TECHNISCHE
UNIVERSITÄT
DRESDEN

Faculty of Environmental Sciences Institute of Cartography

MASTER THESIS

Categorization and visualization of Twitter data

submitted by

Raluca Georgeta Nicola

14.01.1988, Bucharest - 3837394

Institute of Cartography

Technical University of Dresden

24 September 2013

Advisors: Prof. Dr. Ing. Dirk Burghardt, Dr. Alessio Bertone

TECHNISCHE UNIVERSITÄT DRESDEN

Fakultät Umweltwissenschaften Institut für Kartographie

MASTERARBEIT

Kategorisierung und Visualisierung von Twitter Daten

eingereicht von

Raluca Georgeta Nicola

14.01.1988, Bukarest - 3837394

Institut für Kartographie

Technische Universität Dresden

24. September 2013

Betreuer: Prof. Dr. Ing. Dirk Burghardt, Dr. Alessio Bertone

STATEMENT OF AUTHORSHIP

Herewith I declare that I am the sole author of the thesis named „Categorization and visualization of Twitter data“ which has been submitted to the study commission of geosciences today. I have fully referenced the ideas and work of others, whether published or un-published. Literal or analogous citations are clearly marked as such.

(place, date) (signature)

ACKNOWLEDGEMENTS

I would like to express my gratitude to Prof. Dr.-Ing. Dirk Burghardt and Dr. Alessio Bertone for their guidance in conducting a scientific research and for their time and the patience they showed during our brain-storming sessions.

Special thanks to Dipl.-Ing. Ralf Klammer and Dipl.-Ing. Stefan Hahmann for their valuable suggestions and help in the practical part of my thesis.

My gratitude goes also to Dipl.-Ing. Stefan Peters, the coordinator of this Master's program, for his precious help during these 2 years of master, not only in matters of studies but also with administrative issues.

A special thank you goes to my parents, Gabi and Horea, for their constant encouragement and support and without the help of whom I could have not pursued this Master's degree.

I would like to also thank the German Academic Exchange Service (DAAD) for the financial support and the help in accommodating to the German culture and making my stay in Germany extremely enjoyable.

Dresden, 24 September 2013 Raluca Georgeta Nicola

ABSTRACT

With more than 500 million Tweets each day, Twitter has gained tremendous popularity in the past few years. The big amount of user generated content, the fact that some Tweets have a spatial location and the public access to the data through the Twitter API are advantages that make this micro-blogging network suitable for data mining of the messages in order to reveal patterns, interests and to analyze user behavior.

One goal of this thesis is to find an optimal method to classify georeferenced Twitter messages from Saxony during the week corresponding to the beginning of the summer floods in this region. Within this chosen period it is investigated to what extent events like floods are reflected in social media. It will be shown that the possibilities to identify such events are limited as the biggest proportion of Tweets is related to personal conversations or status of the user.

The second goal, after classifying the Tweets, is to create a spatial and temporal visualization of the different topics. In order to do that, an investigation of the available visualization tools was made. Based on these results, a web-mapping application was created, in which the Tweets are displayed as clusters. With this application the users can choose a time period, visualize the corresponding topics for that period and interact to see the content of individual Tweets.

Finally, an evaluation of the web-map is conducted on 16 participants with no cartographic background. The users have tested the application and then filled an online-survey to express their opinions on the application.

KURZFASSUNG

Mit mehr als 500 Millionen Tweets jeden Tag hat Twitter enorme Popularität in den letzten Jahren erreicht. Dieses Microblogging Netzwerk bietet einige Vorteile für Data-Mining aufgrund des großen Anteils von nutzergenerierten Inhalten, der Lokalisierung von manchen Tweets und des freien Zugangs zu den Daten durch Nutzung der Twitter API. Das Ziel könnte zum Beispiel die Erkennung von Mustern sowie die Analyse von Benutzerverhalten sein.

Ein Ziel dieser Arbeit ist es eine optimale Methode zu finden, um georeferenzierte Twitternachrichten zu klassifizieren. Die verwendeten Daten stammen aus Sachsen innerhalb des Zeitraumes, wo Frühjahrsfluten in diesem Gebiet aufgetreten sind. Diese bestimmte Zeitperiode wurde gewählt, weil so untersucht werden kann, inwiefern sich Ereignisse, wie Fluten, in sozialen Medien widerspiegeln. Es wird gezeigt, dass die Möglichkeiten solche Ereignisse zu identifizieren begrenzt sind, da die Mehrheit dieser Daten auf persönliche Konversationen oder aktuelle Zustände des Benutzers bezogen sind.

