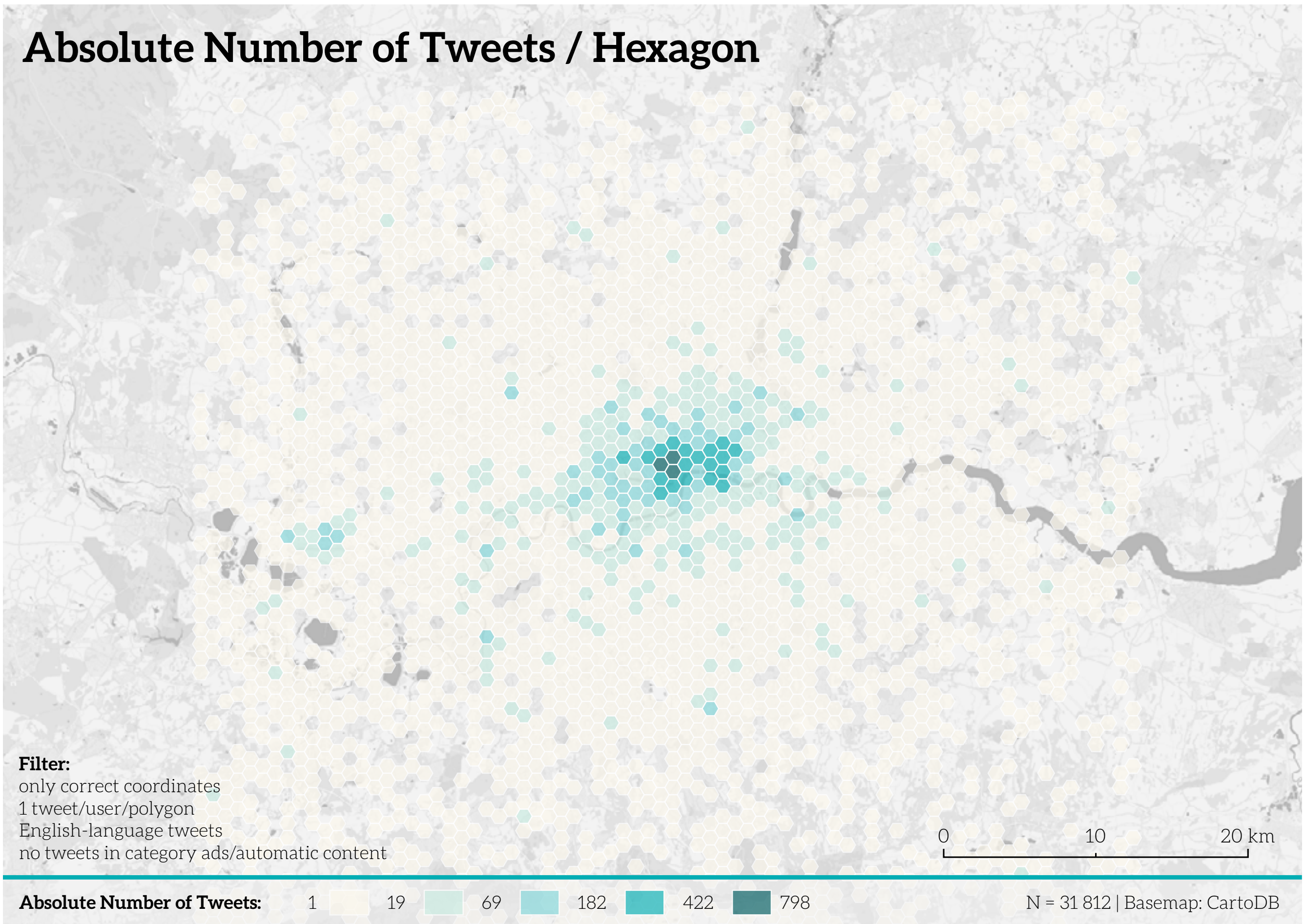

Absolute Number of Tweets / Hexagon
Filter:
only correct coordinates
1 tweet/user/polygon
English-language tweets
no tweets in category ads/automatic content
0
10
20 km
Absolute Number of Tweets:
1
19
69
182
422
798
N = 31 812 | Basemap: CartoDB

Figure 12: Spatial distribution of tweets (own map)

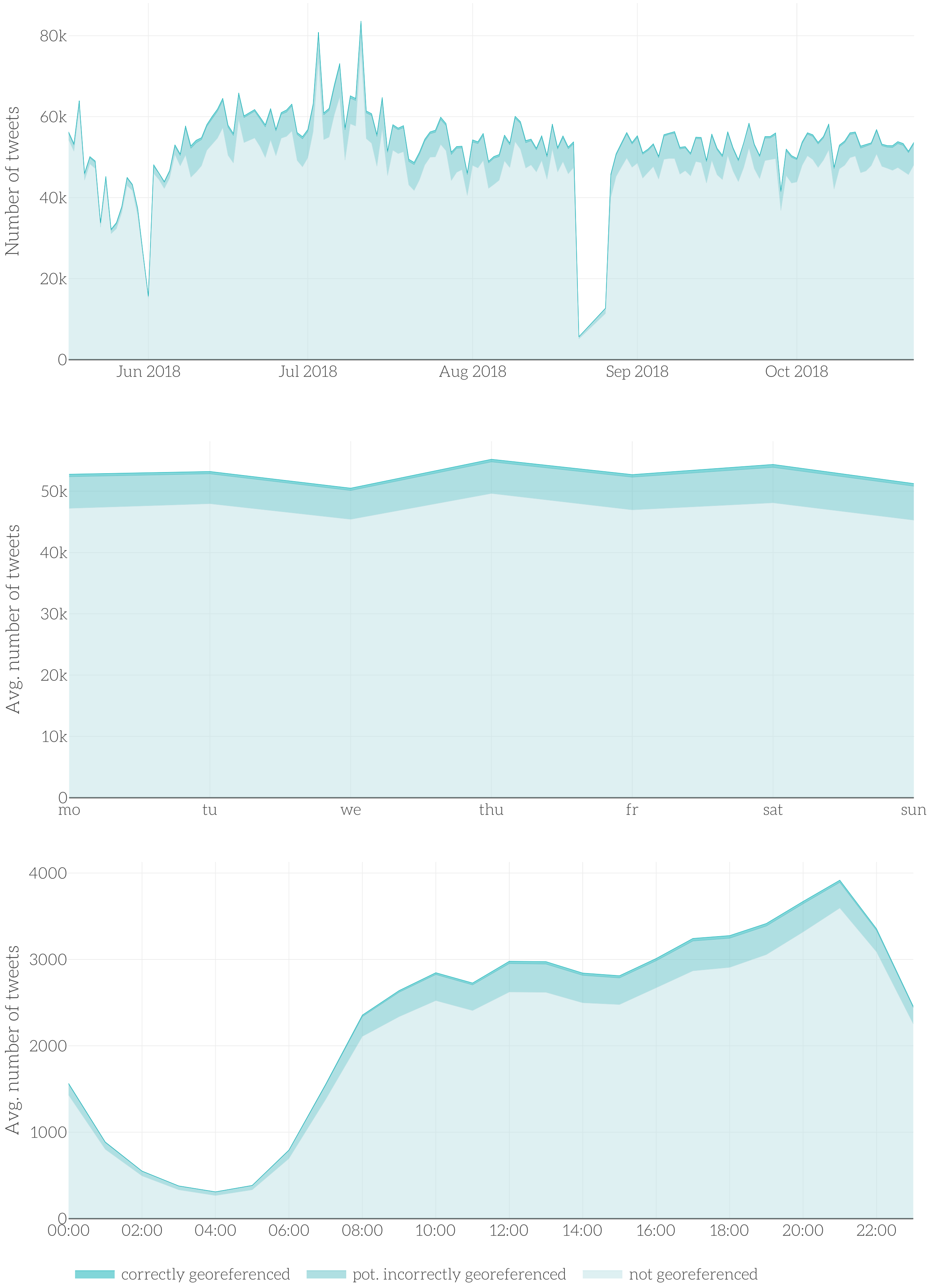


Figure 13: Temporal distribution of the recorded tweets per hour, weekday, and hour (own figure)

amount of Twitter messages apparently. This assumption was also confirmed by the topic model, which indicated a very large number of tweets dealing with football on these days.

The number of tweets during a week remained relatively constant, although (as it is detailed below) there are some differences between single topics. Regarding the number of tweets published during a course of the day, it is unsurprising that most messages are published around evening, while during 1 and 6 pm there is a very low activity recorded.

As a conclusion, Table 7 shows the numbers of tweets included in the different analysis tasks[1]:

4.5. Content

4.5.1 Topics

The analysis of the semantic content of the downloaded Twitter messages was done by applying the Biterm Topic Model approach (see *3.4.3.1. Biterm Topic Model – p. 38*). Compared to other approaches used in the reference projects, BTM promised to provide results in the highest quality. It was developed specifically for the analysis of a large amount of short messages and the developers also tested their model on the basis of tweets (Yan et al. 2013). An external study also indicated a satisfactory reliability of the BTM when working with tweets (Jónsson and Stolee 2017)

The topic modelling algorithm was trained on all English language tweets created between June 2nd and August 31st 2018 containing at least 3 tokens (excluding stopwords, special characters and other elements – for the filtering process see *3.3.5 Further Steps – p. 32*). The tokens of the tweets were lemmatised by the "WordNet Lemmatizer"-Algorithm of the Python-NL-toolkit.

The tokenised tweet set contained 628 836 different words, but only words with at least 10 appearances (58 655) were provided to the training set. Overall, after filtering the messages, 2 840 303 tweets remained in the training set.

The number of topics (k) was set to be 100. This number was also used in other projects, such as the London study by Lansley and Longley (2016). As in model testing by the developers (Yan et al. 2013, p. 1449), topic learning consisted of 1 000 iterations with the Dirichlet priors alpha being 0.005 (50/k) and beta 0.01.

The quality of the topic model was assessed by selecting the top 20 words for each topic and the words ranked from 1001 to 1020 and comparing the coherence manually (as also done by the developers – Yan et al. 2013, p. 1451). The top words for each category are listed in the annex (*Queried Topics – p. 91*).

Task	Numbers					Sum
	Not georeferenced	Potentially incorrectly georeferenced		Correctly goreferenced		
		All	Within the bounding box	All	Within the bounding box	
All tweets	7 458 223	825 473	702 416	66 745	53 825	8 350 441
English-language Tweets	6 076 551	753 854	635 772	59 552	47 521	6 889 957
Topic modelling	4 355 442	615 846	518 618	47 410	38 258	5 018 698
Sentiment detected	3 967 433	451 943	383208	31298	26512	4 450 674

Table 7: Numbers of analysed tweets (own table)

[1] Maps presenting topic distribution patterns also included tweets with no assigned topics in the calculations

Ads/automatic content

Topic	Topic Proportion
Radio programme	0.000990
Social media trends	0.000663
Ticket sales	0.005437
Social media trends in the UK	0.003614
Road traffic	0.007445
Traffic information	0.009650
Ads for social media channels	0.006258
Sales	0.008577
Customer service	0.011034
Job offers	0.003698
Weatherbots	0.003943

Architecture/tourism/travel

Topic	Topic Proportion
Housing / architecture	0.008110
London culture	0.002069
Sea / water	0.002310
Place check-ins	0.008286
Travelling	0.011370

Food & drinks

Topic	Topic Proportion
Desserts	0.005661
Food (restaurants)	0.007393
Food (home-made)	0.006692
Drinking	0.005042

Social Issues

Topic	Topic Proportion
Crime	0.008502
Social controversies	0.014142
World news	0.005390
Religion	0.003539
LGBT themes	0.002299
Social issues	0.013422
Social system	0.016348

Social Media

Topic	Topic Proportion
Social media	0.009906
Ask for support	0.012493

Personal topics

Topic	Topic Proportion
Pets / animals	0.001833
Family	0.018686
Wedding	0.004977
Death	0.003041
Retrospect	0.021406

Informal communication

Topic	Topic Proportion
Anticipation ("can't wait")	0.004657
Negations	0.016916
Anticipation ("looking forward")	0.004752
Congratulations	0.010234
Approval of arguments	0.025602
Superlatives	0.007079
Slang	0.023331
Cursing	0.015749
Appreciation	0.016750
Motivation	0.018188
Acknowledgements / gratitude	0.008281
Birthday wishes	0.006008
Felicitations	0.017237

Leisure

Topic	Topic Proportion
Clubbing	0.012029
Love Island	0.012815
Celebrities	0.014826
Free time	0.013491
Plants / gardens	0.006729

Transport

Topic	Topic Proportion
Airplanes / ships	0.002364
Railways	0.010988

Sustainability/charity

Topic	Topic Proportion
Food bank	0.001483
Charity / volunteering	0.023741
Sustainability	0.004522

Arts/culture

Topic	Topic Proportion
Literature	0.002852
Music	0.011555
Movies	0.006374
Visual arts	0.004839
Series	0.007990

Body/health/appearance

Topic	Topic Proportion
Cosmetics	0.003686
Fashion	0.005516
Body issues	0.011988
Health	0.008104

None

Topic	Topic Proportion
None1	0.026982
None2	0.018745
None3	0.027301
None4	0.022930
None5	0.023730
None6	0.027107
None7	0.008572
None8	0.024580

Politics

Topic	Topic Proportion
Elections	0.010252
Trump	0.005058
Brexit	0.010310

Sports

Topic	Topic Proportion
Cricket	0.003072
Football players	0.010152
Premier league	0.013203
Football	0.014655
Racing	0.002344
Tennis	0.003610
Football World Cup	0.014382
Sport matches	0.005629
Football World Cup England	0.008475
Boxing	0.001412

Everyday activities/topics

Topic	Topic Proportion
Weather	0.003950
Fitness	0.005105
Sleeping/resting	0.013651
Education	0.007478

Business & finances

Topic	Topic Proportion
Business events	0.015016
Finances	0.017221
Business	0.023322
Technology	0.007873

Cultural events

Topic	Topic Proportion
Carnivals/fairs	0.000930
Pop concerts	0.002168
Concerts (more classical but also pop)	0.005327
Event announcements	0.018136
Open air concerts	0.002424

Figure 14: Topics recognised by the BTM (own figure)

Furthermore, the topic model can also be tested on the website by typing in an English language text. The topic model algorithm then presents the five topics with the highest probabilities assigned to the text.

The topic model was subsequently applied to all English language (and tokenised) tweets in the complete downloaded data set.

For topic learning the code provided by the developers of the model was used, while for text preparation and application own codes were written that suited the data structure and the aim of the project well.

The BTM generated 100 topics, named by inspecting their most common words and phrases manually. Later these topics were classified into 18 categories. For 7 topics a clear identification of their content was not possible, these remained unnamed. Figure 14 shows the categorisation with the corresponding topic proportions (related to appearing throughout the whole corpus).

For consistent and simple further processing, it was assumed that all tweets contain one topic, namely the one with the highest proportion. This assignment was tested manually by inspecting the tweets and their assigned topics and proved to be suitable for further analysis.

As BTM is based on word frequency distribution patterns, consequently in some cases it created categories by words that don't necessarily carry a semantic meaning. These topics include frequently used words that were used regardless of the implied meaning of the tweet. For example the topic 'negations' incorporates words like 'no', 'not', 'don't', 'never', etc. regardless of the topic it is about. This means that the sentence "I don't care about politics and I never will" is assigned to the negations topic with a proportion of 0.4, followed by "social controversies" with 0.06. To these topics, the category "informal communication" was assigned. Informal communication makes up the largest topic category, accounting for about 19 per cent of all tweets.

The second largest group make up topics that couldn't be recognised clearly. The words and tweets assigned to this category are very diverse, it is almost impossible to recognise easily identifiable and coherent patterns.

Automatic content by bots or advertisements make up about 5.6 per cent of the surveyed data. These contents are (as mentioned above) often

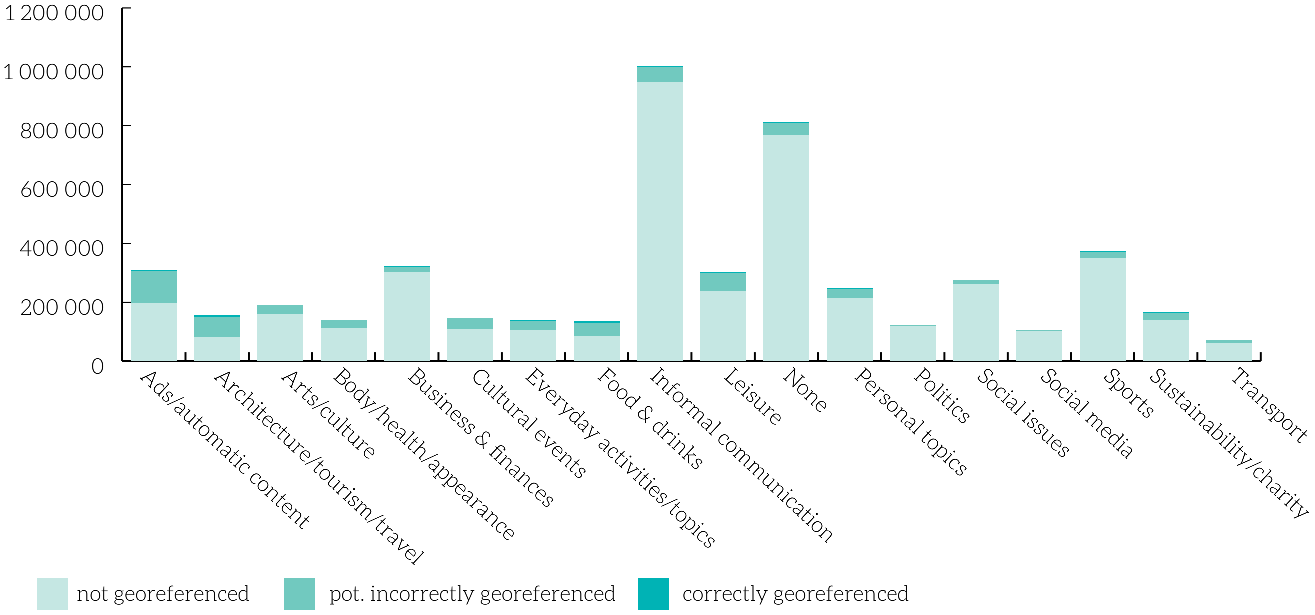


Figure 15: Number of tweets assigned to topic categories (whole dataset, own figure)

correctly recognised by BTM, as a very static structure and the usage of relatively few words are characteristic for such messages. Good examples are weatherbots that post in a regular interval meteorological information from their stations. Others are job offers and advertisement channels, these accounts make up a relatively large share. Messages assigned to other categories don't appear to be machine generated. Therefore, during data analysis, tweets assigned to these categories were disregarded.

Other topics are relatively easy to recognise, and their content covers a broad range of themes, reaching from personal and everyday topics such as health and leisure to political and societal discussions. Sports make up a relatively large percentage of tweets, especially football. The topic model generated 5 distinct topics[2], all dealing with this theme. A reason behind this can be ascribed to the Football World Championships. Days when games took place accounted for a huge increase of Twitter activity.

4.5.2 Sentiment

Assessing the sentiment scores of the messages was simpler, as the VADER library allows defining scores without the need for pre-processing. Scores were assigned to all English language tweets in the complete corpus. However, in further analysis only tweets were included which had non-neutral sentiment analysis results (having a neutrality score of not 1).

As described in the third chapter, the VADER library provides two kinds of indicators. The first are multidimensional metrics defining scores of the text being positive, negative, or neutral (with scores adding up to 1). The second is a "normalized, weighted composite score" (compound) which defines the sentiment of the text in a unidimensional way (Hutto 2018).

In the case study, only the compound score is used, as it is a standardised indicator and enables comparing different messages according to their sentiment and the strength of their sentiment. A value of -1 means a completely negative message, a value 1 a completely positive. The distribution of sentiment scores for selected topics is shown Figure 16.

As the Biterm Topic Model, the assignment of sentiment scores can also be tested on the companion website.

[2] Football players, Premier League, Football World Cup England, Football World Cup

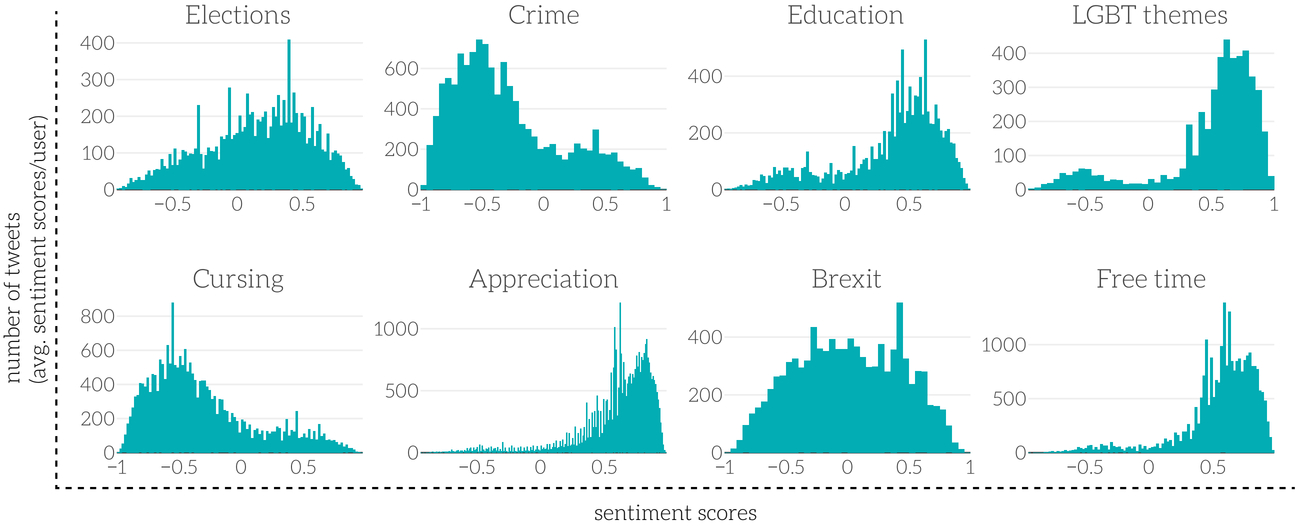


Figure 16: Sentiment scores for selected topics (own figure)

4.6. Internal Quality

Assessing the content of the downloaded Twitter messages doesn't necessarily lead to a simple answer to the question of their applicability in spatial planning. As working with such SMGI constitutes a very novel approach in planning, it is important to evaluate the quality of this data. There are several possible ways to define data quality, though in general most scholars (for example Devillers and Jeansoulin 2006b; Fonte et al. 2017) differentiate between internal and external quality.

Internal quality is commonly defined as a level of similarity between the data produced and the perfect data that should have been produced (Devillers and Jeansoulin 2006b, p. 37). This quality doesn't take specific needs of a certain domain into account. Therefore internal quality is in general easy to standardise. The ISO 19113 standard recommends five criteria as indicators for assessment. On the following pages, an internal quality assessment of the downloaded Twitter data is presented. Afterwards follows the assessment of the external quality. External quality refers to the usability of the data for a specific domain or question, which is in this case the reliability of Twitter messages as a source of data for urban planning.

4.6.1 Completeness

Completeness can be described as the "presence of absence of features, their attributes and relationships" (Devillers and Jeansoulin 2006b, p. 38). As Fonte et al. (2017, p. 144) point out, this is a "major concern" when dealing with VGI. VGI is strongly affected by participation biases, some areas tend to be overrepresented, while others may lack completely of information. This is especially true for Twitter, where cities account for a large share of the posted messages.

But these inequalities do not only occur when comparing cities to each other. Also accross different areas of single cities, tweeting activity shows large differences. Most of the messages are concentrated in the central districts but even when excluding the city core, there seems to be no correlation between tweeting activity and the population numbers.

Furthermore, as described in the third chapter, Twitter doesn't provide access to its complete set of data unless using a paid subscription. The filtering methods of the Streaming API are unclear and it is not known which rules are applied when deciding whether to send a piece of requested data or not.