Nach der Klassifizierung der Tweets, ist das zweite Ziel eine räumliche und temporale Visualisierung von verschiedenen Themen zu erstellen. Dafür wurde zunächst eine Untersuchung von geeigneten Visualisierungswerkzeugen durchgeführt. Basierend auf diesen Ergebnissen, wurde eine Web-Mapping Anwendung hergestellt, welche die Tweets als Cluster visualisiert. Die Benutzer können durch diese Web-Anwendung eine Zeitperiode wählen, bekommen die entsprechenden Tweets visualisiert und erhalten zusätzliche Informationen durch Interaktionen.

Abschließend, wurde eine Evaluierung der Web-Anwendung durchgeführt. Dazu haben 16 Teilnehmer, ohne kartographischen Hintergrund, die Web-Anwendung ausprobiert und schilderten ihre Eindrücke anschließend in einer Online-Umfrage.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

UGC	User generated content
LDA	Latent Dirichlet Allocation
API	Application Programming Interface
Tf-idf	Term frequency - inverse document frequency
GeoTTM	Geographic Twitter Topic Model
NLP	Natural Language Processing
SVG	Scalable Vector Graphics
VML	Vector Markup Language (XML based document for visualization of 2D vector graphics - replaced by SVG)
XML	Extensible Markup Language
JSON	Javascript Object Notation
SHP	Shapefile ESRI specific format
KML	Keyhole Markup Language
.geoJSON	JSON format that stores also geographical attributes

1 INTRODUCTION

1.1 INTRODUCTION AND BACKGROUND

Extracting and interpreting information from user generated content is a trending topic in the scientific community and in the business world. Furthermore, data with a spatial component are even more valuable. This is proved by the numerous web applications that deal with processing and visualization of user generated content.

Twitter data in particular are easily accessible and part of the Tweets are georeferenced. Micro-blogging posts cover a broad range of topics. Although the classification of topics in Twitter data poses many challenges due to the limited size of Tweets, research in this field is abundant. The classification criteria vary depending on content or on user intentions. This master thesis focuses on the classification and visualization of the topics extracted from georeferenced Twitter data according to their content. The visualization method takes into account both spatial and time components. The area of interest was spatially limited to Saxony (Germany) and timely limited to the week of severe floods (31.05.-06.06.2013).

1.2 PURPOSE AND MOTIVATION

The motivation for this thesis was to get an insight into Twitter data analysis and visualization and to explore some of the methods and tools used for this. As the spatial visualization played an important role, only georeferenced Twitter data have been analyzed. The classification of the Tweets was indeed a challenge as Tweets are very short, use abbreviations or slang words. Besides the research for a good method for classifying the Tweets, a personal motivation was also to discover what people Tweet about in Germany and to see if real life events are reflected in the micro-blogging community. For this last reason, the time period for which the Tweets were processed is limited to the week of floods in Saxony.

The first purpose of this thesis is to identify the limitations of the current Twitter spatial tools. For this a comparison of the currently available applications that focus on analysis and visualization of georeferenced Twitter data was made.

After exploring several web applications that deal with analysis and visualization of Tweets, it was concluded that few geospatial visualizations of Twitter data show the topics that are discussed within the Tweets. Even fewer of them visualize topics on a longer time span, so that the user can browse through the data according to time. Derived from this conclusion, the second goal was to find a suitable method for Twitter data classification and to use it for discovering the general topics of discussion in the micro blogging environment. After the classification is done a new challenge is to find a suitable method for visualizing the topics spatially and temporally.

1.3 RESEARCH QUESTIONS

In the beginning of the master thesis, the following research questions were identified:

- How to extract the main topics from georeferenced Tweets? Which methods of topic extraction are available and which ones are the most suitable for Tweet classification?
- How to visualize these topics spatially and temporally in a way that will make them easy to be identified on a map within a user selected time period?

1.4 THESIS STRUCTURE

This thesis is divided in 5 main chapters. The first chapter introduces the topic of this thesis and shows an overview of the work with the expected results. The second chapter presents Twitter as a micro-blogging environment and its specific terminology. The state of the art in the third chapter is meant to show what is the latest research in topic classification of Tweets and in spatial visualization of Twitter data. The analysis of available tools was an important step in choosing the classification method and the visualization and it is presented in chapter 4. The next chapter, chapter 5, describes the workflow of the classification process. In the same chapter the implementation of the visualization web-map and a short evaluation of the results are presented. The conclusion and further research ideas come at the end in chapter 6.

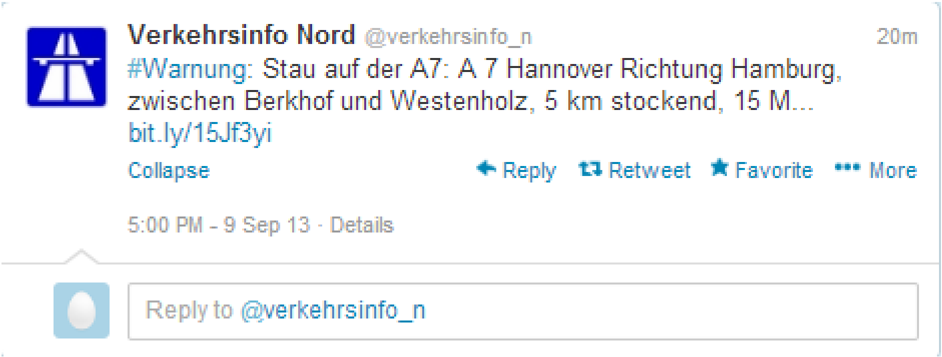
2 TWITTER

2.1 SOCIAL NETWORKS AND USER GENERATED CONTENT

The era of Web 2.0 is associated with the concept of "user generated content". Krumm (2008) define the concept of user generated content as *information, data or media contributed voluntarily by regular people with the purpose of making it available to others in a useful or entertaining way, usually on the web*. Social networks, like Twitter, are made entirely of user generated content. Aggregating all this data and analyzing it to discover patterns or new trends is what gives it real value and it is an ongoing research topic. However, as useful as it could be, user generated content can also be extremely noisy: many posts can be inaccurate, some incomplete or written with abbreviations or in slang and there is a prevalence of spam messages (Jackoway et al., 2011). Therefore research has been done in data preprocessing and filtering to make it usable and fit for analysis.

2.2 HOW TWITTER WORKS – SPECIFIC TERMINOLOGY

Twitter is a social networking website that was launched in 2006 and it has been continuously gaining popularity ever since (Lee et al., 2011). Twitter allows its registered users to post short messages also called **Tweets** (Figure 1). Messages are constrained to 140 characters and can only include text. However, files like photos or videos can be also added as a URL and usually to save space, URLs are shortened to tiny URLs.



Verkehrsinfo Nord @verkehrsinfo_n 20m

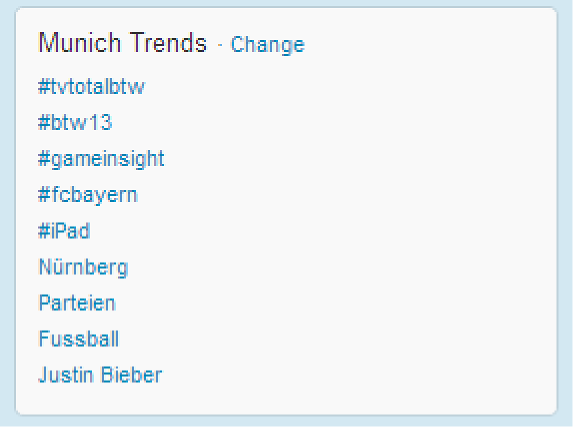
#Warnung: Stau auf der A7: A 7 Hannover Richtung Hamburg, zwischen Berkhof und Westenholz, 5 km stockend, 15 M... bit.ly/15Jf3yi

5:00 PM - 9 Sep 13 · Details

Figure 1: Screenshot of a Tweet containing Hashtags, a tiny URL and the options for replying and retweeting.

Each user creates its own network of contacts by **following** the accounts s/he is interested in (might be friends, family, public persons or institutions that have a Twitter account). Basically following means subscribing to another user's Tweets or updates. The user who follows can be followed back but this is not implied. This non reciprocity of user connections was unique to Twitter, but recently other social networks like Facebook have started using it. When one user follows another, s/he can read the Tweets that the latter posts. By default, Twitter accounts are public and therefore anyone, not necessarily only followers, can see a user's post. However, some users choose to set their posts as private and only their followers can see their posts in this case (Twitter, 2013b). Posts from private accounts will also not be seen in the search result list. Users can reply to other users' posts by using the **reply** function or by beginning the Tweet with the user's id preceded by "@". In this way conversations can be carried out on Twitter. Users also have the option to send each other **direct messages (DM)** which are private and unsearchable. But in order for a user to receive or send a direct message to another user, they both have to follow each other. Direct messages are also restricted to 140 characters.