The completeness of attributes is also diverging highly among features. As a tweet can possibly hold up to 150 attributes, it is highly unlikely that one message would have all attribute fields populated. The most basic attributes, the User ID, the message and a creation timestamp are always provided. Consecutively, connecting messages to specific users and timfeframes is alway possible. In contrast, coordinates were provided only in 9 per cent of the tweets of the case study, most Twitter messages don't provide information about the location they were created.

4.6.2 Logical Consistency

Logical consistency refers to the "degree of adherence to logical rules of data structure, attribution, and relationships" (Devillers and Jeansoulin 2006b, p. 38). The diversity of social media data makes it extremely hard to maintain a consistent data structure, as the content of messages may contain different elements, for example videos, photos, written in different languages by different people expressing different topics.

Twitter data possesses a robust structure presumably making the saving and accessing heterogeneous types of content efficiently possible. However, regarding spatial data, there are some crucial inconsistencies. As described in the third chapter (see *3.5.1 Location/Place/Coordinates – p. 40*) more precisely, Twitter makes a distinction between the location of users and tweets, as well between the correct coordinates of the messages and the tagged place they were posted from.

4.6.3 Positional Accuracy

In general, when using locational attributes of Twitter messages, the accuracy depends on the device used. Assuming most of the georeferenced tweets were created using a mobile phone, determining the user's position was done by applying (and combining) three approaches.

The highest precision is provided when using GPS but in case there is no adequate signal available, Cell Identifier or WLAN based positioning may also be used. This approach is called Assisted GPS (or A-GPS) and is used in most cases by mobile devices (Wang, Wong, and Kong 2012). Typically, GPS-enabled smartphones are able to detect the location with an accuracy of 4.9 metres (National Coordination Office for Space-Based Positioning 2017).

Of course, this accuracy is only valid when users chose to provide the coordinates of their devices directly (see *3.5.1 Location/Place/Coordinates – p. 40*). When a user choses to simply tag a location, their real position might be anywhere in the world. Of course, in many cases (for example when tagging a bar) it is likely that the user is at (or near to) the tagged place, but for a fine-grained analysis only coordinates with a high reliability are suitable.

The number of such tweets is very low. Of the 756 241 georeferenced tweets, 702 416 were potentially incorrectly referenced, only 53 825 didn't share the coordinates with other posts. The most tagged coordinate pair[3] (located near the Trafalgar Square) appears in 142 353 tweets (making up almost 19 per cent of all georeferenced messages).

From a technical point of view, tweets show a high positional accuracy. However, the possibility of geotagging a place doesn't guarantee that this accuracy is also reached in reality, reducing the correctness of locational information significantly.

4.6.4 Temporal Accuracy

Temporal accuracy is probably the most correct and consistent feature of Twitter data. When streaming, the requested data can be downloaded in real time, making information access immediately possible. The creation time of a Tweet (and also of a user profile) is always available, exactly to the second.

Of course, it cannot be guaranteed that users post in real time about the happenings (for example tweets regarding memories of past events make up the tenth most common topic), mostly the temporal distribution of tweets are in line with actual events or seem to be logically consistent.

4.6.5 Thematic Accuracy

Thematic accuracy constitutes a challenge to be measured as an internal quality of social media data. In "traditional" spatial data it refers to the "accuracy of quantitative attributes and the correctness of non-quantitative attributes and the classifications of attributes and their relationships" (Devillers and Jeansoulin 2006b, p. 38).

As the content is generated by users, it is challenging to identify the contained topics immediately, it requires further steps of processing. The content of the data (and its quality) therefore also coheres with the question it is expected to answer. For example when comparing the question of which languages are used in a certain area and which topics are being addressed, the same dataset might probably have different thematic accuracies. Therefore, thematic accuracy might in this case be regarded as an external quality and was assessed as such in the study.

[3] lon: - 0.12731805 | lat: 51.50711486

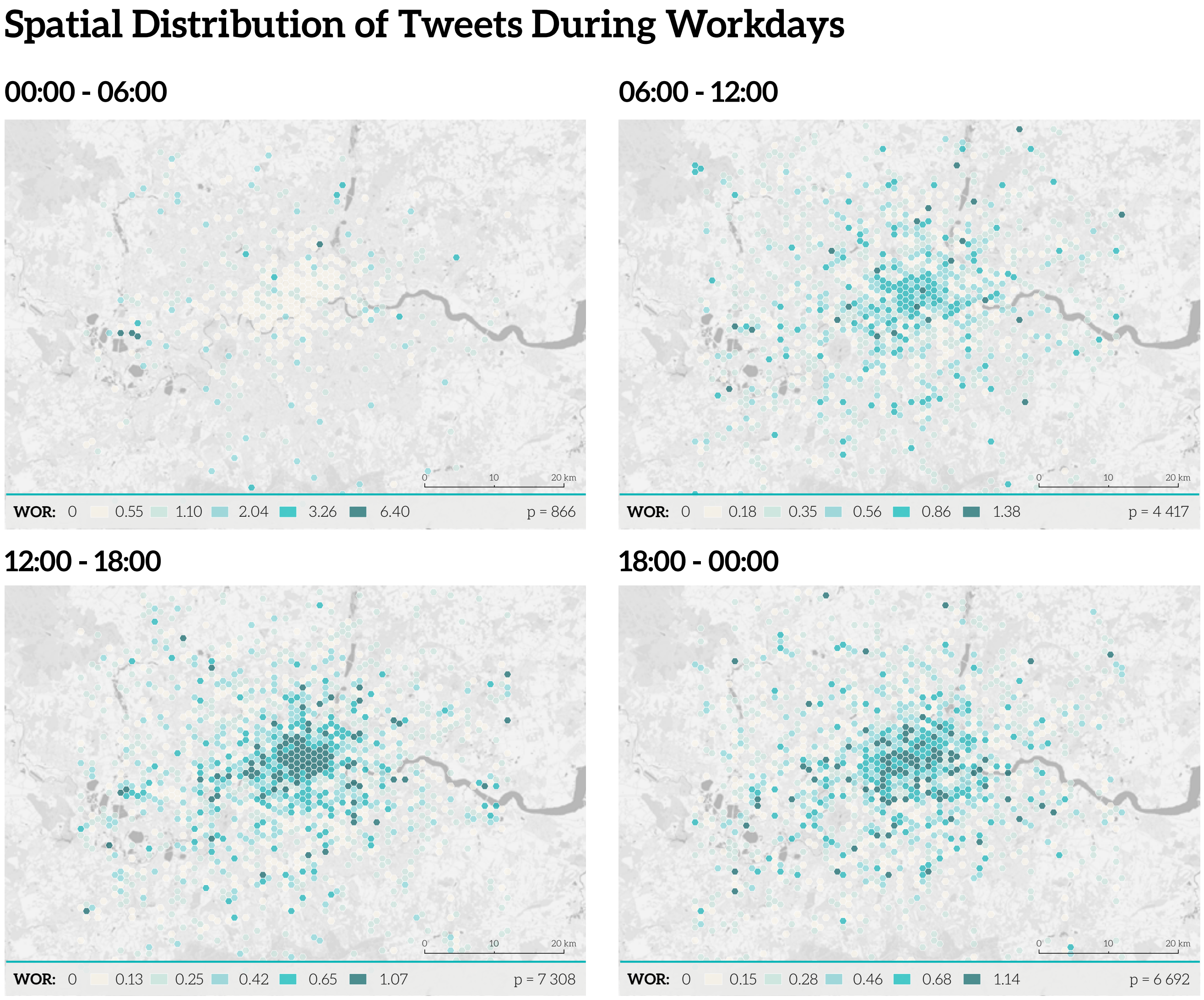
4.7. External Quality / Usability

In contrast to the internal quality, the external quality deals with the concrete usability of the data for concrete questions. Devillers and Jeansoulin (2006b, p. 39) define external quality with reference to the “level of concordance that exists between a product and user needs, or expectations, in a given context”.

This also means that a dataset with good internal quality doesn’t necessarily guarantee to become valuable for every possible question, but also a dataset with low internal quality can prove to be useful in a specific context.

The external quality in this case study refers to spatial planning and the applicability of Twitter data in that framework. More concretely, the data is regarded to be a source of information referring to indicators that are difficult to survey (thus quantify) or where the very high temporal accuracy of Twitter data promises to become the most beneficial.

Measuring the external quality of geographic data is by its definition harder to standardise and often also implies criteria that describe internal quality (Devillers and Jeansoulin 2006b, p. 39). For this case study, some of the data-based and socio-economic and demogra-


Spatial Distribution of Tweets During Workdays
00:00 - 06:00
0
10
20 km
WOR: 0 0.55 1.10 2.04 3.26 6.40
p = 866
06:00 - 12:00
0
10
20 km
WOR: 0 0.18 0.35 0.56 0.86 1.38
p = 4 417
12:00 - 18:00
0
10
20 km
WOR: 0 0.13 0.25 0.42 0.65 1.07
p = 7 308
18:00 - 00:00
0
10
20 km
WOR: 0 0.15 0.28 0.46 0.68 1.14
p = 6 692

Filter: only English-language tweets, 1 tweet/user/hour/polygon, only English-language tweets, no tweets in category ads/automatic content

r: 19 283 | tweets recorded 16 May - 23 Oct 2018 on Mo - Fr | Basemap: CartoDB

Figure 17: Spatial distribution of tweets during workdays (own figure)

phic approaches presented by Fonte et al. (2017) were applied. The quality of the data is assessed through validation of the three thesis through reliable (official) demographic indicators and the logical consistency of the dataset itself and its relation to external events and features.

4.7.1 Thesis I: Twitter activity as indicator

Twitter data may deliver information about the places people gather during different times. It enhances traditional AGI that mostly only contains people's night time residence, but lacks information about where they would spend their days. Also, differences during the course of a week may be recorded, indicating places that are more likely to be frequented during the weekend for example.

The patterns of distribution of people and activities allows to create models inferring these distributions. Consequently, such models can be implemented to predict activities in planned projects as done in the Madrid study (García-Palomares et al. 2018) for example.

Furthermore, tracking the distribution of tweets in real time may also to identify short-term increases in demand for public transport in cases of events such as concerts, gatherings, or demonstrations.

Tweeting activity is expected to represent logical expectations about the location of people (and tweets) at different timeframes. During the day and on workdays the activity is assumed to be concentrated in the more central districts of the city, while at night, people would tweet from their homes, distributed all across London.

The distribution of tweets reflects these assumptions partially. As Figure 17 shows, when compared to daytime, in the night there is a higher activity of Tweeting in the less central areas of the city, comparable to the findings of García-Palomares et al. (2018).

When looking at different days of the week, there is no strong dependence recognisable, Tweeting activity is concentrated in the same areas. This might be attributed to activities by tourists who are likely to spend their time where the most important sights are located.

As shown in Figure 18 patterns of activity also follow the expectations about tweeting activities in relation to land use categories (based on Corine Landcover units). Areas such as leisure and parks experience an about 50% higher Twitter activity on weekends than during workdays. These findings are also in line with the Madrid study.

When looking at specific events, the results become very heterogeneous. Tweeting activity is in many cases in line with the expectations but because of the small numbers of messages it cannot be regarded to be reliable. The amount of georeferenced tweets is too small to draw conclusions that can be regarded as being statistically significant.

Therefore, the first thesis might be accepted and tweeting activity can be used as an indicator about people's location during different times, but only in an aggregated form. When looking at some specific events, the very low number of tweets makes it impossible to draw valuable conclusions. Because this would promise to be a more valuable asset for spatial planning, its usability needs to be questioned.

4.7.2 Thesis II: Twitter topics as indicator

Social media data contains many themes/topics/aspects of our everyday lives that are in general challenging to capture with means of traditional methods. The semantic analysis of discussions on Twitter might help to discover spatial distribution patterns of certain topics that might be interesting for spatial planning. These topics include descriptions of social issues, for example crime, social controversies, or safety. The identification of the spatial (and temporal) distribution patterns of such topics

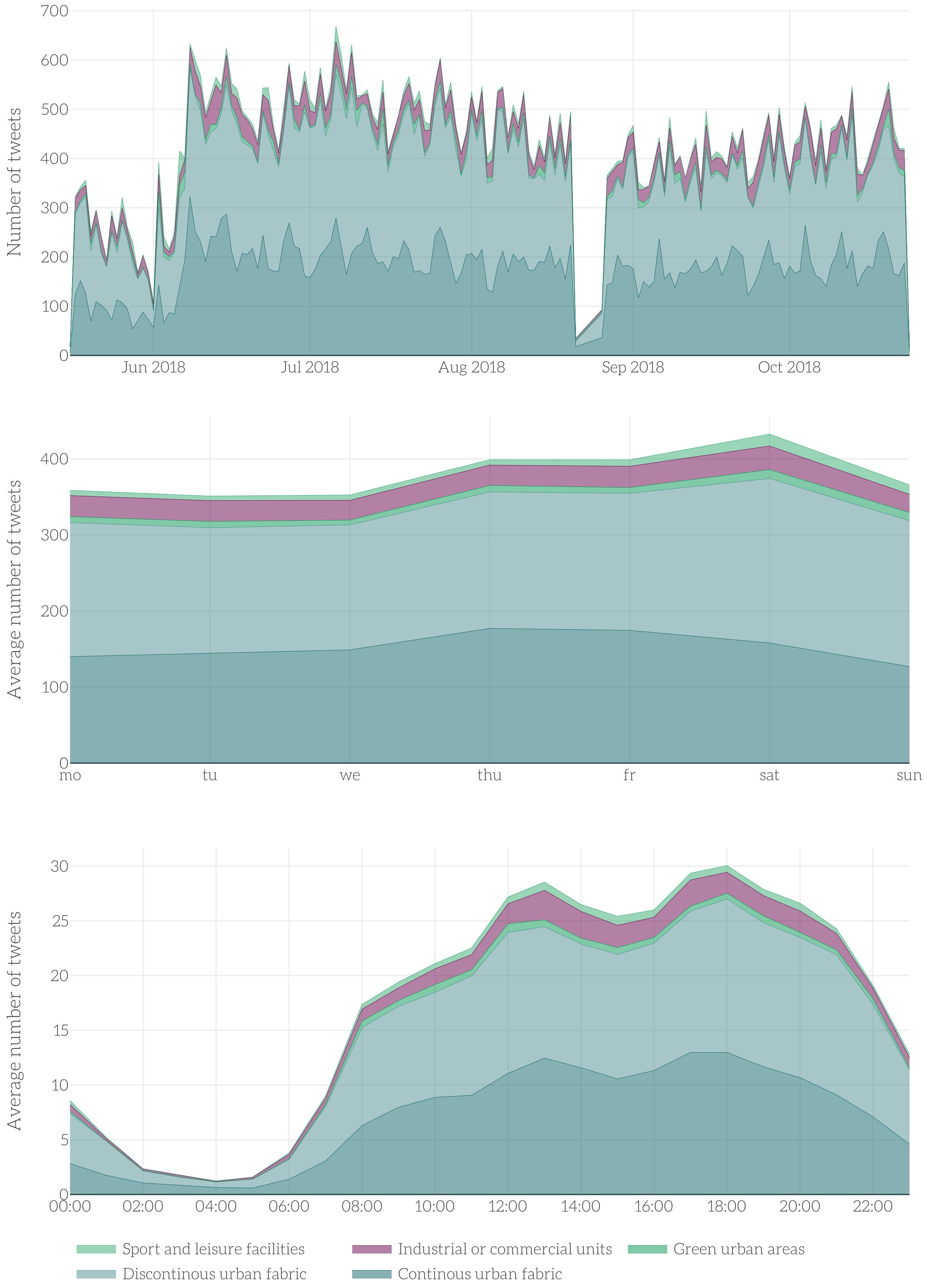


Figure 18: *Tweeting activity (only correct coordinates and 1 tweet/user/time unit) in selected corine categories per day, weekday, and hour (own figure)*

can be assumed to become a valuable asset for spatial planning.

Several studies dealt with the localisation of Twitter topics, among other cities also in London (Lansley and Longley 2016). The case study did in some (though not all) cases replicate the results of these projects.

A small number of topics, such as travelling, are indeed concentrated around transport hubs as airports and train stations but in many cases the topics are distributed over a (seemingly) random pattern. Even if analysing just the polygons with a large number of tweets, outliers tend to distort the distribution.

For a comparison, Figure 19 shows the distribution of Twitter corresponding topics in the case study compared to Lansley and Longley (2016). The distribution of topics reflects the built environment to a certain grade, although because of the small number of tweets the results of the case study have a little validity.

The problem of small numbers quickly becomes a challenge when dealing with the spatial distribution of topics. As an example there are only two hexagon cells in the whole study area having more than three tweets dealing with the topic crime with correct coordinates. In case taking also possibly incorrect coordinates into account, there are still only 45 polygons (and only 6 of them contain more than 10 tweets dealing with the topic).

Therefore, by aggregating multiple topics as an input for the analysis (in case of crime, the topics social controversies and social issues can be connected), it could be possible to enhance the number of relevant tweets in the polygon. However, in many cases it is still not possible to reach a tweet count of more than 10 or 12.

Of course it is also possible to generate larger aggregation areas. By doing so (and taking the area of London's boroughs as reference) a sufficient number of tweets in a given area becomes easier to reach, the scale gets too large to draw valuable conclusions.

As an alternative way of calculation it is also possible to analyse the distribution of the tweets by their distance to certain amenities. For this research, different amenities[4] in London wer queried from OSM and the distance of the tweets to the nearest polygon was calculated. As Figure 20 suggests, in lots of cases topics are located closer to some corresponding amenities:

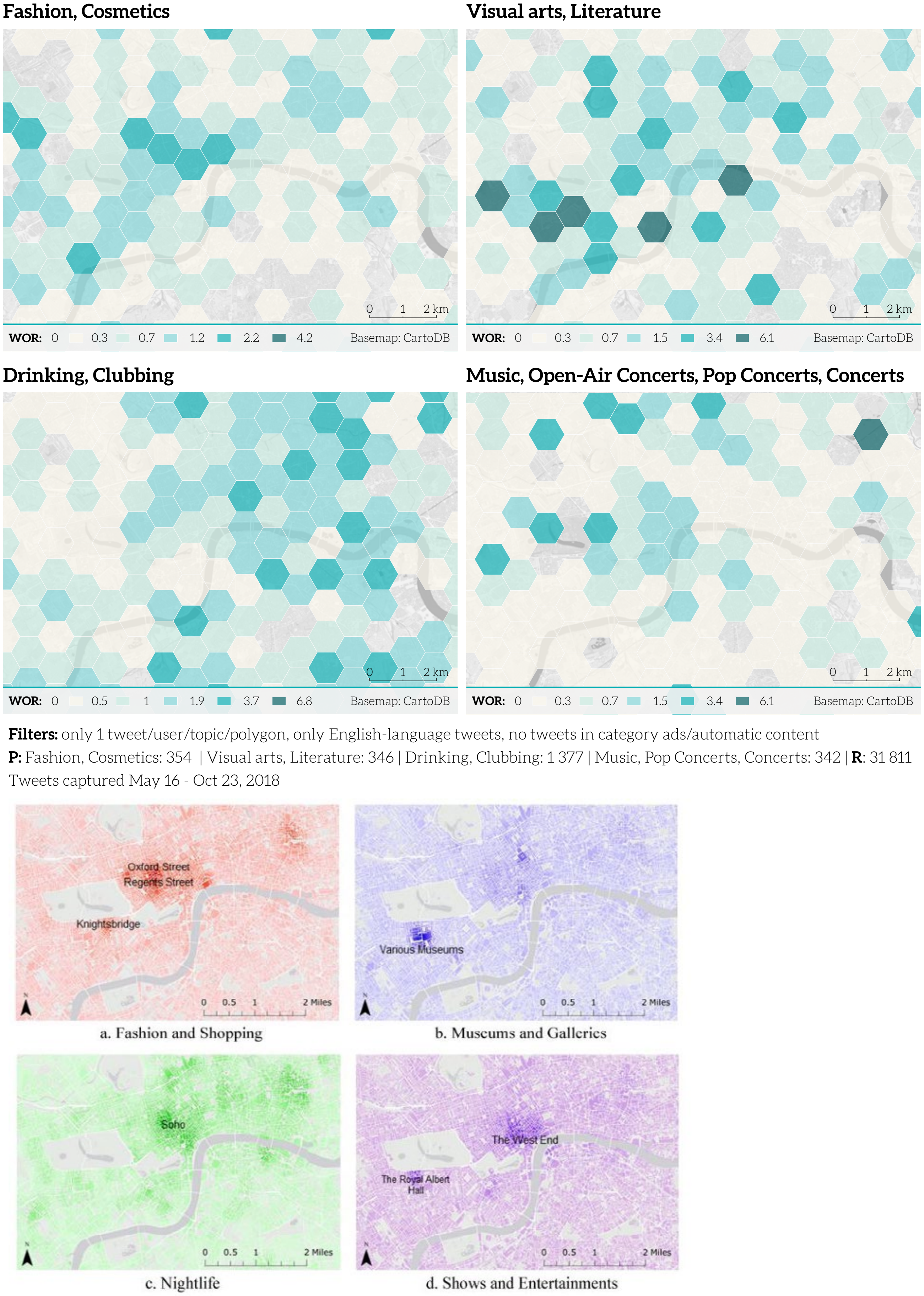
Here, it is also important to mention that this approach also only works when there are enough tweets available and it is highly unlikely to set particular "epicentres" for a certain social phenomenon. Therefore, the spatial distribution of topics can be related to some (physical) real-world phenomena. This means in consequence however that such information probably doesn't provide much novel knowledge. It is interesting to see that there is an increased tweeting activity about drinking near bars but this information is highly unlikely to become valuable for spatial planning. The location of bars or different amenities can be simply spotted on a map.

When taking the temporal distribution of tweets into account, the appearances of certain topics are linked to real-life events. A good example are the Football World Championships (June 15th to July 14th 2018), as shown in Figure 21. Days when matches took place led to an increased volume of tweets containing respective topics on Twitter. Also other sports events, like tennis games or boxing matches are visible in the trends.

Some cultural events lead to an increased activity too, for example concerts, the Notting Hill Carnival[5] or the London Pride. The same is true for politics and socially relevant events,

[4] The map of the downloaded amenities can be found in the annex (*Queried Amenities – p. 105)*

[5] Here it is important to note that on the first 2 days of the carnival there were no tweets recorded


Fashion, Cosmetics
0 1 2 km
WOR: 0 0.3 0.7 1.2 2.2 4.2 Basemap: CartoDB
Visual arts, Literature
0 1 2 km
WOR: 0 0.3 0.7 1.5 3.4 6.1 Basemap: CartoDB
Drinking, Clubbing
0 1 2 km
WOR: 0 0.5 1 1.9 3.7 6.8 Basemap: CartoDB
Music, Open-Air Concerts, Pop Concerts, Concerts
0 1 2 km
WOR: 0 0.3 0.7 1.5 3.4 6.1 Basemap: CartoDB

Filters: only 1 tweet/user/topic/polygon, only English-language tweets, no tweets in category ads/automatic content
P: Fashion, Cosmetics: 354 | Visual arts, Literature: 346 | Drinking, Clubbing: 1 377 | Music, Pop Concerts, Concerts: 342 | **R**: 31 811
Tweets captured May 16 - Oct 23, 2018

a. Fashion and Shopping

b. Museums and Galleries

c. Nightlife

d. Shows and Entertainments

Figure 19: Weighted Odds Ratio for selected topics in central London (own figure) as compared to Lansley and Longley 2016, p. 93

Trump's visit in the United Kingdom and the People's Vote Protest March are clearly visible.

However, not every notable news event seems to appear on Twitter. Some major topics of the summer, such as the cave rescue of the 12 schoolchildren in Thailand or the collapse of the Morandi-bridge in Italy don't appear clearly in these statistics.