Replying should not be confused with mentioning. When mentioning another user the syntax is the same: "@" symbol followed by user name but generally the mention is not in the beginning as it is for replying. Another available option on Twitter is to repost a message that another user posted. This is called **retweeting (RT)** and it is used when a user wants to share a post of another user with his followers (Figure 1). An important concept of Twitter is the **hashtag**. A hashtag is a part of the Tweet that denotes the topic about which the user Tweets. Normally it is preceded by a "#" (Figure 1). This is a useful way of labeling what the post is about and in this way users can create trends in conversations. Twitter.com uses hashtags to create a list of trending topics (Figure 2).



Munich Trends · Change

#tvtotalbtw

#btw13

#gameinsight

#fcbayern

#iPad

Nürnberg

Parteien

Fussball

Justin Bieber

Figure 2: Screenshot of Trending Topics on Twitter.com

This can be useful to see what the other users tweet about and these trends can be local or global. It is not recommended that a user associates a Tweet with too many hashtags or that s/he uses hashtags that are not relevant to the Tweet (Twitter, 2013a).

The **geolocation** of a Tweet can be embedded in the Tweet information if the user chooses to turn the location information on. A user can choose to share his exact location (latitude and longitude) or just the place s/he tweets from (neighborhood or city) (Twitter, 2013c). The place is manually selected by the user from a predefined list of locations supported by Twitter and is used when the user tweets from a desktop or fixed location device. The exact location is extracted from the mobile device's geolocation system and the user does not have to write anything, the location is provided automatically. Research done by (Leetaru et al., 2013) show that only approximately 1-2% of Tweets have exact geolocation.

In the last years, Twitter has gained a lot of popularity and that is what makes it a good source of information. From 6.0 million active users in 2006 (Wolf, 2011) Twitter has reached 200 million active users and 500 million Tweets per day in 2013 (Holt, 2013).

From all the social networks available Twitter was chosen as a source of data because of its freely available API used for data extraction and because of the popularity that the social network gained in the last years. The disadvantage is that although Twitter provides a great deal of user generated data, the micro-blogging platform is at the same time a very noisy medium and it is quite difficult to filter relevant information.

2.3 COLLECTION OF TWITTER DATA

Twitter provides an Application Programming Interface to access its data (Horn, 2010). The API is HTTP based and uses GET requests to retrieve data from Twitter. There are 3 types of APIs that allow for Twitter data retrieval. The Search API allows for queries against recent Tweets and it will find only Tweets that are no more than 7 days old and correspond to the used query (Twitter, 2013e). The REST API is used by many applications to post Tweets or follow someone. The Streaming API retrieves Tweets in real-time starting with the moment when the request is made. Several libraries in C++, Java, Javascript, PHP etc. are used to work with the Twitter API and retrieve data in a specific format and for different platforms (Twitter, 2013d).

The Twitter data collection that was needed for this project was done exclusively by Stefan Hahmann. For this, the Streaming API was used along with the Java library Twitter4J. This allowed access to approximately 1% of all Tweets. The search query was spatially limited with a bounding box filter that includes Germany, Switzerland as well as parts of Austria, the Netherlands and the Czech Republic. After the acquisition, each Tweet was post-processed in order to detect the language using the Apache Tika library (Hahmann & Burghardt, 2013). The Tweets were collected starting September 2012, but for my thesis, the period of interest was the flooding period in Saxony, so only Tweets for a period of 1 week from 31.05.2013 until 06.06.2013 were used.

3 STATE OF ART

The research in the field of information extracted from social networks is very broad and an exhaustive literature review would be impossible. The focus here is on research related specifically to classification and visualization of Twitter data. In the first part of the chapter, methods and literature related to Twitter data topic classification are presented. The second part describes different spatial visualizations of Twitter data by making an overview of existing Twitter visualization web applications.

3.1 TOPIC CLASSIFICATION – USER MOTIVATION ON TWITTER, CHALLENGES AND METHODS IN THE CLASSIFICATION OF TWEETS

Extensive research has been done in trying to understand what drives users to post messages on social networks and what they post about. In the context of providing voluntary geographic data, Goodchild (Goodchild, 2007) has mentioned that the main motivations of users are self-promotion - exposing your personal data to other people including your friends and family and sharing information that might be useful with others. Research in the field of micro-blogs such as Twitter shows that Twitter user intentions are somehow similar. Java et al. (Java et al., 2007) categorized manually the Tweets recorded for a period of 2 months and discovered users intentions quite close to Goodchild's results. According to their research, most posts on Twitter are related to social activity: daily chatter about what people are currently doing and about their daily activities. The second motivation would be sharing and seeking information. In another study (Alhadi et al., 2011) Alhadi et al. classified the Tweets into 8 categories. They combined a manual classification technique with an online survey for finding out the purposes of Twitter users. The classes are more specific than in the previous studies: social interaction with people, give/require feedback and express emotions fall under the umbrella of social activities. The other 5 categories, more related to sharing/seeking information are: sharing resources, promotion and marketing, broadcast alert/urgent information, require/raise funding and recruit worker. There are also Tweets that could not be classified. A similar study for content based categorization was conducted by Naaman et al.. In (Naaman et al., 2010) the authors distinguish 9 categories of message contents derived from a sample of 3379 extracted messages (Table 1).