When taking the general temporal distribution over workdays into account, the results meet the expectations in general. Topics linked to culture or leisure are more frequent on weekends, whereas topics regarding work or business appear more often during the week. Noticeably skewed distributions can mostly be explained by outliers of specific individual events, for example the visit of Donald Trump, which took place on a Friday or the broadcast dates of ITV's reality Love Island.

Comparing the distributions of single topics during the course of one day, there are no large differences observable. In general, most topics have their peaks sometime around the late afternoon or in the evening, while in the night between 1:00 and 6:00 there are only a very few tweets published. Still, some specific events, such as football games or Love Island, have noticeable peaks around the times these events happen.

The value of knowledge extractable from the temporal distribution of tweets is comparable with that of the spatial distribution. The fre-

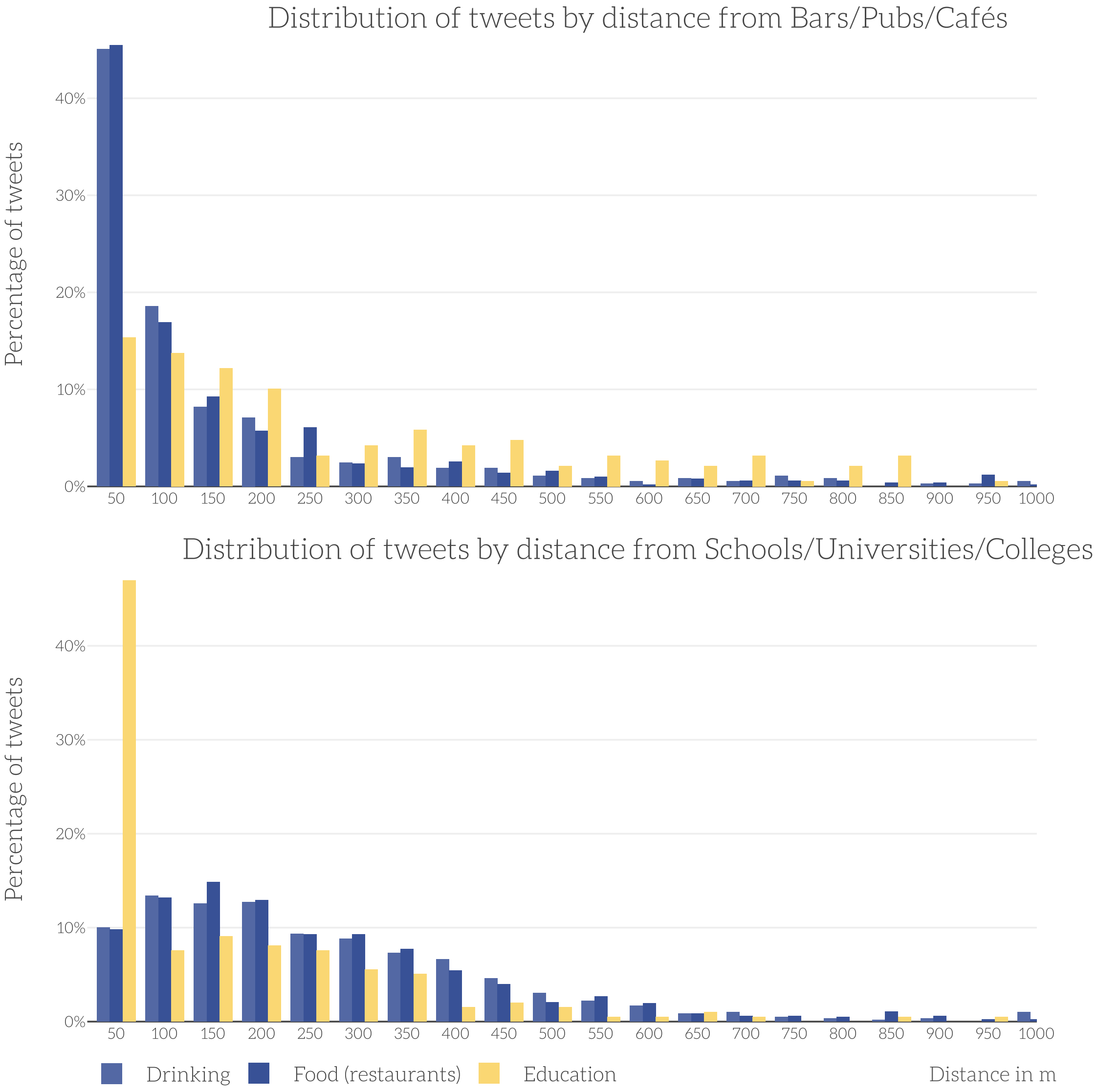


Figure 20: Distribution of tweets by distance from selected amenities (own figure)

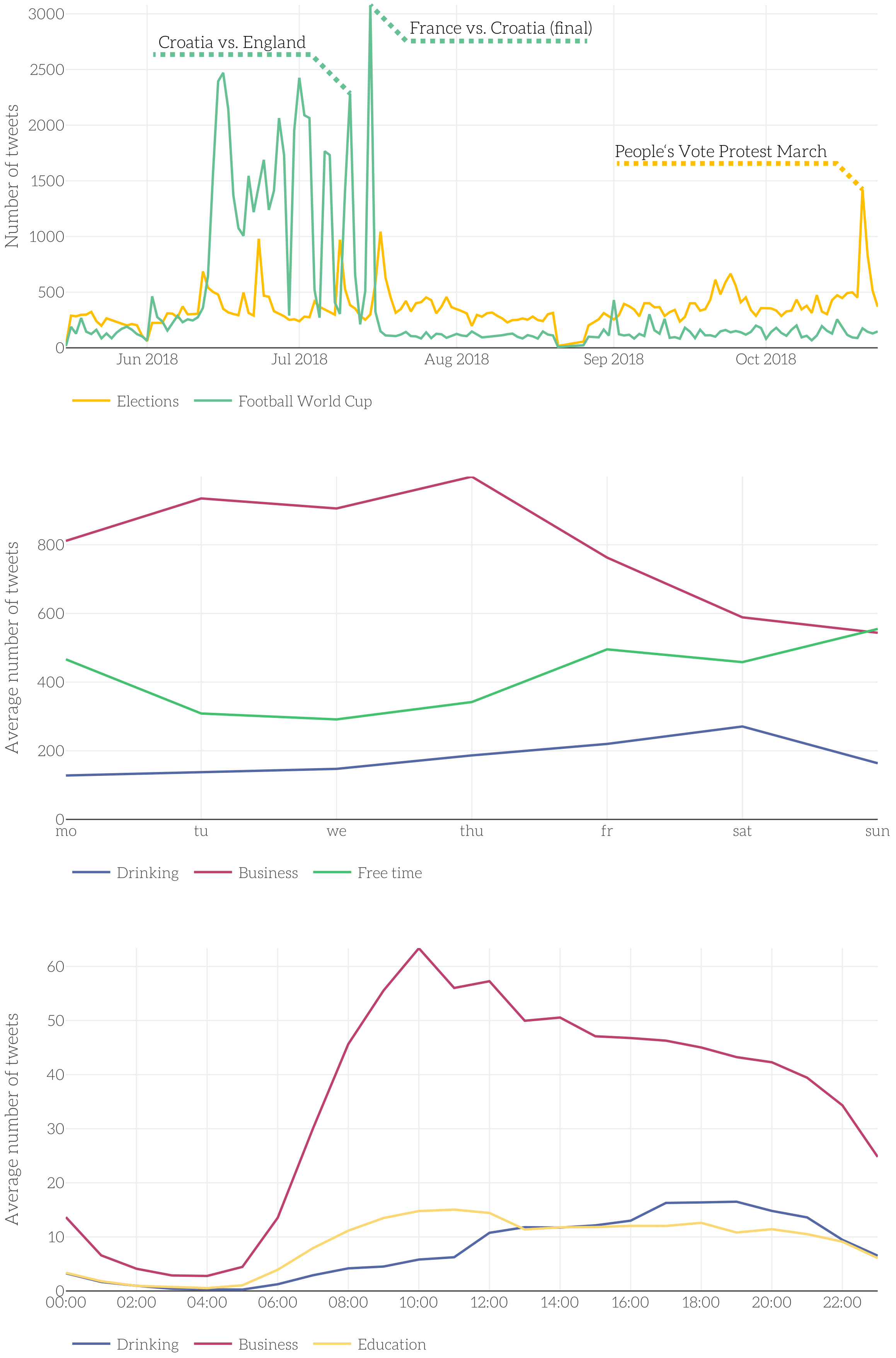


Figure 21: Temporal distribution for selected topics per day, weekday, and hour (Dates: FIFA 2018, People's Vote 2018, own figure)

quency of Twitter topics depicts the temporal happenings of relevant phenomena, still this knowledge doesn't carry much value. It is logical that people will talk about Donald Trump on the day he visits London but this information is almost completely irrelevant for the domain of spatial planning.

An interesting (but because of the lack of data in this case very theoretical) question could be the focus on the spatial distribution of tweets linked to a specific event. There are some studies that follow Twitter activity in case of a catastrophe. A study depicted the effects of Hurricane Sandy on the New York Twitter community (Shelton et al. 2014) for instance.

The case study could also identify some interesting patterns. For example, the demonstration march (People's Vote March) aiming for a second Brexit-referendum could also be captured on Twitter. The areas along the protest's route accounted for a 20-times higher activity of Twitter users talking about either Brexit or elections. Figure 22 shows the increased tweeting activity about these topics (as compared with other days) along the route of the protst march.

Still, the small number of tweets makes it in this case also impossible to draw some more valuable conclusions, especially not knowledge that promises to become for spatial planning.

Another possibility of checking the validity of tweets as indicators is by comparing them to administrative data. For this task, tweets (for example posts containing the topic crime) were aggregated to the level of wards and the frequency of topics was correlated to corresponding statistics (for example crime rate).

The results of the calculations generally don't meet the expectations. In case of comparing the

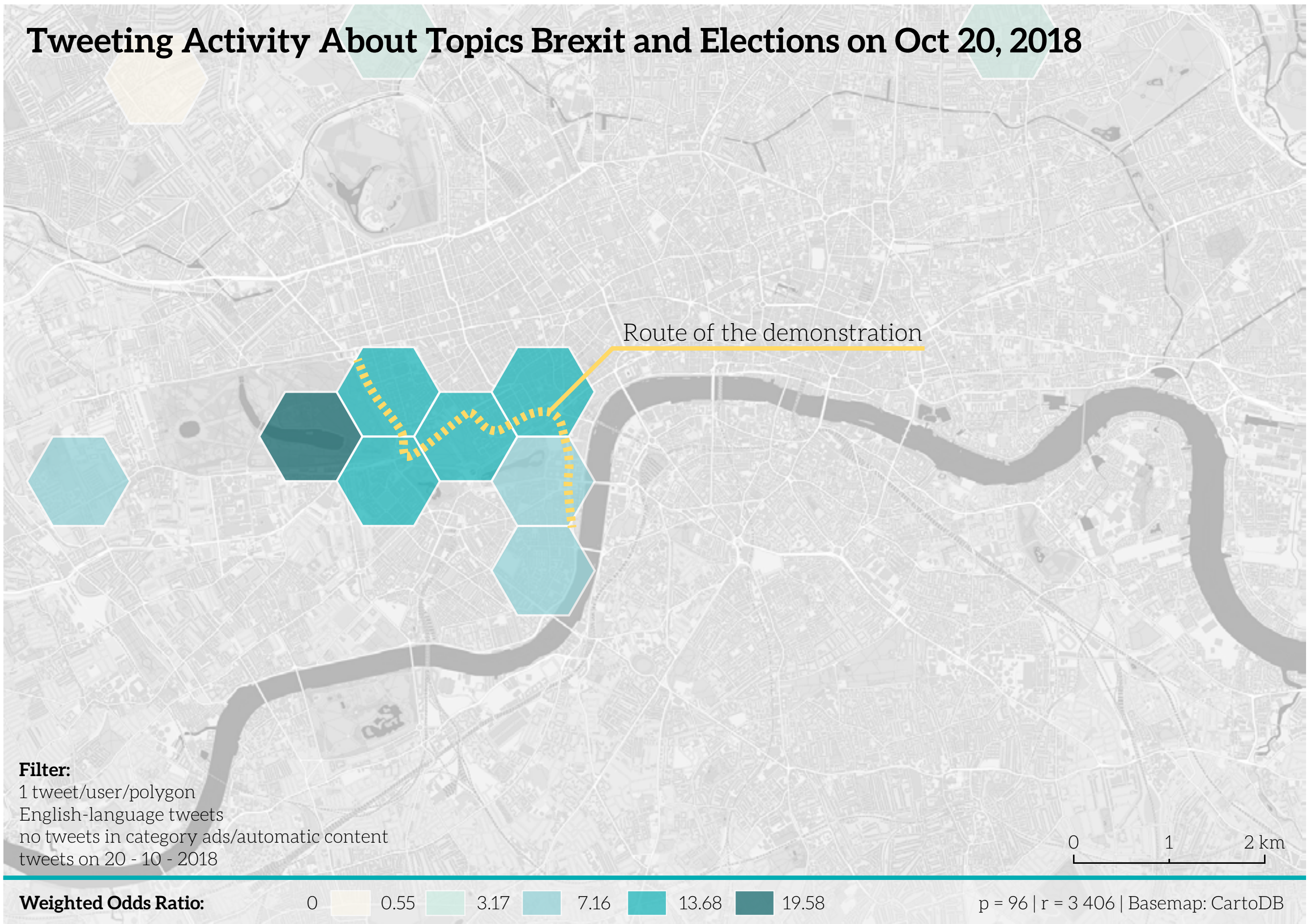

Tweeting Activity About Topics Brexit and Elections on Oct 20, 2018
Route of the demonstration
Filter:
1 tweet/user/polygon
English-language tweets
no tweets in category ads/automatic content
tweets on 20 - 10 - 2018
0
1
2 km
Weighted Odds Ratio:
0
0.55
3.17
7.16
13.68
19.58
p = 96 | r = 3 406 | Basemap: CartoDB

Figure 22: Tweeting activity about topics „Brexit" and „Elections" on the day of People's Vote Protest March (Oct 20, 2018), compared to the demonstration route (own figure - demonstration route: People's Vote 2018)

number (or the other calculation methods – relative numbers, Odds Ratio, Weighted Odds Ratio) of tweets to the average well-being score of the people living in that ward (Figure 23), no connection seems to appear.

Occasionally these results are even outright contradictory. The comparison of the odds ratio of tweets containing the topic crime with the ward-level crime rate results in a negative (even though insignificant) correlation. Thus, the model predicts that a lower number of tweets about crime indicates a higher crime rate.

Therefore, unfortunately the topics of tweets cannot be viewed as a reliable and valuable source of information for spatial planning and the second thesis cannot be accepted.

4.7.3 Thesis III: Sentiment as Indicator

The underlying information about an opinion or a sentiment in a tweet might constitute a valuable enhancement. As Mitchell et al. (2013) state, the happiness scores of tweets might be used as indicators about the general well-being of people. According to their results there is a direct correlation of the average happiness scores of tweets in an area with multiple relevant statistical indicators (Gun Violence, Peace Index containing multiple crime indicators, America's Health Ranking, BRFSS life satisfaction score).

The usage of tweets as indicators for peoples' well-being promises to provide relevant information in a very short time and might constitute an easy alternative to traditional surveys. The indicators can also be linked to certain topics, to get not only the sentiment of the users and their messages in general, but also with regards to some specific topics and themes.

It is important to note that although the basic assumptions of this thesis are in line with those of the referenced well-being study, the methodology differs in many ways. Mitchell et al. (2013, p. 2) used a general sentiment lexicon with predefined happiness scores for each word. For example the word "rainbow" has a score of 8.1 (on a scale from 1 – sad to 9 – happy) regardless to the context it appears in. Then, the sentiment score of a message is calculated by aggregating the distinct sentiment scores

Because of its focus on working with social media texts and its capability of recognising the context of messages into account, the case stu-

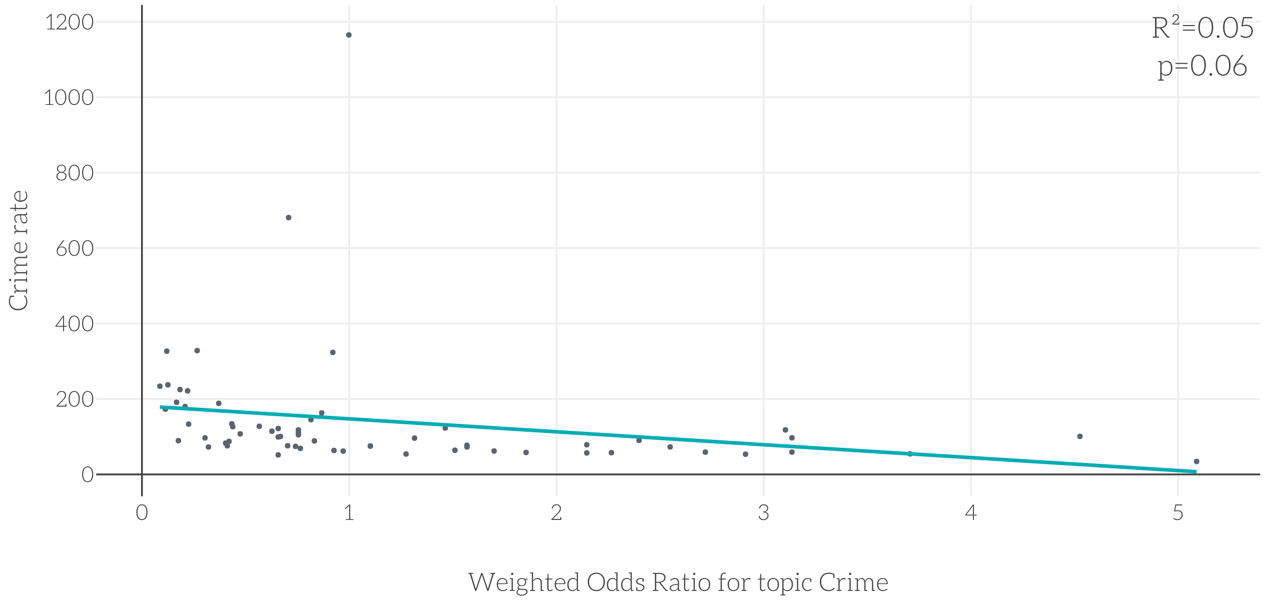


Figure 23: Crime rate per ward (Greater London Authority 2015) as compared to tweeting activity about topic „Crime" (only correct coordinates, 1 tweet/user/ward) (own figure)

dy used the sentiment detection algorithm VADER. VADER identifies less opinion words (for example “rainbow” is detected to be completely neutral), but doesn’t lack context-awareness.

Furthermore, the scale of the study is also different, Mitchell et al. (2013) compared the overall sentiment scores of distinct cities of the USA to each other while this case study compares distinct wards of London as a single city.

Analysing the average sentiment scores of the 100 captured topics in general leads to plausible results. Topics expected to have an obvious sentiment score (cursing and crime can expected to carry more negative, while acknowledgements/gratitude and appreciation more positive sentiments) are indeed perceived according to the expectations. Of course it is more interesting to look at the topics that are not expected to have positive or negative scores.

For example the topic railways tends to carry more negative sentiment scores. For this question it would be interesting to analyse the spatial distribution of sentiment scores to this topic. Doing so might help to recognise concrete places where intervention is needed.

Unfortunately, the small amount of tweets make a reliable analysis of tweets hardly possible. Figure 24 shows the average sentiment scores of the topic railways. Though some noticeable differences between single polgons do appear, the results are very sparse and cannot be regarded to be significant or reliable.

As done by Mitchell et al. (2013) and for the previous thesis in this case study, the sentiment scores can also be correlated with the existing well-being indicators. However, there seems to be no correlation between the overall sentiment of Twitter messages in London‘s wards

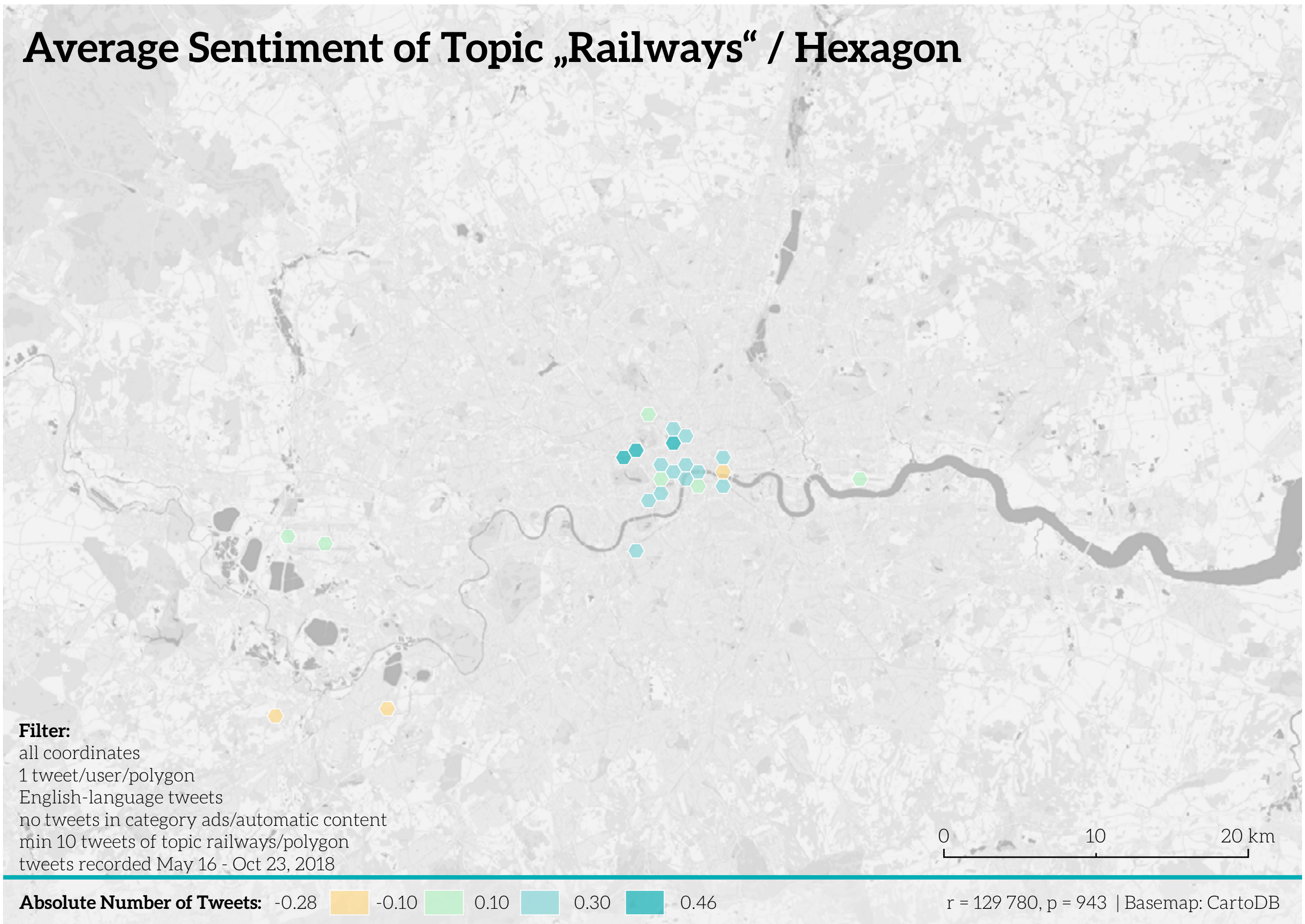

Average Sentiment of Topic „Railways“ / Hexagon
Filter:
all coordinates
1 tweet/user/polygon
English-language tweets
no tweets in category ads/automatic content
min 10 tweets of topic railways/polygon
tweets recorded May 16 - Oct 23, 2018
0
10
20 km
Absolute Number of Tweets: -0.28
-0.10
0.10
0.30
0.46
r = 129 780, p = 943 | Basemap: CartoDB

Figure 24: Average sentiment of tweets dealing with the toic „railways“ (own figure)

and the corresponding statistical indicators (Figure 25), but also other indicators don’t appear to be linked to the scores.

Therefore, also the third thesis needs to be rejected, the sentiment scores of the tweets failed to indicate the overall sentiment of the wards of the city. However, when looking at single topics, they might lead to relevant and interesting conclusions but to prove this assumption, more data would be needed.

4.7.4 Overall External Quality of Twitter Data

As the downloaded data failed to confirm two of thesis sketched above completely and validated the first one also only partially, it can be concluded that tweets didn’t prove to be a valuable asset in spatial planning. Though in many cases it is possible to relate Twitter data to the real world, this is only the case when dealing with phenomena that are anyways easily measurable, mappable, and/or relatively obvious. Topics and questions being potentially valuable for spatial planning are not allowing to draw reliable, reliable and valid conclusions.

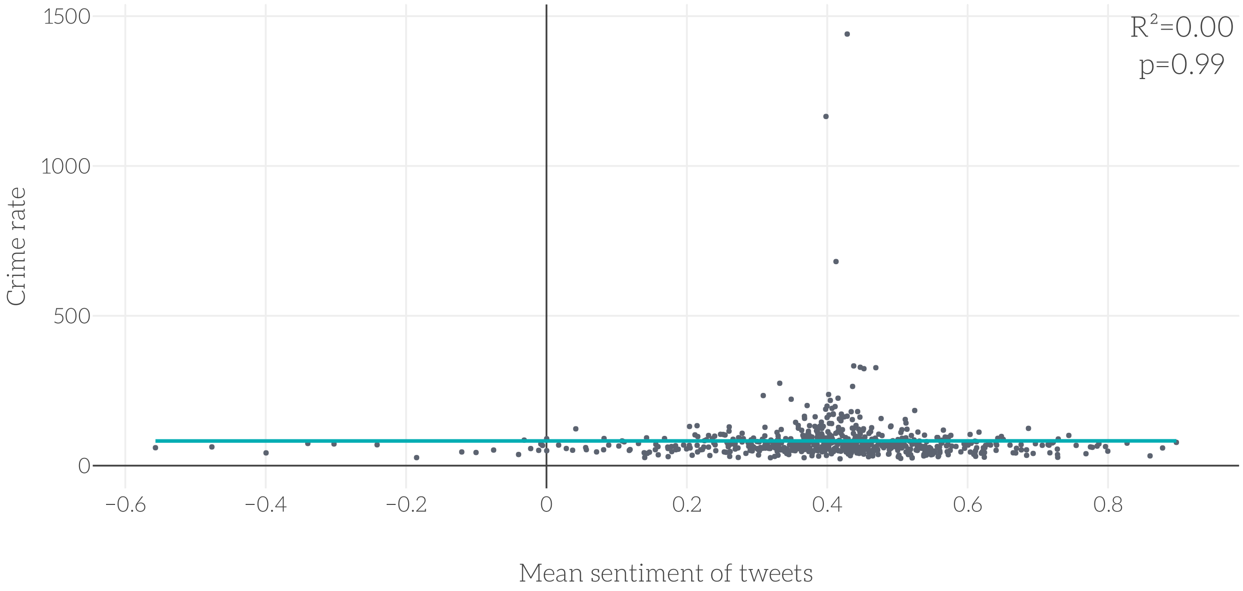


Figure 25: Average sentiment of tweets/ward (only correct coordinates, mean sentiment of tweets/user - own figure) as compared to the crime rate (Greater London Authority 2015)

5. Conclusions

It is difficult to overstate the impact recent technological development exerts on our lives and our cities. Big Data and the surrounding technologies are impressive. Nowadays, data is created and transferred in an amount and speed that was never seen before in history of humanity. This of course also shapes the ways we look at the amazingly complex systems of our cites. The abundant availability of data covers all possible domains and aspects of this system and promises to help us to get insight into these perplexing processes. Nowadays it is possible to access data in real time while sophisticated methodologies enable to process it quickly, efficiently, and flexibly.

Of course, Big Data is not only used to describe the state-of-the art, it can also be used to predict the future. Some authors, such as Andersen (2008) even go that far that Big Data can be used as the sole basis for our decisions. "With enough data, the numbers speak for themselves", is probably the most prominent and provocative statement in his article.

Although his position is rather extreme, the excitement surrounding Big Data has also found its ways into the cities we live in. The emerging processes and technologies led to a phenomenon which Kitchin (2016) calls data-driven urbanism. Big Data promises to shape and form the cities of tomorrow. Under certain circumstances it can even become the determinative factor planning decisions are based on. Therefore it quickly becomes crucial for planners to deal with questions that come up when applying these new technologies.

Most importantly, planners need to assess the values, potentials, but most importantly the shortcomings of Big Data critically. It is commonly assumed that Big Data also contains a big amount of information. However, this assumption has proven to be false in many cases, including the case study of this project. The data needs to be refined, filtered, and sorted out before it is possible to extract information.

Of course not only planners and organisational departments of cities are affected by these technological developments. New devices and technologies have also transformed the way geographic data is created. Even a few years ago, only cartographers were able and capable of mapping our cities and countries. Today, everybody with a mobile phone can create and disseminate geographic data. Lots of initiatives (most prominently OpenStreetMap) have appeared where people can map their surroundings, but geographic data is nowadays also created possibly unknowingly.

By creating a tweet for example, we do not only post a message on the internet but also contribute information about our location, the time we visited that location, and of course a topic we found at that time and place relevant. The aggregation of such data might therefore lead us to knowledge about people's subjective impressions and feelings towards specific topics at specific places. In consequence, this leads to valuable information that might help us to turn our cities into better places.

Of course Social Media Geograhic Information incorporates many challenges of Big Data. It is unstructured, varying in quality, and cannot be processed by traditional means and methods. Furthermore, it is heavily biased, not validated, and therefore it can hardly be regarded to be representative.

Still, its ubiquitous and quick availability, its flexibleness regarding time, space and content make it very promising for planners. A number of studies had set the focus on different aspects of urban life through the lens of social media analysis. It is possible to see where people are at different times (García-Palomares et al. 2018), access information about spatial segregation

(Lamanna et al. 2018), find out what people talk about at certain places (Lansley and Longley 2016), get information about quality of life in cities (Mitchell et al. 2013), and many more. There seems to be no question where SMGI wouldn't be able to provide an answer.

On the other hand, many researchers (for instance Field 2013; Poorthuis et al. 2014) argue for viewing such data critically. As stated above, SMGI is heavily biased, not validated, contains possibly lots of errors, as well as irrelevant or misleading information. Most importantly it wasn't intended to provide the knowledge we expect to access from it.

Therefore, an important motivation for this master's thesis was to find answers to questions about the usability and value of SMGI for spatial planning. SMGI was considered as a source of information about people's location, discussions, and sentiments. To assess these questions, a case study was conducted, based on more than 8.3 million tweets recorded in London from May to October 2018.

As most scholars agree (for example Crampton et al. 2013; Venturini et al. 2017), the raw data is too coarse to deliver any valuable information. It needs to be pre-processed first. The information extraction approaches are generally novel for spatial planners. Most of analysis methods are borrowed from other domains, especially computational linguistics and Natural Language Processing. These two domains have developed means and methods that enable to identify the content and the sentiment of texts (in this case tweets) automatically. As these factors are highly valuable for urban planning, new methodologies can be assumed to be welcomed in the planning community.

5.1. Lessons Learnt

Maybe because of large expectations, the results can be described as sobering. Although the amount of over 8 million downloaded tweets is indeed impressive, only a very small fraction of these tweets proved to be suitable for spatial analysis on a lower level. Only a small percentage of the messages is georeferenced, many of those have potentially incorrect location information.

When looking at distinct users it is clearly visible that a very small fraction of them created a very large share of the content. Furthermore, of course, it can be assumed that the users didn't compose their tweets with the intention of providing valuable information for urban planners.

Topic modelling has shown that Twitter is primarily a platform for informal communication, mainly focusing on themes that are probably irrelevant for (or not directly addressable by) urban planning. For example this means that the whole dataset contained only 79 correctly georeferenced tweets dealing with the topic crime. Therefore this small number also didn't allow to make statistically significant comparisons with the validated and reliable crime rate indicator.

The small number of relevant messages cannot be completely attributed to lost information during the analysis process. The single steps provided satisfactory results. The Biterm Topic Model defined topics accurately and assigned the single tweets correctly. The VADER sentiment analyser also captured the sentiment of the messages properly in general.

The case study used Twitter's complimentary "Streaming API" in order to access the data. This means that only a filtered sample of relevant tweets was provided for analysis. Still, it can't be assumed that the costly "Decahose API" or the even more expensive "Firehose API" would have delivered better results. According to Morstattel et al. (2013) when working with a bounding box (as done in this case study) the "Streaming API" returns the almost complete set of tweets nevertheless.

Even when looking at some more general indicators (and not splitting the anyhow very small dataset into more pieces) thus enhancing its reliability by using higher numbers, Twitter data still lacks validity. Missing metadata about the users' demographic indicators means that we don't know what people generated the downloaded data. Surveys and research projects have indicated that Twitter's user base consists mostly of younger, better educated people living and tweeting mostly in cities. Still, no more concrete and finer information about the portal's users is available.

When assessing Twitter data on its internal quality, several shortcomings appear. By its nature the dataset was often very inconsistent, especially when it comes to location information. The three positional attributes (the user's location, the tweet's location and its coordinates) are in many cases contradictory and ambiguous. Although the bounding box defined a very concrete focus area, messages were still downloaded from places as far as Chile, Australia, or Mauritius.

Furthermore, even though mobile devices allow us to define our location within an accuracy of about 5 metres, in case users geotag their position their real location might be virtually anywhere in the world.

Comparing the spatial and temporal distribution of tweets and their topics to real world phenomena, the results are very diverging. In many cases tweets are able to depict the world accurately. For example, results of the case study indicated that there is a concentration of tweets about education around schools universities, and colleges. The same is true for the topic drinking and bars, pubs, and cafés.

Real world events also tend to appear in the data. Evenings with England playing during the football world cup are clearly visible in Twitter trends, as well as Donald Trump's visit or an increased frequency of the topic drinking on Fridays and Saturdays.

Unfortunately even though these results seem interesting indeed, it is hard to argue that they would carry a high value for spatial planners. Topics possibly being more relevant, those covering social issues and controversies don't represent real-world (and validated) indicators. Regardless which topics are compared to which indicators (no matter which calculation or filtering methods are used), there is no statistically significant and valid correlation appearing. Likewise, also sentiment results are not in line with any of the indicators evaluating the quality of life.

There are several ways to overcome the issue of small numbers. A shortcoming of this case study was the relatively short period of data collection. The referenced research projects recorded tweets for one or sometimes even multiple years. By prolonging the duration of data sampling, a higher number of tweets might possibly facilitate drawing representative conclusions. However, as sketched in the second chapter, and also became apparent in the case study, a larger dataset doesn't necessarily mean more information. Furthermore the need for months or years long duration for data sampling contradicts the assumed temporal flexibility and quick availability of Big Data.

In conclusion the case study indicated that the external quality of Twitter data as a source of information for spatial planners is rather low. It depicts some phenomena quite accurately, but seems to fail in case of concentrating on not so obvious questions.

Referring back to the theoretical discussions it became clear that numbers can't speak for themselves. Or, more accurately, they can, as it was possible to extract a really appealing amount of topics and indicators. Still, the results are in many (and most importantly in the most relevant cases) misleading.

The novel chances of digital research sketched by researchers of the Sciences Po médialab (Venturini et al. 2017) are to some extent also applicable to the findings of the case study. The downloaded dataset signalised a really immense breadth and depth regarding its content. It contained information not only about the time, place, and location of users but also about a number of different topics, sentiments, and much more information could have been extracted when focusing on different assets.

The borders between qualitative and quantitative weren't dissolved completely. The methodoloogy in this case can be characterised more by analysing qualitative assets through quantitative means. All the methodologies of extracting information from the texts underlie a relatively clear quantitative, probabilistic approach.

Similarly, although technically it was possible to aggregate the levels of analysis very flexibly, on a small scale results are negatively affected by the sparsity problem of Twitter data, while on a large scale, results tend to become too generic for concrete analysis questions. Furthermore, because of the uncertainty in location information, the level of aggregation becomes quickly too inaccurate. As shown above, when downloading tweets (supposedly being located in London), also messages from all around the world appear.

5.2. Further Steps

It is important to note that these findings are only true for this concrete case study. The flexible availability of data allows also drawing the scale on a higher level. A comparison of distinct cities with each other might potentially yield better results. There, the problem of inaccurate coordinates might become easy to tackle when it is irrelevant where a Tweet is located inside a city.

Still, many methods that were applied in the case study, might become relevant in spatial planning. Most importantly topic modelling, which can be used not only for language processing, but also in cases where there is a need for creating a model about a phenomenon linked to processes that are not directly measurable. It can be used for modelling land use (Rimal, Zhang, Keshtkar, Wang, and Lin 2017) and similar domains, such as energy or transport planning.

Generally speaking, the most important finding of this study was the need for reflecting phenomenon of Big Data critically. Despite the large amounts of data generated constantly, for a specific question, only a small fraction of these are relevant. The amount of information doesn't say anything about its quality, therefore it becomes important not to be "blended" by large numbers. The focus should be laid on the quality of the data, as well as is suitability for answering a certain research question, and its overall value combining these two factors.

5.2.1 The Perfect Data

The evaluation of the downloaded Twitter messages has signalised some serious flaws in their data quality. The probably most crucial aspect for planners, locational quality, doesn't meet the most basic requirements. Furthermore, there is no information about the users available apart from their usernames and some basic attributes. As planning makes concrete decisions, such metadata is often crucial to get a reliable and valid picture of a certain situation.

The lack of metadata could be compensated through other qualities. For example it would be irrelevant to know which genders the recorded users have if the data revealed information about the general sentiment towards a topic in a general neighbourhood. Then, the missing data could be supplemented by traditional methods, such as doing questionnaires or conducting interviews. Twitter data still doesn't possess enough other qualities that could possibly counterbalance these shortcomings.

Of course this doesn't mean that Twitter data is worthless. It is just not suitable for the domain of urban planning. The provided attributes are very valuable for marketing, for example. The real-time streaming of data allows the tracking of certain company (or offer) related keywords, the NLP-techniques presented in this work enable the assessment messages on their content and sentiment. In combination with other APIs, companies have the tools to implement, run, and monitor successful marketing campaigns.

However, when it comes to spatial planning, Twitter is too broad for becoming a valuable asset. Even if ignoring the issues surrounding the localisation of the messages, just the content of tweets is generally speaking often irrelevant for planners. Although informal communication or sports should not be completely ignored as they are also parts of a complex urban system, these are topics where planning exercises only little influence.

5.2.2 The Best Fitting Data

Alternatively, many other sources and companies exist that could possibly provide more specific (and accurate) urban data. Airbnb knows a lot about tourists and other visitors spending time in our cities, Foodora and Deliveroo see the best and fastest biking routes in the cities by tracking the couriers during deliveries. Various bike and scooter-sharing companies also record accurate data about movements of people. So do mobile virtual network operators, and the number of organisations and companies dealing with data generation and analysis doesn't seem to stop growing.

Although in this project the quality of data of these companies was not assessed, it can be assumed that they have access to more accurate and reliable location data, as it is one of the basic factors for their success. For Twitter, in contrast, it is sufficient to know the approximate location of its users, but if they are a few metres farther, or maybe in the next neighbourhood, it doesn't make a difference.

In contrast, the car-sharing business Uber needs to know the accurate location both of its customers as their drivers. The company is therefore able to track the accurate location of its drivers and customers accurately in real-time. In addition, Uber must also have information about road and traffic conditions, redirections, and a lot of more factors that are relevant for operating this ride-sharing service. This means, when assessing people's mobility behaviour it can be expected that Uber can provide a dataset that is big and suitable enough for more concrete planning tasks.

5.2.3 The Valuable Data

In fact, Uber offers access to its data for city officials and planners. The company set up a website called Uber Movement containing the data for 2 billion trips made in a number of selected cities around the world in an aggregated form. The dataset, which contains also trips in London, delivers information about the GPS-tracked routes the customers of the company take (Forrest 2018).

In order to present the value of their dataset the company conducted several research projects in different cities. They used the data for the assessment of traffic flows following floods in Nairobi (Uber Movement Team 2018b), during the Hindu festival Dhanteras in Delhi (Uber Movement Team 2018c), or during the closure of the Tower Bridge in London (Uber Movement Team 2018a).

For traffic planning such data promises to become extremely valuable. Uber's data may be expected to exhibit an adequate internal as well as external quality. As the company implies, their aggregated data "will inform decisions about how to adapt existing infrastructure and investing in future solutions to make our cities more efficient" (Uber Movement Team n.d.).

While the data provided is very promising indeed, its context of generation should not be disregarded. The company profits from a good road infrastructure, therefore it is clear that they publish their data with hopes of improving the quality of the streets in a city. Although the data does include a lot of traffic information, it certainly lacks data about public transportation. Furthermore, it also doesn't include details about the reasons people choose Uber over other means of transport, or even the question why people move from A to B.

Therefore, even if the data is correct, reliable, and valid, it won't necessarily lead to better solutions by itself. It can probably clearly tell which roads are congested at which time, possibly leading cities to enhance the capacities at neuralgic points. However, improving road conditions might result in an increase in road traffic, leading again to congested roads. This phenomenon is also known as "induced demand effect" in traffic planning (Litman 2018).

Of course (as in case of Twitter) this doesn't mean that Uber's data is bad. It just shows a little aspect of a much larger topic, probably released by the company with certain intentions. The effect of companies' interests is an additional factor to regard when dealing with Big Data. Such information is extremely valuable, those who have access to such data probably won't release it by pure altruism.

As an example, the city of Vienna has tried since 2017 to retrieve data from Airbnb about their hosts, in order to have a basis for taxation, and of course also to know which effect the portal's apartments have on the real estate market of the city. In 2018 the local government stopped negotiations after Airbnb has shown no signs of complying with the claims (Schenk and Bauer 2018).

In conclusion, it must be noted that planners and city authorities not only need to assess the quality of Big Data, they also need to reflect the possible shortcomings resulting from its generation processes. In addition, working with Big Data needs to be embedded into the planning processes at the correct stage.

5.3. Recommendations

Planning is an inherently complex process. It experienced multiple changes during the course of the last decades. Schönwandt (1999) identified three generations of planning theory.

The first generation was characterised by a rational approach and a very rigid structure of the planning process. Planners were assumed to able to capture and quantify all aspects of a problem. Based on the problem and its specific circumstances, planners expected to apply some methods from a systematic set of rules and strategies, in order to find an objective solution to the problem (Schönwandt 1999, p. 25f).

The second generation, emerging during the late 60s to early 70s, has recognised the shortcomings of such approaches. It redefined the aim of planning to handle complex, and "ill-structured" (definition by Simon 1973) problems. Problems are "ill-structured" by definition because they are unique and cannot be delimited completely. Every "ill-structured" problem is just a symptom of another problem, therefore it becomes also almost completely impossible to find an ultimate and a completely verifiable solution for such a problem (Rittel 1972 as cited in Schönwandt 1999, p. 26).

However, also the second generation didn't succeed to completely dismiss the basic assumptions of the first generation. Although it certainly stopped seeing planning being capable of developing completely correct, rational, and objective solutions, in many cases it just transformed the methodologies. Furthermore the missing capability of repeating problem solving strategies on new problems was attributed to the nature and uniqueness of the problem. The planner was still regarded as an objective and rational entity. (Schönwandt, Voermanek, Utz, Grunau, and Hemberger 2013, p. 14).

The third generation of planning theory stopped perceiving the planner as a definite and objective entity. The “planning world” is a part of the “life world”, therefore the planner interacts with other entities and is consequently also influenced by the environment. The process of planning runs as a cycle through both worlds, with defining problems, methods, and the implementation plan in the planning world, but all the other steps being carried out in the everyday world (Schönwandt 1999, p. 27ff, translation of the terms as in Förster, Engler, Fabich, Lechner, et al. 2015, p. 2).

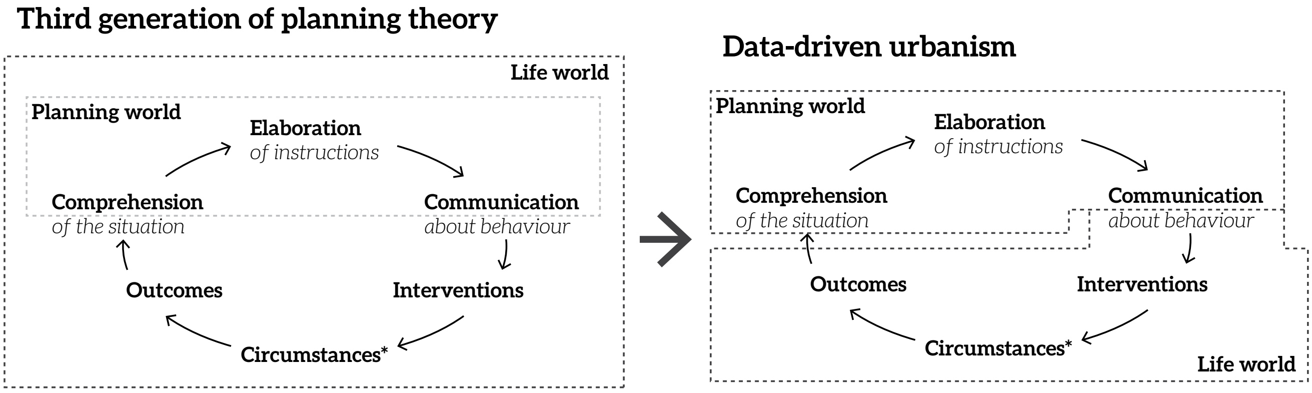
Data driven urbanism seems to redraw the border between the “planning world” and the “life world”. It doesn’t break the planning cycle, as it still doesn’t induce a deterministic approach. In contrast, the data enables to continuously track and monitor the effects of a strategy or measure, with modifications and fine-tuning available in real-time. Figure 26 presents the mechanism of data driven urbanism as compared to Schönwandt‘s third generation of planning theory.

The planner doesn’t need to be seen as part of the life world anymore. It is completely indifferent where the planner is located, interpreting a dataset and defining solutions for a given problem in a given city be done anywhere. This is also true for the case study, as I have never been to London but still managed to describe multiple aspects of the city.

Technical solutions can deliver objective and precise information. But by also using manual methods, for example observation, a lot more knowledge is possibly provided. Gehl and Svarre (2016, p. 6) present an automatic sensor that can quickly and efficiently count the number of cyclists along a bike path. At one day, there are no cyclists recorded. A human observer standing next to the sensor can see that a delivery truck parked on the bike path, meaning that cyclists drove around the sensor. In such case the observer can just take a picture of the situation and continue counting cyclists, while the automatic sensor would just record no bike traffic at all. Therefore, even though flexibility is a main asset of Big Data, sometimes it lacks this aspect completely.

Nonetheless, applying innovative and quantitative technical methods isn’t inherently wrong. Dangers of Big Data analysis don’t result simply from the data itself. They result from the ways how planners (and others) treat such data. When done correctly, handling of large amounts of data can indeed become a highly valuable asset for planning. It is just essential for planners to assess the value of the concrete data correctly, and to percieve Big Data analysis as one of the many means of getting information.

The opportunistic approach in trying to find traces in a provided dataset is potentially trea-



*“Setting“ in Förster et al. 2015

Figure 26: Schönwandt‘s (1999, p. 28) planning cycle of the third generation of planning theory (own figure - translation based on Förster et al. 2015, p. 2)

cherous. If the data is generated by a third organisation, its best usability can be expected only for the specific domain that organisation is interested in. Twitter as a portal for communication can depict discussions very well, but although its data contains locational information, it is not suitable for assessing such questions. In case of Uber the situation is similar. The data reveals a large amount of (pre-processed) information about a city's road system but nothing more.

Therefore, in many cases it might even become more fruitful generating the own data, applying a methodology that matches both the phenomenon and the specific question the best. Here it is even useful to learn from other companies generating Big Data. For example, after completing an Uber-ride, the app asks the passenger to rate their trip (Uber Technologies n.d.). Airbnb does the same after checking out from the visited apartment (Airbnb n.d.).

The implementation of such techniques is very simple. Neither portals want the users to fill in multiple-pages long surveys. They can simply assess the quality of the offered service by rating it on a 1-5 star scale, with the possibility of adding a short comment. Such a system could be implemented on public transport vehicles of the city. By installing small computers that do nothing more but sending a score with a locational and a temporal attribute, authorities could possibly see cases where problems appear in real-time. If there is an accumulation of low scores, the concrete issue can be identified by focusing on that area with traditional methods.

Similarly, it is not necessarily needed to resort to Uber's dataset to get information about traffic. Most busses are equipped with GPS-devices and can also provide a very similar information. By connecting the satisfaction level of passengers the city could generate a dataset with a similar quality as Uber's.

Generally speaking, Big Data is most valuable in early stages of situation analysis and monitoring. It can provide a rough view of the situation, but it doesn't help understanding the questions behind the reasons the observed phenomena appear. Or, as Mazzocchi (2015, p. 1252) puts it, "in most cases, understanding the *why* [BC: emphasis in original] is crucial for reaching a level of knowledge that can be used with confidence for practical applications and for making reliable predictions".

Therefore data needs to be treated as an asset and but not as the basis of a problem solving process. When thinking of planning, it is in addtition crucial to understand it as an iterative and communicative (Selle 1997) activity. Making concrete decisions premise an intensive discussion with the ones involved. Furthermoe, the discussion should also not be just about persuading people about a certain idea or a plan, but also understanding the needs of the people involved and gathering their local knowledge about and subjective sentiments towards a specific situation.

Social media sites can become highly valuable assets for cities as platforms of discussion. But not in means of just analysing the large quantities of data generated on such portals. They can be used as communication channels while actively listening to users, addressing their needs, and of course discussing plans, strategies, and all the relevant topics and themes coming up during a planning processes.

Although these recommendations seem really simple and reasonable, both Kitchin (2016) and Townsend (2014) agree that planning becomes increasingly stronger based on a very quantitative and data-based knowledge. Latter even concludes, when it comes to applying new technologies in urban planning, "evidence that we are moving in the wrong direction is everywhere" (Townsend 2014, p. 283).

For this reason, it is crucial that planning keeps its theoretical and qualitative foundation. While incorporating technological and methodological advancements into its set of skills, instruments, and approaches, new technologies should not replace everything. Big Data is an extremely helpful asset, but not the sole source of ultimate knowledge.

Addressing a specific question, small data, or "slow data" as Townsend (Townsend 2014, p. 316ff) puts it, might often be more useful. It is more reliable to set up a research plan and/or a survey with specific questions and indicators in mind that can be assumed to exploit the sought phenomenon more accurately. In contrast, Big Data requires an approach to look for interesting traces and information, but only with a very probabilistic approach. Some aspects might be recorded, while others not. In addition, also the recorded data has in general no guarantee to be valid, representative, or even correct. "Slow data must be collected, sparingly and by design, not harvested opportunistically from data exhaust". (Townsend 2014, p. 318).

Even if planning is in many cases very abstract, and deals with broad and general strategies, it still makes very concrete decisions. Therefore the data such decisions are based on also needs to be concrete. Furthermore, a danger in assuming that Big Data will guarantee us a clear insight into complex processes lies in the possibility of falling back to the holistic, God's Eye-view of planning, that was characteristic for decades (and in some cases is still until now). No matter how big the data is, it can't be big enough to model the complexity of our cities completely and accurately.

Or, as Townsend (2014, p. 317) cites the American sociologist William Bruce Cameron (1967, p. 13):

> *"[...] not everything that can be counted counts, and not everything that counts can be counted."*

References

List of Figures

Figure 1: Effects of changing the drawing order of points on a map (Field 2013) 18

Figure 2: Text morphology (own figure based on Trost 2004, p. 26) 29

Figure 3: Hidden Markov Chain (own figure) 31

Figure 4: Document pre-processing (own figure) 32

Figure 5: Sentiment analysis (own figure) 35

Figure 6: Factors considered by VADER (own figure based on Hutto and Gilbert 2014, p. 6f) 36

Figure 7: Functioning of the Biterm Topic Model (own figure based on Yan et al. 2013) 39

Figure 8: Software architecture (own figure) 44

Figure 9: Study area (own map) 46

Figure 10: Aggregated numbers of users and Twitter messages (all recorded tweets - own figure) 47

Figure 11: Location of tweets outside London (own map) 49

Figure 12: Spatial distribution of tweets (own map) 50

Figure 13: Temporal distribution of the recorded tweets per hour, weekday, and hour (own figure) 51

Figure 14: Topics recognised by the BTM (own figure) 53

Figure 15: Number of tweets assigned to topic categories (whole dataset, own figure) 54

Figure 16: Sentiment scores for selected topics (own figure) 55

Figure 17: Spatial distribution of tweets during workdays (own figure) 58

Figure 18: Tweeting activity (only correct coordinates and 1 tweet/user/time unit) in selected corine categories per day, weekday, and hour (own figure) 60

Figure 19: Weighted Odds Ratio for selected topics in central London (own figure) as compared to Lansley and Longley 2016, p. 93 62

Figure 20: Distribution of tweets by distance from selected amenities (own figure) 63

Figure 21: Temporal distribution for selected topics per day, weekday, and hour (Dates: FIFA 2018, People's Vote 2018, own figure) 64

Figure 22: Tweeting activity about topics „Brexit" and „Elections" on the day of People's Vote Protest March (Oct 20, 2018), compared to the demonstration route (own figure - demonstration route: People's Vote 2018) 65

Figure 23: Crime rate per ward (Greater London Authority 2015) as compared to tweeting activity about topic „Crime" (only correct coordinates, 1 tweet/user/ward) (own figure) 66

Figure 24: Average sentiment of tweets dealing with the toic „railways" (own figure) 67

Figure 25: Average sentiment of tweets/ward (only correct coordinates, mean sentiment of tweets/user - own figure) as compared to the crime rate (Greater London Authority 2015) 68

Figure 26: Schönwandt's (1999, p. 28) planning cycle of the third generation of planning theory (own figure - translation based on Förster et al. 2015, p. 2) 75

List of Tables

Table 1: NS-SEC Analytic classes (Office for National Statistics 2010, p. 3) 13

Table 2: Ethical principles (own table) 21

Table 3: Data handling (own table) 21

Table 4: Tweet object (own table based on Twitter Inc. n.d.-c) 26

Table 5: User object (own table based on Twitter Inc. n.d.-b) 26

Table 6: Stemming and lemmatising of texts (own table) 30

Table 7: Numbers of analysed tweets (own table) 52

List of Formula

Formula 1: BTM application (Yan et al. 2013, p. 1448) 40

Formula 2: Topic proportions for biterm b (Yan et al. 2013, p. 1448) 40

Formula 3: Proprortion of biterm b in document d (Yan et al. 2013, p. 1448) 40

Formula 4: Odds Ratio (Poorthuis et al. 2014, p. 8) 42

Formula 5: Weighted Odds Ratio (Poorthuis et al. 2014, p. 8) 42

Literature

› Ahmed, Wasim, Bath, Peter A., & Demartini, Gianluca. (2017). Chapter 4: Using Twitter as a Data Source: An Overview of Ethical, Legal, and Methodological Challenges, (February), 79–107. doi:10.1108/S2398-601820180000002004

› Airbnb. (n.d.). How do reviews work? Airbnb Help Center. https://www.airbnb.co.uk/help/article/13/how-do-reviews-work. Accessed 25 November 2018

› Anderson, Chris. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired. https://www.wired.com/2008/06/pb-theory/. Accessed 20 October 2018

› Balahur, Alexandra. (2013). Sentiment Analysis in Social Media Texts. In Proceedings ofthe 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, (pp. 120–128). Atlanta: Association for Computational Linguistics.

› Battisby, Alison. (2018, April 2). The latest UK Social Media Statistics for 2018. Avocado Social. https://www.avocadosocial.com/the-latest-uk-social-media-statistics-for-2018/. Accessed 20 October 2018

› Beckett, Andy. (2003, September 8). Santiago dreaming. The Guradian. https://www.theguardian.com/technology/2003/sep/08/sciencenews.chile. Accessed 25 November 2018

› Beninger, Kelsey, Fry, Alexandra, Jago, Natalie, Lepps, Hayley, Nass, Laura, & Silvester, Hannah. (2014). Research using Social Media; Users' Views. London: NatCen Social Research.

› Bing, Liu. (2011). Information Retrieval and Web Search. In Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (pp. 211–268). doi:DOI 10.1007/978-3-642-19460-3_6

› Bird, Steven, Klein, Ewan, & Loper, Edward. (2009). Natural Language Processing with Python. Journal of Endodontics (Vol. 28). doi:10.1097/00004770-200204000-00018

› Blei, David M. (2012). Probabilistic Topic Models. Communications of the ACM, 55(4), 77–84. doi:10.1145/2133806.2133826

› Bloomberg, Michael. (2014). Foreword. In The Responsive City. Engaging Communities Through Data-Smart Governance (pp. v–vi). San Francisco: Jossey-Bass.

› boyd, danah, & Crawford, Kate. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information Communication and Society, 15(5), 662–679. doi:10.1080/1369118X.2012.678878

- Cadwalladr, Carole. (2017, May 7). The great British Brexit robbery: how our democracy was hijacked. The Guradian. https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy. Accessed 24 October 2018
- Cadwalladr, Carole, & Graham-Harrison, Emma. (2018, March 17). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach | News | The Guardian. The Guradian. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election. Accessed 24 October 2018
- Cameron, William Bruce. (1967). Informal Sociology: A Casual Introduction to Sociological Thinking. New York: Random House.
- Campagna, Michele. (2014). The Geographic Turn in Social Media: Opportunities for Spatial Planning and Geodesign. In ICCSA (pp. 598–610). doi:10.1007/b98054
- Campagna, Michele. (2016). Social Media Geographic Information : Why social is special when it goes spatial ? European Handbook of Crowdsourced Geographic Information., 45–54. doi:http://doi.org/10.5334/bax
- Campagna, Michele, Floris, Roberta, & Massa, Pierangelo. (2015). Planning Support Systems and Smart Cities. doi:10.1007/978-3-319-18368-8
- Campagna Michele, Steinitz Carl, Di Cesare Elisabetta Anna, Chiara Cocco, Ballal Hrishikesh, & Canfield, Tess. (2016). Collaboration in Planning: the Geodesign approach. Rozwój Regionalny I Polityka Regionalna, 35, 55–72.
- Capps, Kriston. (2014, August 14). Twitter Made an Amazing Map of Twitter Going Nuts Over Ferguson. CityLab. https://www.citylab.com/equity/2014/08/twitter-made-an-amazing-map-of-twitter-going-nuts-over-ferguson/376119/. Accessed 23 October 2018
- Carr, Daniel B., Olsen, Anthony R., & White, Denis. (1992). Hexagon Mosaic Maps for Display of Univariate and Bivariate Geographical Data. Cartography and Geographic Information Systems, 19(4), 228–236.
- Chen, Edwin. (2011, August 22). Introduction to Latent Dirichlet Allocation. Edwin Chen's Blog. http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/. Accessed 29 October 2018
- Christensson, Per. (2016, June 20). API Definition. TechTerms. https://techterms.com/definition/api. Accessed 22 October 2018
- Cocchia, Annalisa. (2014). Smart and Digital City: A Systematic Literature Review. In R. P. Demeri & C. Rosenthal-Sabroux (Eds.), Smart City. How to Create Public and Economic Value with High Technology in Urban Space (1st ed., pp. 13–43). Cham: Springer. doi:10.1007/978-3-319-33681-7
- Crampton, Jeremy W., Graham, Mark, Poorthuis, Ate, Shelton, Taylor, Stephens, Monica, Wilson, Matthew W., & Zook, Matthew. (2013). Beyond the geotag: situating ' big data ' and leveraging the potential of the geoweb. Cartography and Geographic Information Science, 40(2), 130–139. doi:10.1080/15230406.2013.777137
- Davies, Mark. (n.d.). The Corpus of Historical American English (COHA), Google Books (Standard), and the Google Books (BYU / Advanced) corpus. BYU Google Books. https://google-

books.byu.edu/compare-googleBooks.asp. Accessed 30 October 2018

- De Mauro, Andrea, Greco, Marco, & Grimaldi, Michele. (2016). A formal definition of Big Data based on its essential features. Library Review, 65(3), 122–135. doi:10.1108/LR-06-2015-0061
- Devillers, Rodolphe, & Jeansoulin, Robert. (2006a). Fundamentals of Spatial Data Quality. Fundamentals of Spatial Data Quality. doi:10.1002/9780470612156
- Devillers, Rodolphe, & Jeansoulin, Robert. (2006b). Spatial Data Quality: Concepts. In Fundamentals of Spatial Data Quality (pp. 31–42).
- European Union. Regulation 2016/679. , 2014 Official Journal of the European Communities (2016). Brussels: European Parliament and European council. doi:http://eur-lex.europa.eu/pri/en/oj/dat/2003/l_285/l_28520031101en00330037.pdf
- Farias, D. I. Hernánde., & Rosso, P. (2016). Irony, Sarcasm, and Sentiment Analysis. Sentiment Analysis in Social Networks. Elsevier Inc. doi:10.1016/B978-0-12-804412-4.00007-3
- Field, Kenneth. (2013, June 20). 3 billion tweets on a map. Cartography. Design. Research. http://cartonerd.blogspot.com/2013/06/3-billion-tweets-on-map.html. Accessed 24 October 2018
- FIFA. (2018). 2018 FIFA World Cup RussiaTM - Matches. FIFA.com. https://www.fifa.com/worldcup/matches/. Accessed 27 November 2018
- Fiorina, Carly. (2004). Information: the currency of the digital age. San Francisco.
- Fonte, Cidália Costa, Antoniou, Vyron, Bastin, Lucy, Estima, Jacinto, Arsanjani, Jamal Jokar, Bayas, Juan-Carlos Laso, et al. (2017). Assessing VGI Data Quality. In Mapping and the Citizen Sensor (1st ed., pp. 137–163). Ubiquity Press. doi:https://doi. org/10.5334/bbf.g
- Forrest, Conner. (2018, January 9). Uber Movement gives urban planners access to data from 2 billion trips. TechRepublic. https://www.techrepublic.com/article/uber-movement-gives-urban-planners-access-to-data-from-2-billion-trips/. Accessed 25 November 2018
- Förster, Agnes, Engler, Carina, Fabich, Stephanie, Lechner, Sarah, Ramisch, Theresa, & Schöpf, Susanne. (2015). Beyond the usual suspects: Uncovering the network of civic and private sector actors in Munich's urban development. In AESOP Annual Congress. Prague.
- Francis, W. Nelson, & Kucera, Henry. (1979). Brown corpus manual. Brown University.
- García-Palomares, Juan Carlos, Salas-Olmedo, María Henar, Moya-Gómez, Borja, Condeço-Melhorado, Ana, & Gutiérrez, Javier. (2018). City dynamics through Twitter: Relationships between land use and spatiotemporal demographics. Cities, 72(June 2017), 310–319. doi:10.1016/j.cities.2017.09.007
- Gehl, Jan, & Svarre, Birgitte. (2016). Leben in Städten. Basel: Birkhäuser Verlag.
- Goodchild, Michael F. (2007). Citizens as sensors: The world of volunteered geography. GeoJournal, 69(4), 211–221. doi:10.1007/s10708-007-9111-y
- Greater London Authority. (2015). Ward Profiles and Atlas. London Datastore. https://data.london.gov.uk/dataset/ward-profiles-and-atlas. Accessed 26 November 2018
- Greenfield, Adam. (2013). Against the smart city. New York: Do projects.
- GSMA. (2017, June 13). Number of Mobile Subscribers Worldwide Hits 5 Billion. Gsma. https://www.gsma.com/newsroom/press-release/number-mobile-subscribers-worldwide-hits-5-billi-

on/. Accessed 20 October 2018

- Hays, David G. (1967a). Automatic Translation. In Introduction to Computational Linguistics (pp. 206–222).
- Hays, David G. (1967b). Techniques for Linguistic Research. In Introduction to Computational Linguistics (pp. 180–191).
- Huang, Haosheng, & Gartner, Georg. (2016). Using mobile crowdsourcing and geotagged social media data to study people's affective responses to environments. In European Handbook of Crowdsourced Geographic Information (pp. 383–397). doi:10.5334/bax
- Hunston, Susan. (2006). Corpus Linguistics. In K. Brown (Ed.), Encyclopedia of language and linguistics (2nd ed., pp. 234–248). Amssterdam: Elsevier.
- Hutto, C. J. (2018). VADER-Sentiment-Analysis (Readme). GitHub. https://github.com/cjhutto/vaderSentiment. Accessed 28 October 2018
- Hutto, C. J., & Gilbert, Eric. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM-2014). doi:10.1210/en.2011-1066
- Jónsson, Elías, & Stolee, Jake. (2017). An Evaluation of Topic Modelling Techniques for Twitter, 1–11.
- Joseph, Kenneth, Landwehr, Peter M., & Carley, Kathleen M. (2014). Two 1%s don't make a whole: Comparing simultaneous samples from Twitter's Streaming API. In Social Computing, Behavioral-Cultural Modeling, and Prediction (pp. 75–83).
- Kemp, Simon. (2018, October 17). The State of the Internet in Q4 2018. We are Social. https://wearesocial.com/blog/2018/10/the-state-of-the-internet-in-q4-2018. Accessed 22 October 2018
- Kholkovskaia, Olga. (2017). Role of the Brown Corpus in the History of Corpus Linguistics, 1–4.
- Kitchin, Rob. (2016). The ethics of smart cities and urban science. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2083). doi:10.1098/rsta.2016.0115
- Lamanna, Fabio, Lenormand, Maxime, Salas-Olmedo, María Henar, Romanillos, Gustavo, Gonçalves, Bruno, & Ramasco, José J. (2018). Immigrant community integration in world cities. PLoS ONE, 13(3), 1–19. doi:10.1371/journal.pone.0191612
- Lansley, Guy, & Longley, Paul A. (2016). The geography of Twitter topics in London. Computers, Environment and Urban Systems, 58, 85–96. doi:10.1016/j.compenvurbsys.2016.04.002
- Leetaru, Kalev H., Wang, Shaowen, Cao, Guofeng, Padmanabhan, Anand, & Shook, Eric. (2013). Mapping the global Twitter heartbeat: The geography of Twitter. First Monday, 18(5–6), 1–34.
- Leonardo, Piazza, Mooney, Peter, & Minghini, Marco. (2017). A Review of OpenStreetMap Data. In Mapping and the Citizen Sensor (pp. 37–59). London: Ubiquity Press.
- Linckels, Serge, & Meinel, Christoph. (2011). E-Librarian Service. User-Friendly Semantic Search in Digital Libraries. Berlin, Heidelberg: Springer.
- Litman, Todd. (2018). Generated Traffic and Induced Travel: Implications for Transportation Planning, 71(4), 38–47.
- Liu, Bing. (2012). Sentiment Analysis and Opinion Mining. San Rafael: Morgan & Claypool.

doi:10.1007/978-1-4899-7502-7_907-1

- Lopez, German. (2016, January 27). The 2014 Ferguson protests over the Michael Brown shooting, explained. VOX. https://www.vox.com/cards/mike-brown-protests-ferguson-missouri/mike-brown-shooting-facts-details. Accessed 24 October 2018
- Luo, Tiejian, Chen, Su, Xu, Guandong, & Zhou, Jia. (2013). Sentiment Analysis. In Trust-based Collective View Prediction (pp. 53–68).
- Manning, Christopher D., Raghavan, Prabhakar, & Schütze, Hinrich. (2009). An Introduction to Information Retrieval. Camebridge: Camebridge University Press.
- Marr, Bernard. (2018, May 21). How much data do we create every day? The Mind-Blowing Stats Everyone Should Read. Forbes. https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#5a35fed460ba. Accessed 20 October 2018
- Mazzocchi, Fulvio. (2015). Could Big Data be the end of theory in science?: A few remarks on the epistemology of data-driven science. EMBO reports, 16(10), 1250–1255. doi:10.15252/embr.201541001
- Mikheev, Andrei. (2004). Text Segmentation. In R. Mitkov (Ed.), The Oxford Handbook of Computational Linguistics (pp. 201–218).
- Mitchell, Lewis, Frank, Morgan R., Harris, Kameron Decker, Dodds, Peter Sheridan, & Danforth, Christopher M. (2013). The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place. PLoS ONE, 8(5). doi:10.1371/journal.pone.0064417
- Mitkov, Ruslan (Ed.). (2004). The Oxford Handbook of Computational Linguistics (1st ed.). New York: Oxford University Press.
- Mocanu, Delia, Baronchelli, Andrea, Perra, Nicola, Gonçalves, Bruno, Zhang, Qian, & Vespignani, Alessandro. (2013). The Twitter of Babel: Mapping World Languages through Microblogging Platforms. PLoS ONE, 8(4). doi:10.1371/journal.pone.0061981
- Morstatter, Fred, Pfeffer, Jürgen, Liu, Huan, & Carley, Kathleen M. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose, 400–408. doi:10.1007/978-3-319-05579-4_10
- National Coordination Office for Space-Based Positioning. (2017). GPS Accuracy. GPS.gov. https://www.gps.gov/systems/gps/performance/accuracy/. Accessed 12 November 2018
- Office for National Statistics. (2010). Standard Occupational Classification 2010. Volume 3 The National Statistics Socio-economic Classification: (Rebased on the SOC2010) User Manual (Vol. 3). Houndmills: Palgrave Macmillan. http://www.ons.gov.uk/ons/guide-method/classifications/current-standard-classifications/soc2010/soc2010-volume-3-ns-sec--rebased-on-soc2010--user-manual/index.html
- OpenStreetMap Contributors. (n.d.). OpenStreetMap stellt Kartendaten für tausende von Webseiten, Apps und andere Geräte zur Verfügung. https://www.openstreetmap.org/about. Accessed 25 November 2018
- People's Vote. (2018). People's Vote and The Independent March for the Future. People's Vote.

https://www.peoples-vote.uk/march. Accessed 26 November 2018

- Pérez, Fernando, Granger, Brian E., & Hunter, John D. (2011). Python : An Ecosystem for Scientific Computing. Computing in Science & Engineering, 13(2), 13–21. doi:https://doi.org/10.1109/MCSE.2010.119
- Poorthuis, Ate, Zook, Matthew, Shelton, Taylor, Graham, Mark, & Stephens, Monica. (2014). Using Geotagged Digital Social Data in Geographic Research. In Key Methods in Geography (Pre-publication version of chapter).
- Pozzi, Federico Alberto, Fersini, Elisabetta, Messina, Enza, & Liu, Bing. (2016). Challenges of Sentiment Analysis in Social Networks: An Overview. In Sentiment Analysis in Social Networks (Vol. 1, pp. 1–11). Elsevier Inc. doi:10.1016/B978-0-12-804412-4.00001-2
- Python Software Foundation. (2018a, October 20). Download Python. Python.
- Python Software Foundation. (2018b, October 25). Common string operations. Python 3.6.7 documentation. https://docs.python.org/3.6/library/string.html. Accessed 31 October 2018
- Resnick, Brian. (2016, May 12). Researchers just released profile data on 70,000 OkCupid users without permission. VOX. https://www.vox.com/2016/5/12/11666116/70000-okcupid-users-data-release. Accessed 24 October 2018
- Ribeiro, Filipe N., Araújo, Matheus, Gonçalves, Pollyanna, Gonçalves, Marcos André, & Benevenuto, Fabrício. (2016). SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Sci.ence, 5(23). doi:10.1140/epjds/s13688-016-0085-1
- Rimal, Bhagawat, Zhang, Lifu, Keshtkar, Hamidreza, Wang, Nan, & Lin, Yi. (2017). Monitoring and Modeling of Spatiotemporal Urban Expansion and Land-Use/Land-Cover Change Using Integrated Markov Chain Cellular Automata Model. ISPRS International Journal of Geo-Information, 6(288), 1–21. doi:10.3390/ijgi6090288
- Ritchie, Rene. (2018, January 9). 11 years ago today, Steve Jobs introduced the iPhone. iMore. https://www.imore.com/history-iphone-original. Accessed 25 November 2018
- Rittel, Horst. (1972). On the Planning Crisis: Systems Analysis of the “First and Second Generations.” Bedriftsøkonomen, 8, 390–396.
- Rogers, Simon. (2012, October 23). Twitter reveals London’ s ethnic groupings. The Guradian. https://www.theguardian.com/news/datablog/gallery/2012/oct/23/twitter-ehtnicity-london-map. Accessed 24 October 2018
- Rogers, Simon. (2013, December 12). Beyonce releases her new album online. CARTO. https://srogers.carto.com/viz/337d9194-6458-11e3-85b5-e5e70547d141/public_map
- Rogers, Simon. (2014a, May 19). FA Cup final 2014 tweets: mapped. The Guradian. https://www.theguardian.com/football/datablog/ng-interactive/2014/may/19/fa-cup-final-2014-tweets-mapped. Accessed 24 October 2018
- Rogers, Simon. (2014b, May 23). Animated map : local and EU elections in tweets. The Guradian. https://www.theguardian.com/news/datablog/ng-interactive/2014/may/23/animated-map-local-and-eu-elections-in-tweets. Accessed 24 October 2018
- Rogers, Simon. (2014c, August 13). How news of #Ferguson spread across Twitter. CARTO. https://srogers.carto.com/viz/4a5eb582-23ed-11e4-bd6b-0e230854a1cb/embed_map. Accessed 23

October 2018

- Rosen, Aliza, & Ihara, Ikuhiro. (2017, September 26). Giving you more characters to express yourself. Twiter Blog. https://blog.twitter.com/official/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself.html. Accessed 20 October 2018
- Rosen, Guy. (2018, September 28). Security Update. Facebook Newsroom. https://newsroom.fb.com/news/2018/09/security-update/. Accessed 24 October 2018
- Roth, Joel, & Johnson, Rob. (2018, July 24). New developer requirements to protect our platform. Twiter Blog. https://blog.twitter.com/developer/en_us/topics/tools/2018/new-developer-requirements-to-protect-our-platform.html. Accessed 22 October 2018
- Rzeszewski, Michal, & Beluch, Lukasz. (2017). Spatial Characteristics of Twitter Users—Toward the Understanding of Geosocial Media Production. ISPRS International Journal of Geo-Information, 6(8), 236. doi:10.3390/ijgi6080236
- Samuelsson, Christer. (2004). Statistical Methods. In R. Miitkov (Ed.), The Oxford Handbook of Computational Linguistics (pp. 358–375).
- Schenk, Julia, & Bauer, Anna-Maria. (2018, October 31). Stadt Wien bricht Ortstaxe-Verhandlungen mit Airbnb ab. Kurier. https://kurier.at/chronik/wien/stadt-wien-bricht-ortstaxe-verhandlungen-mit-airbnb-ab/400311060. Accessed 22 November 2018
- Schönwandt, Walter. (1999). Grundriss einer Planungstheorie der «dritten Generation». disP - the Planning Review, 35(136/137), 25–35. doi:10.1080/02513625.1999.10556696
- Schönwandt, Walter, Voermanek, Katrin, Utz, Jürgen, Grunau, Jens, & Hemberger, Christoph. (2013). Komplexe Probleme Lösen. (Jovis, Ed.). Berlin.
- Scola, Nancy. (2018, June 29). Google is building a city of the future in Toronto. Would anyone want to live there? Politico. https://www.politico.com/magazine/story/2018/06/29/google-city-technology-toronto-canada-218841. Accessed 25 November 2018
- Scott, David W. (1985). Averaged Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions. The Annals of Statistics, 13(3), 1024–1040.
- Selle, Klaus. (1997). Planung und Kommunikation. disP - the Planning Review, 129(1997), 40–47. doi:10.1080/02513625.1997.10556645
- Seward, Zachary M. (2014, August 12). Twitter admits that as many as 23 million of its active users are automated. Quartz. http://qz.com/248063/twitter-admits-that-as-many-as-23-million-of-its-active-users-are-actually-bots/. Accessed 16 November 2018
- Shelton, Taylor. (2017). Spatialities of data: mapping social media 'beyond the geotag.' GeoJournal, 82(4), 721–734. doi:10.1007/s10708-016-9713-3
- Shelton, Taylor, Poorthuis, Ate, Graham, Mark, & Zook, Matthew. (2014). Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data.' Geoforum, 52, 167–179. doi:10.1016/j.geoforum.2014.01.006
- Shelton, Taylor, Poorthuis, Ate, & Zook, Matthew. (2015). Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. Landscape and Urban Planning, 142, 198–211. doi:10.1016/j.landurbplan.2015.02.020
- Simon, Herbert A. (1973). The structure of ill structured problems. Artificial Intelligence, 4,

181–201. doi:10.1016/0004-3702(73)90011-8

- Sloan, Luke. (2017). Who Tweets in the United Kingdom? Profiling the Twitter Population Using the British Social Attitudes Survey 2015. Social Media + Society, 3(1), 205630511769898. doi:10.1177/2056305117698981
- Sloan, Luke, Morgan, Jeffrey, Burnap, Pete, & Williams, Matthew. (2015). Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. PLoS ONE, 10(3), 1–20. doi:10.1371/journal.pone.0115545
- Sloan, Luke, Morgan, Jeffrey, Housley, William, Williams, Matthew, Edwards, Adam, Burnap, Pete, & Rana, Omer. (2013). SocResOnline Knowing the Tweeters.pdf. Sociological Research Online, 18(7).
- Stampler, Laura. (2013, December 16). This Map of Beyoncé-Related Tweets in Real Time After the Album Dropped Is Flawless. Time. http://newsfeed.time.com/2013/12/16/this-map-of-beyonce-related-tweets-in-real-time-after-the-album-dropped-is-flawless/. Accessed 23 October 2018
- Townsend, Anthony M. (2014). Smart Cities (1st ed.). London, New York: W. W. Norton & Company.
- Trost, Harald. (2004). Morphology. In R. Mitkov (Ed.), The Oxford Handbook of Computational Linguistics (pp. 25–47).
- Twitter Inc. (n.d.-a). Registration form. https://twitter.com/i/flow/signup. Accessed 20 October 2018
- Twitter Inc. (n.d.-b). User object. Twitter Developers. https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/user-objec. Accessed 22 October 2018
- Twitter Inc. (n.d.-c). Tweet objects. Twitter Developers. https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object. Accessed 22 October 2018
- Twitter Inc. (n.d.-d). Docs. Twitter Developers. https://developer.twitter.com/en/docs.html. Accessed 22 October 2018
- Twitter Inc. (n.d.-e). Filter realtime Tweets. Twitter Developers. https://developer.twitter.com/en/docs/tweets/filter-realtime/guides/basic-stream-parameters. Accessed 22 October 2018
- Twitter Inc. (n.d.-f). Pricing. Twitter Developers. https://developer.twitter.com/en/pricing.html. Accessed 22 October 2018
- Twitter Inc. (n.d.-g). Configure Search Tweets: Full Archive. Twitter Developers. https://developer.twitter.com/en/pricing/search-fullarchive. Accessed 22 October 2018
- Twitter Inc. (n.d.-h). Consuming streaming data. Twitter Developers. https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data. Accessed 22 October 2018
- Twitter Inc. (n.d.-i). Introduction to Tweet JSON. Twitter Developers. https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json.htm. Accessed 22 October 2018
- Twitter Inc. (2017, November 3). Developer Policy. Developer Terms. https://developer.twitter.com/en/developer-terms/policy. Accessed 20 October 2018
- Twitter Inc. (2018a, May 25). Twitter Terms of Service. https://twitter.com/en/tos. Accessed 18

October 2018

- Twitter Inc. (2018b, May 25). Developer Agreement. Developer Terms. https://developer.twitter.com/en/developer-terms/agreement. Accessed 20 October 2018
- Twitter Inc. (2018c, May 25). Twitter Privacy Policy. Twitter. https://twitter.com/privacy. Accessed 24 October 2018
- Tzoukermann, Evelyne, Klavans, Judith L., & Strzalkowski, Tomek. (2004). Information Retrieval. In R. Mitkov (Ed.), The Oxford Handbook of Computational Linguistics (pp. 529–544).
- Uber Movement Team. (n.d.). FAQs. Uber Movement. https://movement.uber.com/faqs. Accessed 22 November 2018
- Uber Movement Team. (2018a, March 15). Examining the Impact of the London Tower Bridge Closure. Medium. https://medium.com/uber-movement/examining-the-impact-of-the-london-tower-bridge-closure-5b7626e44915. Accessed 22 November 2018
- Uber Movement Team. (2018b, April 11). How March Floods Affected Nairobi Travel Times. Medium. https://medium.com/uber-movement/how-march-floods-affected-nairobi-travel-times-eaf850285004?lang=en-GB. Accessed 22 November 2018
- Uber Movement Team. (2018c, May 29). Examining the Impact of Traffic as Delhi Shops on Dhanteras. Medium. https://medium.com/uber-movement/examining-the-impact-of-the-london-tower-bridge-closure-5b7626e44915. Accessed 22 November 2018
- Uber Technologies. (n.d.). Rating a driver. Uber Help. https://help.uber.com/riders/article/einen-fahrer-bewerten?nodeId=478d7463-99cb-48ff-a81f-0ab227a1e267. Accessed 25 November 2018
- van Dijk, Jan A. G. M. (2006). Digital divide research, achievements and shortcomings. Poetics, 34, 221–235. doi:10.1016/j.poetic.2006.05.004
- Van Mierlo, Trevor. (2014). The 1% rule in four digital health social networks: An observational study. Journal of Medical Internet Research, 16(2). doi:10.2196/jmir.2966
- Varol, Onur, Ferrara, Emilio, Davis, Clayton A., Menczer, Filippo, & Flammini, Alessandro. (2017). Online Human-Bot Interactions: Detection, Estimation, and Characterization. In Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017). Montreal, Canada: Association for the Advancement of Artificial Intelligence. doi:10.5260/chara.16.4.57
- Venturini, Tommaso, Jacomy, Mathieu, Meunier, Axel, & Latour, Bruno. (2017). An unexpected journey: A few lessons from sciences Po médialab's experience. Big Data & Society, 4(2), 205395171772094. doi:10.1177/2053951717720949
- Voutilainen, Atro. (2004). Part-of-Speech Tagging. In R. Mitkov (Ed.), The Oxford Handbook of Computational Linguistics (pp. 219–248).
- Wang, Xiaoli, Wong, Albert Kai-sun, & Kong, Yongping. (2012). Mobility Tracking using GPS, Wi-Fi and Cell ID. In The International Conference on Information Network 2012 (pp. 171–176). Bali, Indonesia: IEEE.
- We are Flint. (2018, February 15). Social Media Demographics 2018. https://weareflint.co.uk/press-release-social-media-demographics-2018/. Accessed 20 October 2018
- Wiedemann, Gregor, & Niekler, Andreas. (2016). Analyse qualitativer Daten mit dem "Leipzig

Corpus Miner." In Text Mining in den Sozialwissenschaften (pp. 63–88).

- Yan, Xiaohui, Guo, Jiafeng, Lan, Yanyan, & Cheng, Xueqi. (2013). A biterm topic model for short texts. Proceedings of the 22nd international conference on World Wide Web - WWW '13, 1445–1456. doi:10.1145/2488388.2488514
- Yang, Wei, & Mu, Lan. (2015). GIS analysis of depression among Twitter users. Applied Geography, 60, 217–223. doi:10.1016/j.apgeog.2014.10.016
- Ziegler, Chris. (2012, April 25). This was the original "Google Phone" presented in 2006. The Verge. http://www.theverge.com/2012/4/25/2974676/this-was-the-original-google-phone-presented-in-2006. Accessed 25 November 2018

Programming Components

- Hutto, C. J., & Gilbert, Eric. (2018). VADER Sentiment Analyser (3.2.1). https://github.com/cjhutto/vaderSentiment. Accessed 27 November 2018
- Matsubara, Yutaka. (2016). PyMySQL (0.9.2). https://pymysql.readthedocs.io/. Accessed 27 November 2018
- NLTK Project. (2018). Natural Language Toolkit (3.2.5). http://www.nltk.org/. Accessed 27 November 2018
- Plot.ly. (2018a). Plotly (3.4.1). https://plot.ly/python/%0ADash. Accessed 27 November 2018
- Plot.ly. (2018b). Dash (0.30.0). Plot.ly. (2018). Dash. Accessed 27 November 2018
- Ronacher, Armin. (2018). FLASK (1.0.2). http://flask.pocoo.org/. Accessed 27 November 2018
- Yan, Xiaohui. (2018). Biterm Topic Model (0.5). https://github.com/xiaohuiyan/BTM. Accessed 27 November 2018

Appendix

Queried Topics

topic1 (Elections)		
Rank	Word	Proportion
1	vote	0.0258192
2	brexit	0.0207082
3	party	0.018063
4	labour	0.0179369
5	people	0.0176984
6	leave	0.0133321
7	tory	0.0121876
8	voted	0.0107359
9	election	0.00811906
10	government	0.00743697

topic2 (Housing / architecture)		
Rank	Word	Proportion
1	house	0.018835
2	room	0.0140528
3	door	0.0100112
4	home	0.00993097
5	open	0.0086774
6	london	0.00818183
7	window	0.00697923
8	garden	0.00655523
9	flat	0.00652595
10	space	0.00578638

topic3 (Superlatives)		
Rank	Word	Proportion
1	best	0.069993
2	ever	0.0626175
3	thing	0.0300609
4	seen	0.0250557
5	never	0.0221152
6	world	0.021827
7	time	0.0198282
8	life	0.017566
9	worst	0.0163063
10	first	0.0138875

topic4 (Cosmetics)		
Rank	Word	Proportion
1	hair	0.0388205
2	colour	0.0150003
3	nail	0.0120937
4	gold	0.0114905
5	skin	0.0107227
6	makeup	0.0103364
7	look	0.00981422
8	beauty	0.00966639
9	london	0.00875793
10	love	0.00817852

topic5 (Visual arts)		
Rank	Word	Proportion
1	london	0.0274472
2	exhibition	0.0161831
3	artist	0.0151386
4	museum	0.0133676
5	design	0.0115365
6	gallery	0.0113349
7	work	0.009097
8	painting	0.008701
9	love	0.00701167
10	drawing	0.00686271

topic6 (Road traffic)		
Rank	Word	Proportion
1	road	0.0143862
2	bike	0.0127515
3	driver	0.0103337
4	ride	0.00944317
5	london	0.00801636
6	today	0.00729587
7	mile	0.00696752
8	driving	0.00656829
9	traffic	0.00650215
10	cycling	0.00608049

topic7 (Sleeping/resting)		
Rank	Word	Proportion
1	work	0.0229009
2	hour	0.0224395
3	night	0.0212815
4	time	0.0190036
5	last	0.017254
6	morning	0.0166154
7	sleep	0.0130666
8	good	0.013019
9	week	0.0107242
10	today	0.0103698

topic8 (Ads for social media channels)

Rank	Word	Proportion
1	video	0.035721
2	check	0.0195699
3	link	0.0146869
4	youtube	0.0122798
5	live	0.0114831
6	post	0.0110039
7	blog	0.00900579
8	watch	0.00875145
9	channel	0.00852381
10	follow	0.0084662

topic9 (Birthday wishes)

Rank	Word	Proportion
1	happy	0.0952087
2	birthday	0.0822039
3	love	0.0175724
4	today	0.0162991
5	friend	0.0162406
6	year	0.0145107
7	best	0.0139282
8	hope	0.0135887
9	celebrating	0.0100865
10	wish	0.00936502

topic10 (Carnivals/fairs)

Rank	Word	Proportion
1	hill	0.0898697
2	carnival	0.0472791
3	notting	0.0445113
4	london	0.0202222
5	festival	0.0133873
6	nottinghillcarnival	0.0105065
7	party	0.0102806
8	vibe	0.00804938
9	nottinghill	0.00776695
10	boom	0.00765398

topic11 (Crime)

Rank	Word	Proportion
1	police	0.0162529
2	fire	0.00850769
3	crime	0.00777423
4	people	0.00612835
5	year	0.00612421
6	child	0.00607145
7	news	0.00580869
8	attack	0.00520041
9	court	0.00478765
10	officer	0.00467385

topic12 (Social controversies)

Rank	Word	Proportion
1	corbyn	0.0108881
2	right	0.0107674
3	labour	0.0098138
4	party	0.00840428
5	racist	0.00815236
6	dont	0.00657924
7	people	0.00601692
8	think	0.00560203
9	left	0.00533891
10	tory	0.00521326

topic13 (Pets / animals)

Rank	Word	Proportion
1	dog	0.0153953
2	little	0.0143706
3	isle	0.0143131
4	love	0.0137097
5	wight	0.0123929
6	cat	0.0108988
7	repost	0.0104822
8	puppy	0.00927551
9	dogsofinstagram	0.00903608
10	cute	0.00826033

topic14 (Cricket)

Rank	Word	Proportion
1	cricket	0.0317774
2	england	0.0227757
3	test	0.0172713
4	lord	0.0141487
5	match	0.0130621
6	india	0.0116295
7	great	0.0115466
8	ground	0.0105458
9	good	0.0103084
10	wicket	0.0102541

topic15 (Anticipation („can't wait"))

Rank	Word	Proportion
1	cant	0.162397

2	wait	0.0692474
3	believe	0.033156
4	even	0.0162382
5	still	0.0137955
6	couldnt	0.00905354
7	year	0.00835886
8	tomorrow	0.00823427
9	find	0.00797943
10	people	0.00787749

topic16 (Negations)

Rank	Word	Proportion
1	dont	0.0882657
2	know	0.0448802
3	want	0.02734
4	think	0.0250236
5	people	0.0213576
6	even	0.0159134
7	really	0.0141373
8	didnt	0.0137894
9	doesnt	0.0113289
10	need	0.00947792

topic17 (Anticipation („looking forward“))

Rank	Word	Proportion
1	looking	0.0957412
2	forward	0.0871868
3	good	0.0663873
4	look	0.0320975
5	luck	0.0236745
6	seeing	0.0219632
7	great	0.0211178
8	really	0.0200189
9	today	0.0184593
10	tomorrow	0.013444

topic18 (Literature)

Rank	Word	Proportion
1	book	0.0706551
2	read	0.0438873
3	reading	0.0213046
4	review	0.0158507
5	writing	0.0133901
6	article	0.010757
7	blog	0.0101072
8	story	0.00972538
9	write	0.00911254
10	today	0.00761586

topic19 (Congratulations)

Rank	Word	Proportion
1	well	0.0641354
2	done	0.0499367
3	team	0.0205811
4	congratulation	0.0186697
5	proud	0.0186301
6	great	0.0174604
7	year	0.0140654
8	award	0.0104986
9	amazing	0.0102296
10	work	0.00969066

topic20 (Approval of arguments)

Rank	Word	Proportion
1	think	0.0101982
2	people	0.00916453
3	dont	0.00890163
4	thats	0.00763115
5	right	0.00730365
6	agree	0.00685346
7	point	0.00663559
8	need	0.00581288
9	make	0.00539363
10	wrong	0.00500633

topic21 (London culture)

Rank	Word	Proportion
1	london	0.0404834
2	palace	0.0356076
3	harry	0.0286227
4	king	0.0238827
5	potter	0.0204284
6	studio	0.0196773
7	theatre	0.0181836
8	tour	0.0161382
9	buckingham	0.0123912
10	queen	0.0120432

topic22 (Music)

Rank	Word	Proportion
1	song	0.0269732
2	music	0.0215195
3	album	0.0179712
4	love	0.0148203
5	listen	0.00922121
6	track	0.0071795

7	listening	0.00663672
8	tune	0.00643042
9	video	0.00614266
10	good	0.00585109

topic23 (Radio programme)

Rank	Word	Proportion
1	radio	0.0694823
2	live	0.0352056
3	hospital	0.0338876
4	kingston	0.0337814
5	july	0.0293852
6	show	0.0285006
7	june	0.0270234
8	paul	0.0251216
9	sunday	0.0225387
10	chris	0.019027

topic24 (Food bank)

Rank	Word	Proportion
1	free	0.146606
2	london	0.125405
3	unitedkingdom	0.101335
4	foodwaste	0.0778036
5	pret	0.0215681
6	zerowaste	0.0194214
7	sample	0.0121177
8	sale	0.0108521
9	organic	0.0107634
10	samplesale	0.0100123

topic25 (Airplanes / ships)

Rank	Word	Proportion
1	flight	0.0355296
2	london	0.0318604
3	heathrow	0.0317713
4	airport	0.029038
5	plane	0.015687
6	terminal	0.0142275
7	flying	0.0117801
8	security	0.0108777
9	raf100	0.01051
10	track	0.00950356

topic26 (Railways)

Rank	Word	Proportion
1	train	0.0517905
2	service	0.0156206
3	london	0.0152587
4	station	0.0147128
5	hour	0.0114571
6	line	0.00954782
7	time	0.00932448
8	minute	0.007978
9	home	0.00717428
10	cancelled	0.00712064

topic27 (Football players)

Rank	Word	Proportion
1	player	0.0163407
2	season	0.0112014
3	arsenal	0.0104095
4	signing	0.00932219
5	chelsea	0.00916537
6	club	0.00838477
7	sign	0.0082912
8	think	0.0071086
9	transfer	0.00699857
10	fan	0.00684956

topic28 (Food (restaurants))

Rank	Word	Proportion
1	food	0.0336459
2	london	0.0242421
3	lunch	0.013009
4	restaurant	0.0121181
5	dinner	0.00993903
6	pizza	0.00878765
7	delicious	0.00854024
8	vegan	0.00832614
9	burger	0.00765648
10	today	0.00753873

topic29 (Social media trends)

Rank	Word	Proportion
1	photo	0.0986906
2	took	0.0407535
3	place	0.0398055
4	trndnl	0.0352759
5	trend	0.0310228
6	posted	0.0290608
7	hashtag	0.0254529
8	worldwides	0.0243468
9	top20	0.0232539
10	hour	0.0196592

topic30 (Clubbing)

Rank	Word	Proportion
1	night	0.0464204
2	last	0.0296097
3	tonight	0.0203148
4	london	0.0163632
5	party	0.0143449
6	saturday	0.0127106
7	friday	0.0120693
8	come	0.0106302
9	event	0.0100393
10	join	0.0100137

topic31 (Pop concerts)

Rank	Word	Proportion
1	london	0.0411848
2	stadium	0.0384478
3	wembley	0.0314352
4	night	0.0266616
5	queen	0.0176854
6	last	0.0161954
7	taylor	0.0157176
8	tour	0.0144787
9	concert	0.0113125
10	twickenham	0.0110372

topic32 (Concerts (more classical but also pop))

Rank	Word	Proportion
1	music	0.0218535
2	festival	0.0182985
3	london	0.0177489
4	royal	0.0150504
5	tonight	0.0149563
6	live	0.0116076
7	show	0.0114178
8	night	0.010817
9	hall	0.00987956
10	stage	0.00867309

topic33 (Charity / volunteering)

Rank	Word	Proportion
1	great	0.0348089
2	today	0.0244975
3	team	0.0123401
4	work	0.0121522
5	thanks	0.0113143
6	thank	0.0107814
7	people	0.0102874
8	amazing	0.00988527
9	lovely	0.0083032
10	event	0.00802933

topic34 (Movies)

Rank	Word	Proportion
1	film	0.0287785
2	movie	0.0186688
3	star	0.00937717
4	watch	0.00859356
5	london	0.00753264
6	good	0.00710083
7	cinema	0.00612959
8	time	0.00612821
9	best	0.00580676
10	watching	0.00579296

topic35 (Slang)

Rank	Word	Proportion
1	dont	0.0128578
2	know	0.0104651
3	shit	0.0101151
4	even	0.00844569
5	cant	0.00786533
6	need	0.00765829
7	fuck	0.00740563
8	girl	0.00731437
9	people	0.00725932
10	really	0.0072563

topic36 (Business events)

Rank	Word	Proportion
1	great	0.0163955
2	london	0.01054
3	today	0.0097567
4	business	0.00930384
5	event	0.00791304
6	looking	0.00674896
7	talk	0.00656618
8	social	0.00634941
9	future	0.00584031
10	conference	0.00544194

topic37 (Cursing)

Rank	Word	Proportion
1	fucking	0.0188952
2	people	0.0151554
3	fuck	0.0144577
4	shit	0.0129696
5	dont	0.0107436

6	youre	0.00749753
7	cunt	0.00665963
8	stop	0.00597088
9	absolute	0.0057938
10	look	0.00527039

topic38 (Appreciation)

Rank	Word	Proportion
1	great	0.0197201
2	amazing	0.0166628
3	night	0.0152053
4	last	0.0138954
5	show	0.0121653
6	absolutely	0.010692
7	brilliant	0.00984747
8	really	0.0095712
9	beautiful	0.00831853
10	thank	0.00818408

topic39 (Ticket sales)

Rank	Word	Proportion
1	ticket	0.0687453
2	london	0.015358
3	anyone	0.0122615
4	show	0.0119462
5	saturday	0.0118896
6	going	0.0101319
7	tomorrow	0.0096727
8	sunday	0.00913263
9	still	0.00866372
10	tonight	0.00847453

topic40 (Motivation)

Rank	Word	Proportion
1	life	0.0236795
2	love	0.0174218
3	people	0.0142032
4	never	0.012903
5	always	0.0116753
6	time	0.00720583
7	make	0.00709216
8	dont	0.00688465
9	come	0.00676711
10	take	0.00672648

topic41 (None1)

Rank	Word	Proportion
1	think	0.00986141
2	people	0.00944533
3	know	0.00652591
4	thing	0.00652493
5	dont	0.00613102
6	really	0.00582613
7	word	0.00542603
8	thats	0.00533635
9	many	0.00518668
10	love	0.00470636

topic42 (Premier league)

Rank	Word	Proportion
1	game	0.02698
2	season	0.0219729
3	football	0.0203525
4	team	0.0192997
5	league	0.0176893
6	player	0.0113382
7	club	0.011117
8	first	0.0110724
9	good	0.0104527
10	play	0.00949063

topic43 (Football)

Rank	Word	Proportion
1	player	0.0154346
2	goal	0.0147996
3	game	0.0120858
4	kane	0.00999203
5	play	0.00929872
6	team	0.00821765
7	need	0.00748892
8	england	0.00706033
9	world	0.00702132
10	harry	0.00641685

topic44 (Sea / water)

Rank	Word	Proportion
1	beach	0.0211838
2	west	0.0211572
3	east	0.0205149
4	sussex	0.0153045
5	south	0.0149244
6	river	0.0138793
7	kent	0.0137197
8	thames	0.0136475
9	sunset	0.0109909
10	norfolk	0.00936055

topic45 (Social media trends in the UK)

Rank	Word	Proportion
1	london	0.202401
2	united	0.180063
3	kingdom	0.16444
4	trending	0.0205803
5	kent	0.0104907
6	love	0.00473847
7	bromley	0.00356176
8	summer	0.00316061
9	england	0.00308524
10	pride	0.0029418

topic46 (Finances)

Rank	Word	Proportion
1	money	0.0142173
2	year	0.0083771
3	time	0.00750966
4	people	0.00606699
5	much	0.00557503
6	price	0.00525472
7	good	0.00513211
8	company	0.00509584
9	dont	0.00499469
10	business	0.00495178

topic47 (Trump)

Rank	Word	Proportion
1	trump	0.0706889
2	president	0.0185291
3	donald	0.0125607
4	london	0.0112398
5	protest	0.00835643
6	putin	0.00799318
7	russia	0.00742311
8	american	0.00635075
9	america	0.00622561
10	news	0.00609005

topic48 (Love Island)

Rank	Word	Proportion
1	loveisland	0.0685111
2	alex	0.0200316
3	jack	0.0152403
4	georgia	0.0151647
5	girl	0.0142717
6	love	0.0139772
7	laura	0.0134864
8	megan	0.0123435
9	shes	0.0119632
10	adam	0.0103158

topic49 (None2)

Rank	Word	Proportion
1	said	0.0116233
2	never	0.00798165
3	know	0.00786619
4	someone	0.0070777
5	word	0.00641451
6	time	0.00607518
7	well	0.00607377
8	thought	0.00591936
9	name	0.00521018
10	think	0.00504497

topic50 (Weather)

Rank	Word	Proportion
1	rain	0.0355552
2	today	0.0310713
3	weather	0.0284343
4	last	0.0260199
5	heat	0.0195823
6	london	0.0191305
7	summer	0.015766
8	wind	0.0137944
9	heatwave	0.0128353
10	pressure	0.0102674

topic51 (None3)

Rank	Word	Proportion
1	look	0.0136325
2	dont	0.00638731
3	take	0.00616396
4	think	0.00611498
5	time	0.00606374
6	back	0.00577755
7	right	0.00545301
8	know	0.00463764
9	love	0.00446747
10	need	0.00446554

topic52 (Racing)

Rank	Word	Proportion
1	goodwood	0.0267274
2	race	0.0252103
3	silverstone	0.0177147
4	festival	0.0160402
5	speed	0.015808

6	ascot	0.0122643
7	racing	0.0105936
8	circuit	0.010455
9	car	0.00961216
10	porsche	0.00840222

topic53 (None4)

Rank	Word	Proportion
1	time	0.026391
2	need	0.0149213
3	good	0.0122182
4	going	0.0109732
5	great	0.0106056
6	keep	0.010154
7	thing	0.0100166
8	dont	0.00864838
9	work	0.00847725
10	think	0.00822478

topic54 (Social media)

Rank	Word	Proportion
1	tweet	0.0243711
2	twitter	0.0223635
3	people	0.0138482
4	read	0.0125811
5	account	0.0110707
6	post	0.0108861
7	news	0.0083146
8	dont	0.00826399
9	follow	0.00773212
10	medium	0.00716828

topic55 (Fashion)

Rank	Word	Proportion
1	dress	0.015808
2	wearing	0.0157315
3	wear	0.014826
4	summer	0.0135684
5	fashion	0.0132416
6	shirt	0.0113686
7	london	0.0112188
8	style	0.01046
9	short	0.00839256
10	white	0.00820128

topic56 (None5)

Rank	Word	Proportion
1	back	0.0208944
2	home	0.0183101
3	london	0.00943111
4	today	0.00851755
5	going	0.00780828
6	work	0.00758805
7	come	0.00684763
8	time	0.00669006
9	away	0.00646686
10	coming	0.00587178

topic57 (World news)

Rank	Word	Proportion
1	world	0.0048019
2	moon	0.00479048
3	cave	0.00427343
4	boy	0.00426038
5	news	0.00417557
6	history	0.00414458
7	life	0.00401735
8	year	0.00398636
9	animal	0.00391133
10	found	0.00380205

topic58 (Place check-ins)

Rank	Word	Proportion
1	london	0.171695
2	greater	0.0227078
3	bridge	0.0154258
4	city	0.0140036
5	tower	0.0106455
6	view	0.00923385
7	street	0.00816188
8	england	0.00715889
9	summer	0.00682032
10	station	0.00659425

topic59 (Religion)

Rank	Word	Proportion
1	lord	0.0125351
2	jesus	0.0104546
3	khan	0.0101492
4	church	0.0099655
5	pakistan	0.00981654
6	allah	0.00871671
7	christ	0.00663871
8	muslim	0.00651458
9	imran	0.00522358
10	name	0.00510193

topic60 (Tennis)

Rank	Word	Proportion
1	wimbledon	0.0511068
2	tennis	0.0288127
3	final	0.0181147
4	court	0.0167833
5	match	0.0146778
6	golf	0.0137869
7	club	0.0121025
8	centre	0.0101673
9	today	0.00934947
10	great	0.00823708

topic61 (LGBT themes)

Rank	Word	Proportion
1	u200d	0.246421
2	pride	0.0663033
3	london	0.0305225
4	love	0.022427
5	happy	0.0179668
6	lgbt	0.00952375
7	prideinlondon	0.00936337
8	darling	0.00795811
9	amazing	0.00780154
10	rainbow	0.00767171

topic62 (Event announcements)

Rank	Word	Proportion
1	week	0.0292726
2	next	0.0258615
3	back	0.0180056
4	year	0.0140343
5	time	0.0135182
6	come	0.0131951
7	today	0.0126441
8	tomorrow	0.0125514
9	going	0.0121003
10	excited	0.0112728

topic63 (Sustainability)

Rank	Word	Proportion
1	water	0.0255832
2	food	0.0136566
3	plastic	0.0125506
4	drink	0.0115632
5	bottle	0.011361
6	need	0.00777683
7	dont	0.00591475
8	coffee	0.00525
9	free	0.004801
10	clean	0.00441809

topic64 (Football World Cup)

Rank	Word	Proportion
1	world	0.0299418
2	england	0.0219393
3	worldcup	0.0213209
4	game	0.0211539
5	team	0.0145895
6	france	0.0133827
7	final	0.0122462
8	croatia	0.00964906
9	goal	0.00908817
10	belgium	0.00851136

topic65 (Family)

Rank	Word	Proportion
1	love	0.0220328
2	girl	0.0138043
3	friend	0.0120948
4	little	0.010639
5	baby	0.00827129
6	know	0.00776281
7	look	0.00774916
8	shes	0.00774021
9	year	0.00690875
10	family	0.00632729

topic66 (Fitness)

Rank	Word	Proportion
1	training	0.0195486
2	session	0.0160956
3	class	0.0156806
4	fitness	0.0154188
5	u200d	0.0133367
6	body	0.0117316
7	today	0.0115542
8	week	0.0107207
9	yoga	0.0100611
10	morning	0.00963914

topic67 (Traffic information)

Rank	Word	Proportion
1	london	0.0384804
2	road	0.0212687
3	street	0.0196827
4	park	0.0113544

5	station	0.0107784
6	today	0.0106234
7	high	0.00747879
8	town	0.00705585
9	green	0.0066229
10	centre	0.00659008

topic68 (Sport matches)

Rank	Word	Proportion
1	goal	0.0190003
2	time	0.0169121
3	first	0.015283
4	ball	0.0140335
5	half	0.0130277
6	second	0.0117641
7	shot	0.0110816
8	away	0.0110519
9	take	0.0102007
10	great	0.00991173

topic69 (Brexit)

Rank	Word	Proportion
1	brexit	0.0410497
2	deal	0.0126819
3	boris	0.00852373
4	government	0.00819442
5	minister	0.00770216
6	tory	0.00713823
7	theresa	0.00656151
8	johnson	0.00635078
9	plan	0.00618186
10	trade	0.00560429

topic70 (Series)

Rank	Word	Proportion
1	love	0.0554563
2	watching	0.0265829
3	watch	0.0244529
4	island	0.0209823
5	show	0.0176614
6	episode	0.0171594
7	season	0.0127796
8	watched	0.0102919
9	series	0.0102908
10	last	0.0099672

topic71 (Food (home-made))

Rank	Word	Proportion
1	chicken	0.0144864
2	salad	0.0110028
3	cheese	0.00919069
4	tomato	0.00911973
5	lunch	0.00814337
6	today	0.00697778
7	potato	0.00683849
8	sauce	0.00670445
9	fresh	0.00615779
10	beef	0.00580562

topic72 (Social issues)

Rank	Word	Proportion
1	people	0.0478405
2	woman	0.0242093
3	black	0.0161113
4	dont	0.0124995
5	white	0.0119582
6	many	0.0114365
7	young	0.00782465
8	think	0.0070395
9	girl	0.0062242
10	need	0.00597253

topic73 (Football world cup England)

Rank	Word	Proportion
1	england	0.0696254
2	home	0.0411014
3	coming	0.0326439
4	worldcup	0.0241884
5	football	0.0240795
6	come	0.0230241
7	itscominghome	0.0115188
8	world	0.0112563
9	game	0.010534
10	threelions	0.00964985

topic74 (None6)

Rank	Word	Proportion
1	people	0.0160562
2	going	0.0117834
3	youre	0.0102458
4	know	0.00976998
5	feel	0.00925552
6	thing	0.00814317
7	really	0.00804677
8	cant	0.00780527
9	getting	0.0076485

10	stop	0.00728691

topic75 (None7)

Rank	Word	Proportion
1	make	0.0710451
2	feel	0.0237905
3	better	0.0221724
4	much	0.0200208
5	look	0.0178682
6	made	0.0177831
7	sure	0.0172013
8	making	0.0110134
9	even	0.0109846
10	good	0.00996272

topic76 (Open air concerts)

Rank	Word	Proportion
1	park	0.138925
2	hyde	0.0322346
3	london	0.0296772
4	summer	0.0126091
5	finsbury	0.0120151
6	hydepark	0.00881662
7	time	0.00840006
8	regent	0.00804146
9	festival	0.0073641
10	queen	0.00691856

topic77 (Wedding)

Rank	Word	Proportion
1	wedding	0.0370103
2	beautiful	0.0280236
3	love	0.016262
4	photo	0.0120671
5	look	0.0119523
6	gorgeous	0.0105764
7	today	0.00931
8	lovely	0.00894791
9	stunning	0.00844983
10	model	0.00804535

topic78 (Body issues)

Rank	Word	Proportion
1	head	0.00879355
2	back	0.00813462
3	dont	0.00699654
4	hand	0.00686372
5	feel	0.00665166
6	heart	0.00583424
7	face	0.00582837
8	today	0.00560163
9	still	0.00534188
10	need	0.00510267

topic79 (Death)

Rank	Word	Proportion
1	news	0.0160809
2	soul	0.0148532
3	heart	0.0146857
4	family	0.0140675
5	aretha	0.0114764
6	hear	0.011228
7	queen	0.0110143
8	rest	0.0107138
9	peace	0.0101708
10	legend	0.00935909

topic80 (Acknowledgements / gratitude)

Rank	Word	Proportion
1	thank	0.0696732
2	much	0.0539873
3	thanks	0.0493028
4	love	0.0305402
5	really	0.0136171
6	please	0.0133994
7	support	0.0125094
8	lovely	0.0112902
9	great	0.0102495
10	amazing	0.00958571

topic81 (Sales)

Rank	Word	Proportion
1	available	0.0150942
2	free	0.0118867
3	book	0.0110243
4	sale	0.0107587
5	shop	0.0102532
6	today	0.010045
7	online	0.00884427
8	store	0.00708875
9	ticket	0.00705799
10	summer	0.00704056

topic82 (Health)

Rank	Word	Proportion
1	health	0.020984
2	mental	0.0131497
3	care	0.0113265

4	people	0.0103281
5	help	0.00995371
6	need	0.00916909
7	child	0.00738281
8	issue	0.00656129
9	patient	0.00622379
10	life	0.00611852

topic83 (Customer service)

Rank	Word	Proportion
1	email	0.0169898
2	service	0.0153667
3	please	0.0146309
4	still	0.0110188
5	customer	0.0103277
6	order	0.0101252
7	call	0.00983817
8	card	0.00863521
9	today	0.00773757
10	phone	0.00691486

topic84 (Social system)

Rank	Word	Proportion
1	country	0.0123367
2	people	0.00959545
3	world	0.00929086
4	right	0.00744883
5	state	0.00618261
6	british	0.00617131
7	need	0.00444498
8	government	0.00405483
9	child	0.00352693
10	dont	0.0035011

topic85 (Travelling)

Rank	Word	Proportion
1	london	0.0321585
2	time	0.0130539
3	holiday	0.0108234
4	back	0.0103252
5	week	0.0102323
6	trip	0.00939293
7	travel	0.00938984
8	place	0.00911055
9	love	0.00870748
10	year	0.00804213

topic86 (Drinking)

Rank	Word	Proportion
1	drinking	0.0534841
2	beer	0.0356788
3	wine	0.0130432
4	drink	0.0106543
5	nice	0.00946855
6	pale	0.00893322
7	london	0.008159
8	photo	0.00755392
9	good	0.00739698
10	festival	0.00712844

topic87 (Job offers)

Rank	Word	Proportion
1	hiring	0.0782561
2	england	0.0598393
3	london	0.0538295
4	careerarc	0.0503576
5	latest	0.0438273
6	click	0.0404101
7	apply	0.0350894
8	opening	0.0272831
9	job	0.0233692
10	work	0.0193793

topic88 (Celebrities)

Rank	Word	Proportion
1	john	0.00732997
2	great	0.00584653
3	year	0.00559197
4	david	0.00499207
5	james	0.00492205
6	thanks	0.00387178
7	good	0.00374658
8	think	0.00371097
9	well	0.00356916
10	look	0.00331519

topic89 (Business)

Rank	Word	Proportion
1	need	0.0102674
2	work	0.00844189
3	great	0.00728527
4	good	0.00603245
5	think	0.00519724
6	change	0.00451896
7	also	0.00387652
8	really	0.00385578
9	well	0.00377127

10	help	0.00359925

topic90 (Technology)

Rank	Word	Proportion
1	phone	0.0164902
2	using	0.00765081
3	time	0.00693589
4	apple	0.00655162
5	dont	0.0063606
6	cant	0.00595287
7	google	0.00588584
8	iphone	0.00574509
9	work	0.00570153
10	need	0.00548258

topic91 (Retrospect)

Rank	Word	Proportion
1	year	0.0373327
2	last	0.0345719
3	time	0.0325686
4	first	0.0184752
5	week	0.0148628
6	today	0.0120593
7	night	0.0110021
8	still	0.0092442
9	didnt	0.00800293
10	month	0.00724254

topic92 (Weatherbots)

Rank	Word	Proportion
1	fine	0.075303
2	rain	0.0695997
3	slowly	0.0651087
4	forecast	0.0624409
5	today:00mm	0.0570406
6	weather	0.0430797
7	falling	0.0394022
8	settled	0.0390969
9	uv:0	0.0373585
10	rising	0.0348533

topic93 (Boxing)

Rank	Word	Proportion
1	live	0.117675
2	stream	0.0558439
3	fight	0.0486887
4	whyte	0.0398814
5	parker	0.0331362
6	link	0.0320306
7	boxing	0.0308816
8	watch	0.0267885
9	fury	0.0216084
10	online	0.0207761

topic94 (Free time)

Rank	Word	Proportion
1	morning	0.0270156
2	summer	0.0199852
3	london	0.0180891
4	good	0.0179234
5	weekend	0.0155285
6	sunday	0.0140346
7	lovely	0.0124521
8	today	0.0112002
9	happy	0.0109707
10	beautiful	0.0106323

topic95 (Education)

Rank	Word	Proportion
1	school	0.0336918
2	year	0.0225181
3	student	0.0163714
4	today	0.0131741
5	child	0.0092617
6	teacher	0.00835976
7	summer	0.00819513
8	class	0.00793407
9	course	0.00761421
10	kid	0.0075025

topic96 (Felicitations)

Rank	Word	Proportion
1	good	0.0422589
2	hope	0.03355
3	well	0.0264376
4	great	0.0181104
5	youre	0.0180145
6	thanks	0.0134469
7	back	0.0110838
8	soon	0.0109623
9	love	0.0109348
10	really	0.00934845

topic97 (None8)

Rank	Word	Proportion
1	think	0.0158651
2	well	0.0142311
3	good	0.0139297

4	know	0.0129604
5	thats	0.0123769
6	never	0.0109107
7	really	0.0107726
8	people	0.00834815
9	always	0.00773427
10	thought	0.00752308

topic98 (Plants / gardens)

Rank	Word	Proportion
1	garden	0.0268162
2	flower	0.0140393
3	beautiful	0.010916
4	tree	0.0106402
5	nature	0.00975027
6	today	0.00886424
7	summer	0.00824611
8	park	0.00782923
9	plant	0.00738752
10	green	0.00734962

topic99 (Ask for support)

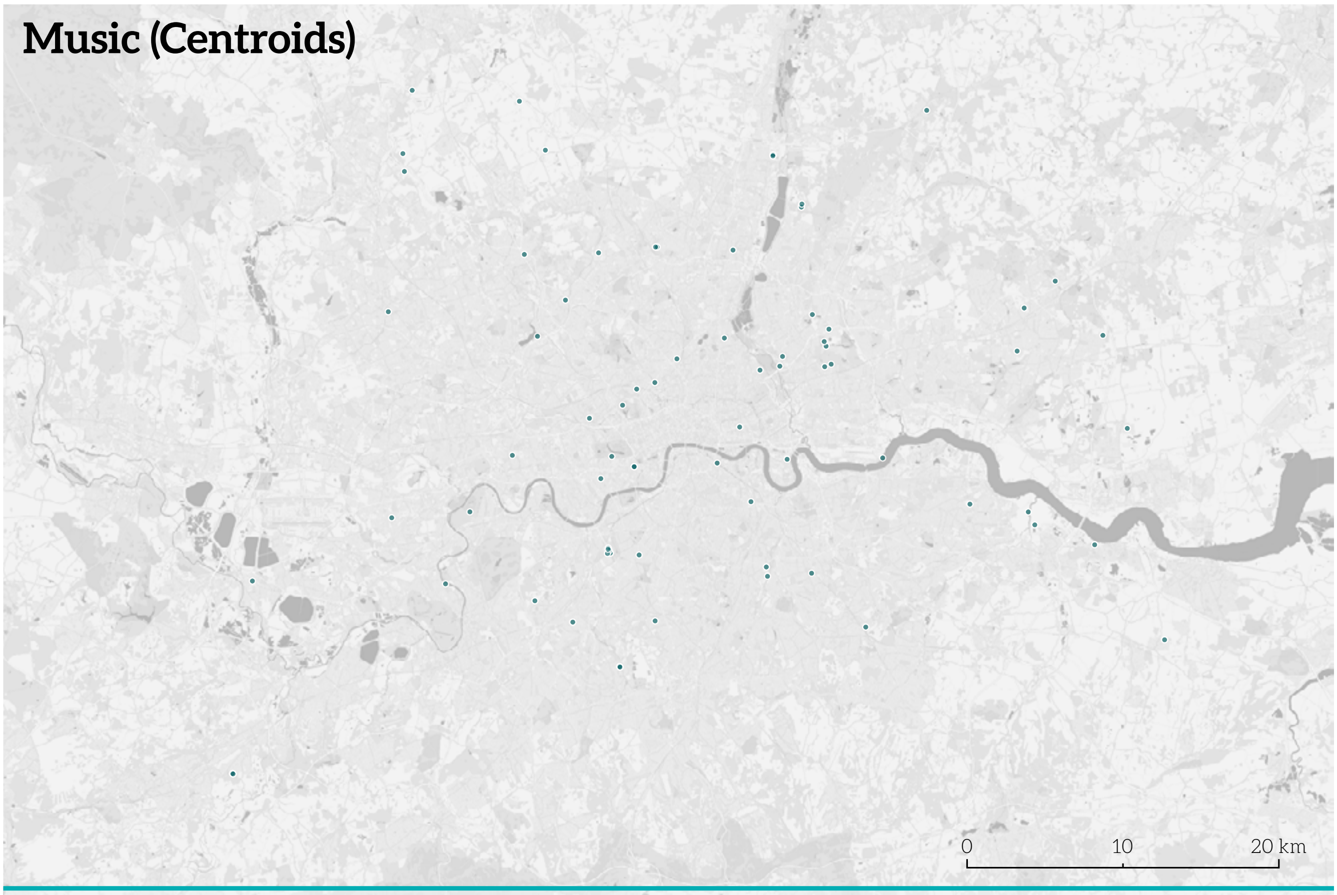
Rank	Word	Proportion
1	please	0.0495782
2	need	0.028813
3	dont	0.0188008
4	help	0.0185409
5	someone	0.0170384
6	want	0.0162139
7	anyone	0.0126639
8	know	0.0118373
9	come	0.0118225
10	give	0.0100819

topic100 (Desserts)

Rank	Word	Proportion
1	cake	0.0200705
2	chocolate	0.0167234
3	coffee	0.016551
4	cream	0.0139167
5	vegan	0.00858613
6	breakfast	0.00737152
7	fruit	0.00711524
8	strawberry	0.0067782
9	made	0.00672383
10	milk	0.00662132

Queried Amenities

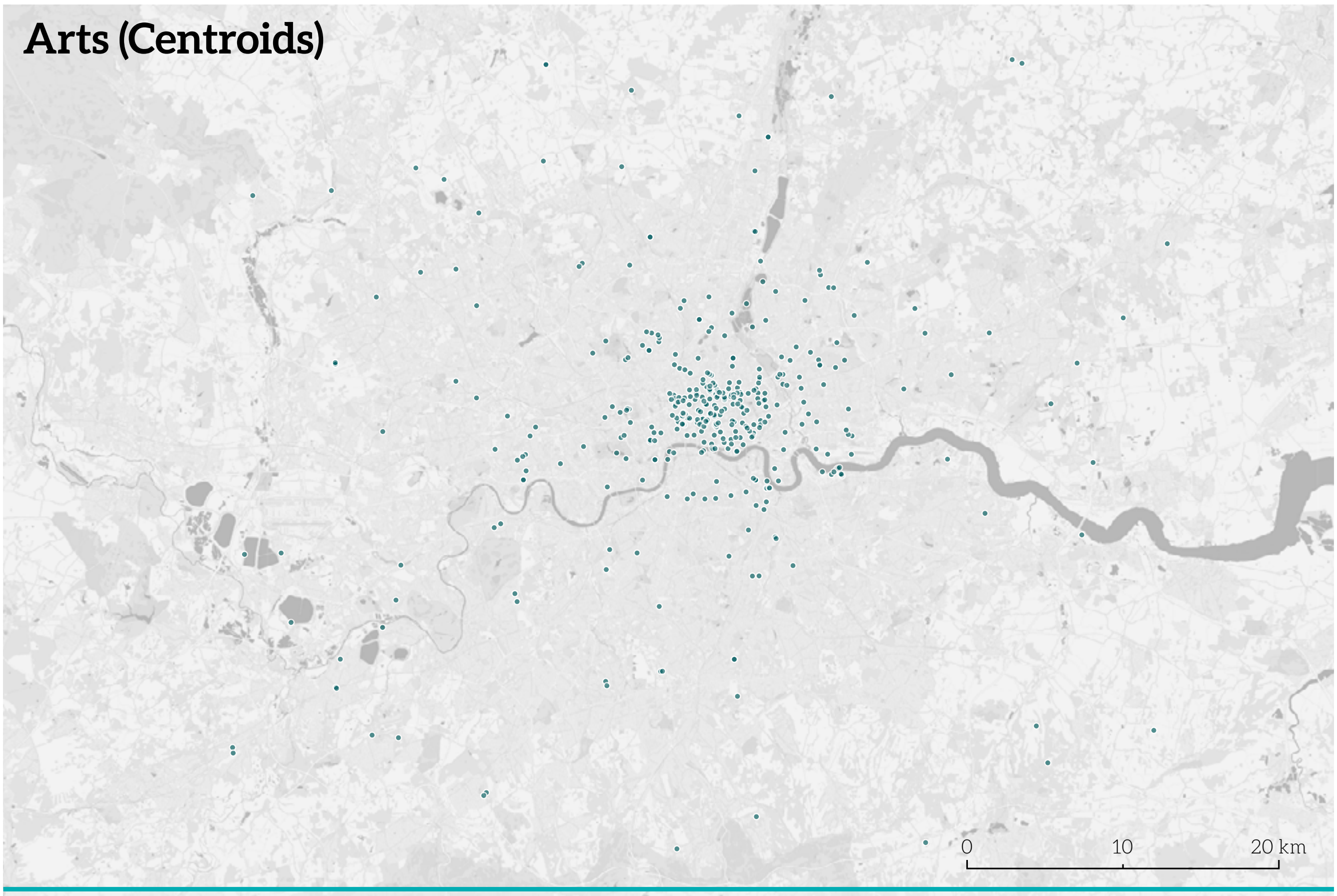
Music (Centroids)



Features: Amenity : concert_hall, music_venue, stadium (O2-Arena)

Source: OpenStreetMap Contributors | Basemap: CartoDB

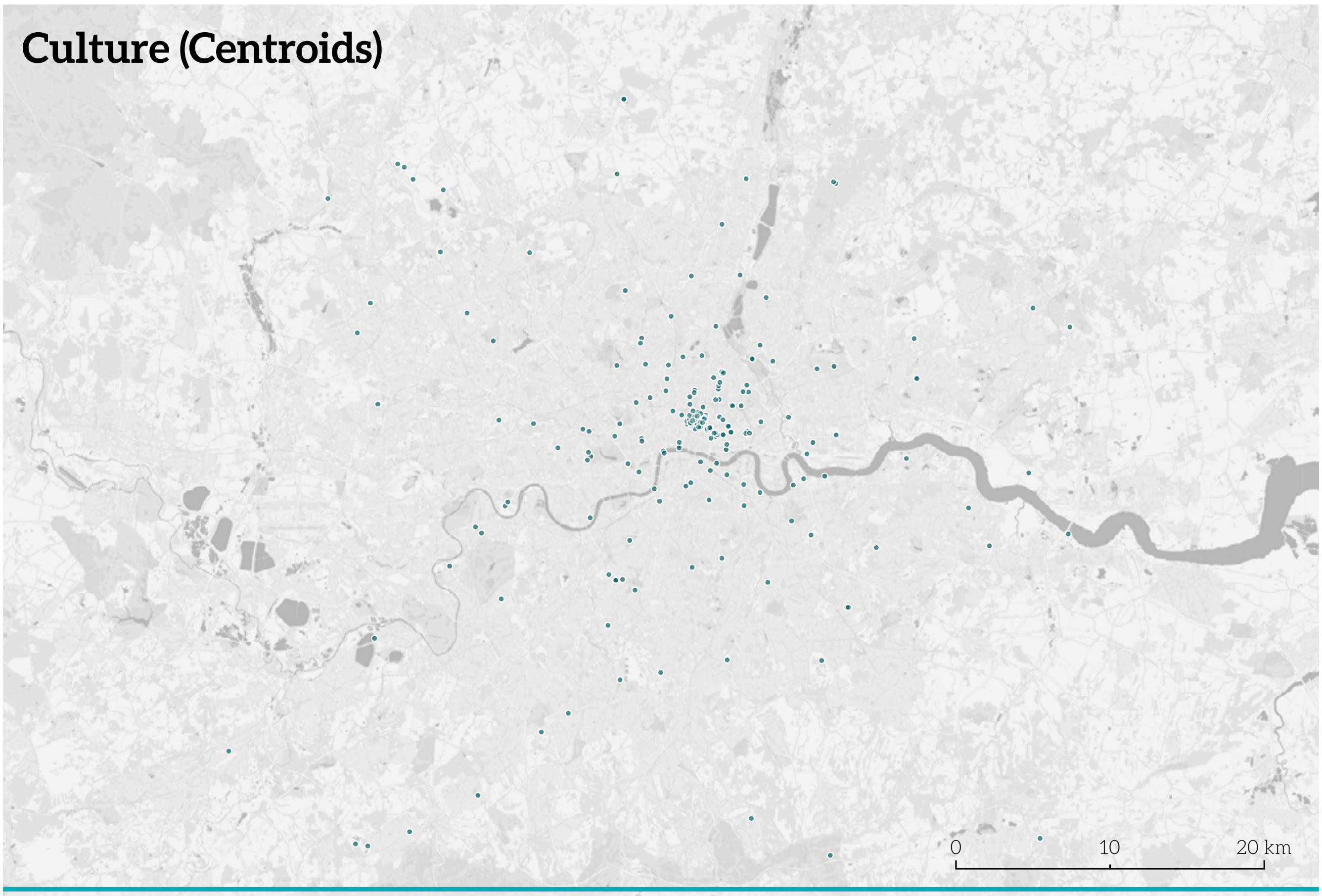
Arts (Centroids)



Features: Amenity : arts_centre / Tourism : museum, gallery

Source: OpenStreetMap Contributors | Basemap: CartoDB

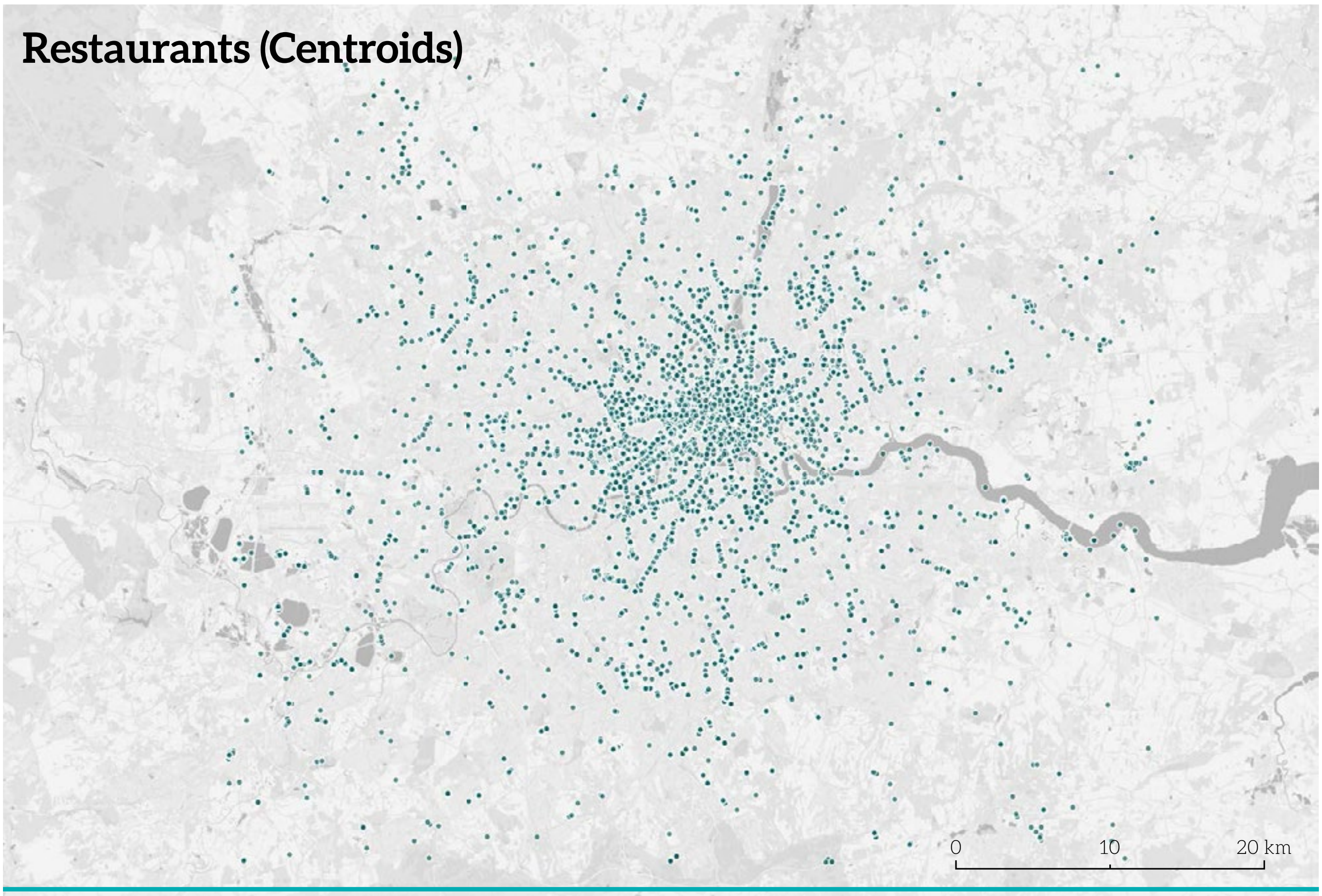
Culture (Centroids)



Features: Amenity : concert_hall, theatre

Source: OpenStreetMap Contributors | Basemap: CartoDB

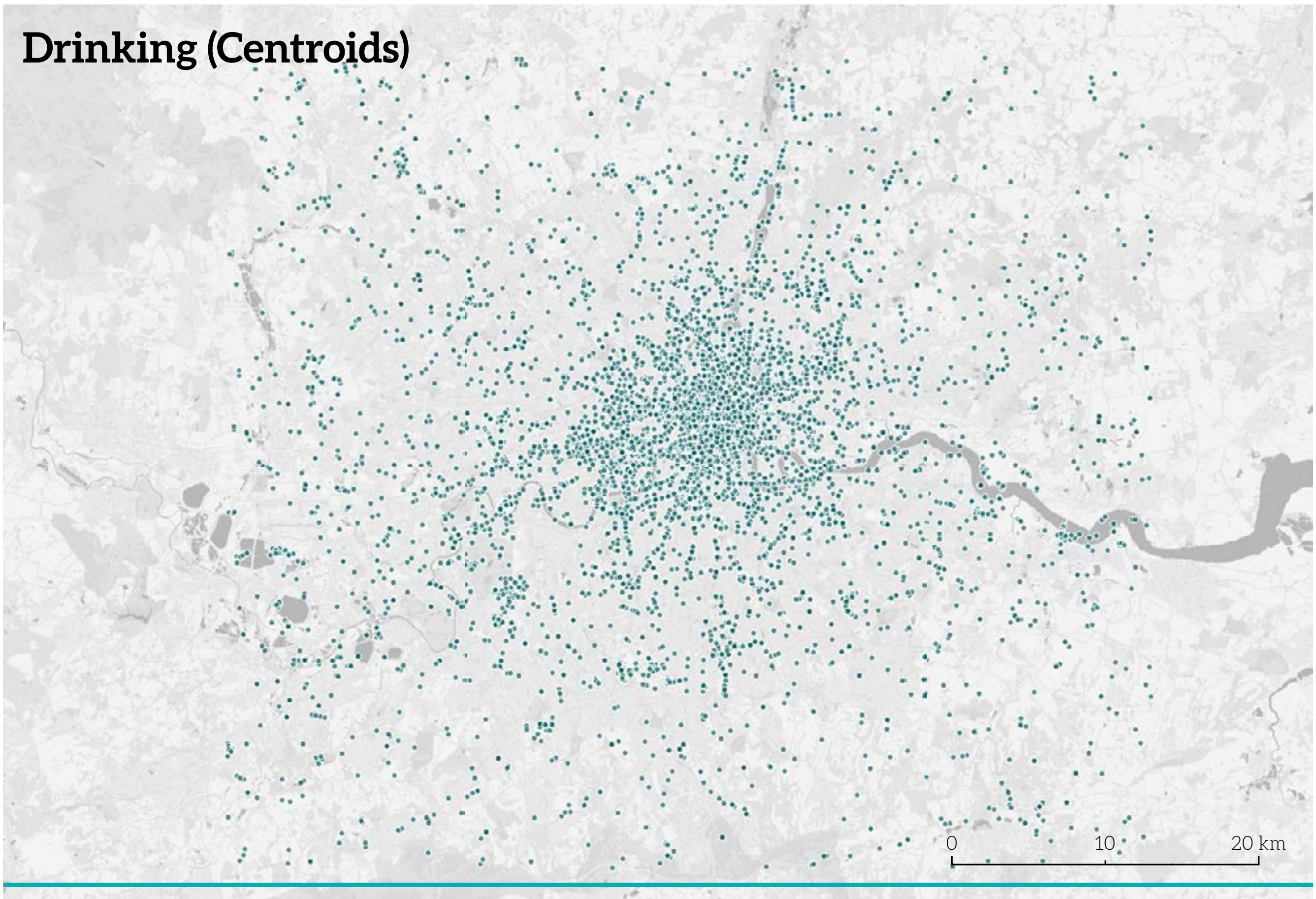
Restaurants (Centroids)



Features: Amenity : restaurant, fast_food, food_court

Source: OpenStreetMap Contributors | Basemap: CartoDB

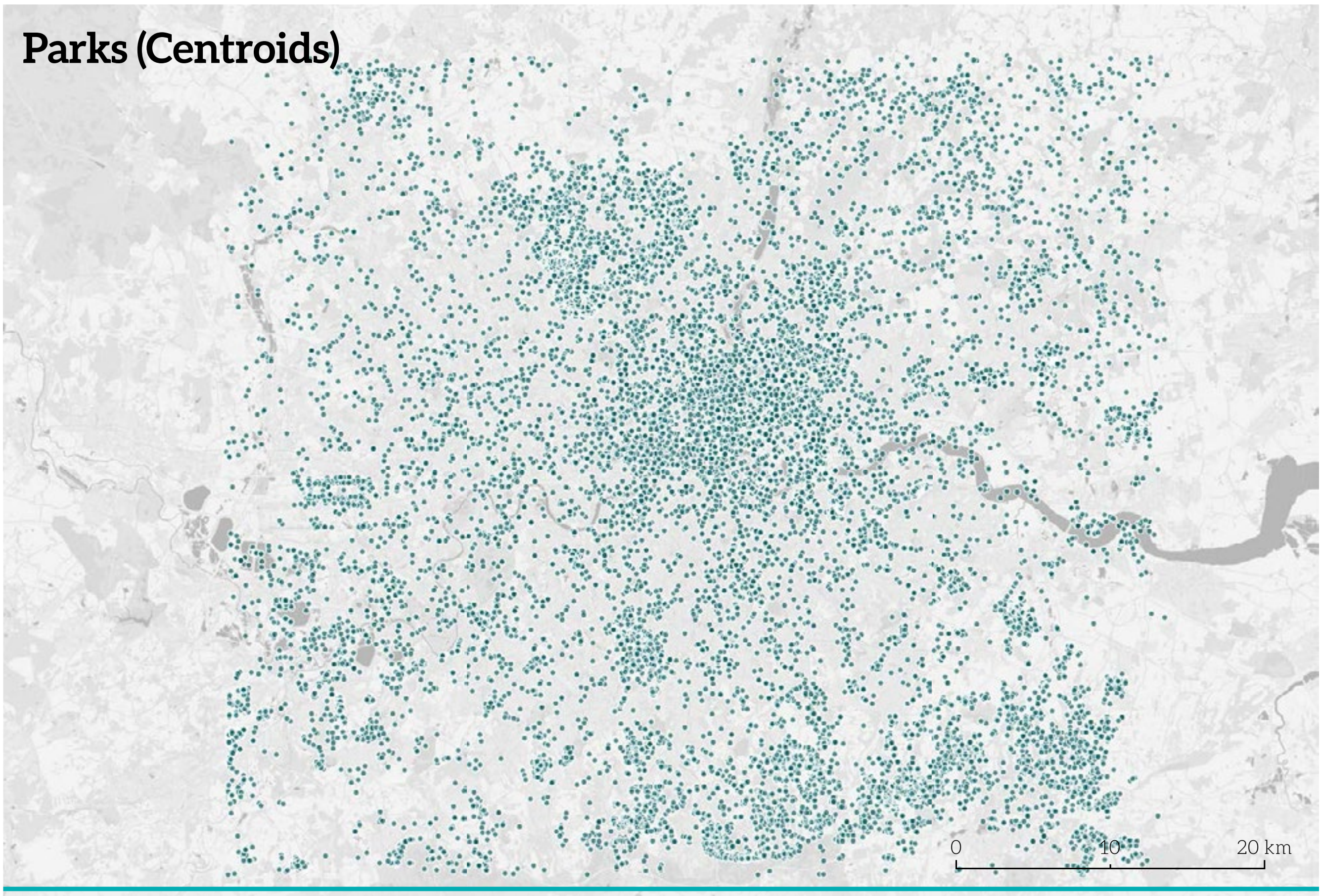
Drinking (Centroids)



Features: Amenity : bar, pub, cafe, biergarten

Source: OpenStreetMap Contributors | Basemap: CartoDB

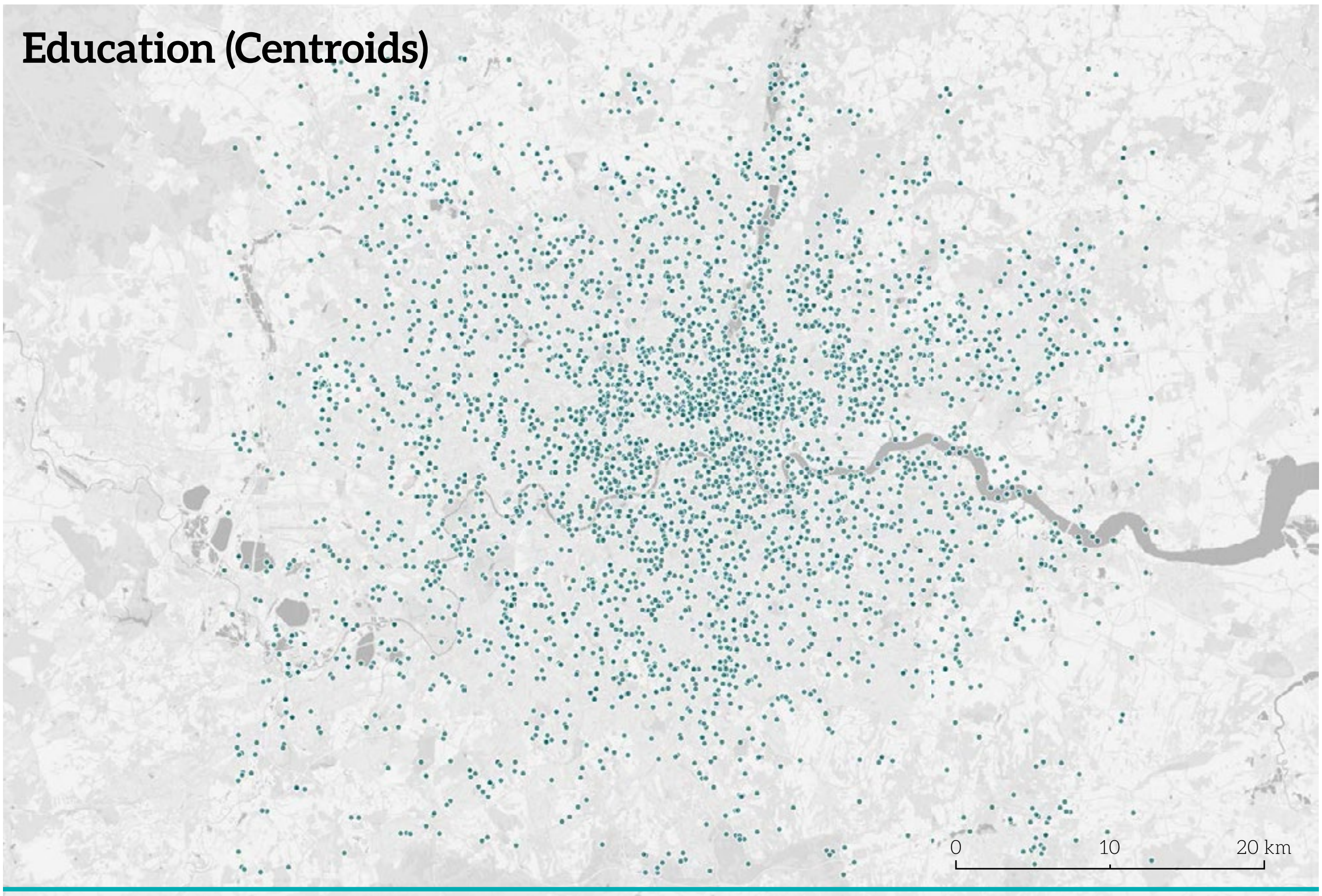
Parks (Centroids)



Features: Leisure : garden, park / Landuse: meadow, forest, grass

Source: OpenStreetMap Contributors | Basemap: CartoDB

Education (Centroids)



Features: Amenity : university, college, school

Source: OpenStreetMap Contributors | Basemap: CartoDB