Code	**Example**
Information Sharing(IS)	"15 Uses of Wordpress <URL>"
Self Promotion(SP)	"Check out my blog <URL>"
Opinions/Complaints (OC)	"Illmatic = greatest rap album ever"
Statements and random thoughts (RT)	"I miss NY but I love LA"
Me now(ME)	"tired and upset"
Question to followers (QF)	"what should my next video be about?"
Presence Maintenance (PM)	"I'm backkk!"
Anecdote me (AM)	"oh yes, I won an electric steamboat machine and a steam iron at the block party lucky draw this morning"
Anecdote others (AO)	"Most surprised <user> dragging himself up pre 7 am to ride his bike"

Table 1: Message categories according to Naaman et al. (Naaman et al., 2010)

An interesting fact is that 41% of the sample data used in the study was categorized as "Me now", which means that most people tweet about their current status or activity. About 20% of the messages shared information, which clearly shows this is not the main activity on Twitter.

In conclusion, social networks and Twitter in particular are mostly used for sharing and retrieving information of interest and for social interaction with other people. Online social interaction can be driven by the same common interests with another group of people, sharing personal information about oneself with friends or with all the other users.

The methods used for the classification of Tweets vary greatly. The fact that the Tweets are only 140 characters long make the task of classifying them or extracting content very difficult. Some of the Tweets contain words that are either abbreviated or belong to slang vocabulary. This makes an automatic classification of Tweets almost impossible.

Some of the methods used for discovering underlying topics in Twitter data are presented next. (Lee et al., 2011) use the Bag of Word approach to classify the trending topics provided by Twitter into more general classes. For this they defined

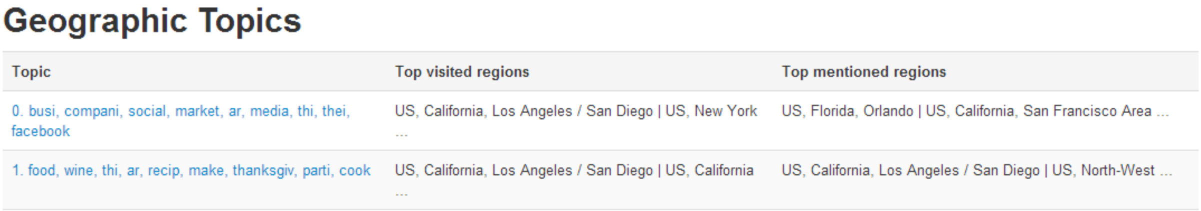
18 general classes: arts and design, books, business, charity and deals, fashion, food and drink, health, holidays and dates, humor, music, politics, religion, science, sports, technology, tv and movies, other news and other. They manually assigned the trending topics on Twitter to one of the categories above and built documents with the trend definition and the Tweets that belong to that trend. Then, for each document they created a word vector with the most frequent term and their tf-idf (term frequency - inverse document frequency) weights. Tf-idf is a method of assigning a weight to a word based on how frequent the word is not only in one document but in all the documents (Rajaraman et al., 2012). If the word appears many times in one document but does not appear at all in the others, then it will have a big weight. On the other hand, if the word appears often in all the documents, its importance is decreased. This method can also be used to remove stop words.

(Zhao et al., 2011) compare the topics on Twitter with the Topics in New York Times. For this they use standard LDA (Latent Dirichlet Allocation) to discover the topics in the newspaper and a customized algorithm of LDA for the Twitter topics. LDA is a probabilistic topic modeling algorithm that was developed by (Blei et al., 2003) in order to help with document modeling and text classification. LDA assumes that a collection of documents share the same set of topics, but in each document multiple topics can be found in different proportions. Each topic is a distribution over words and the LDA algorithm retrieves the most probable words for each topic (Blei, 2012). LDA used for classification of Tweets where each Tweet is considered a document is not so effective because the Tweets are too short. (Hong & Davidson, 2010) have proposed the solution to aggregate all Tweets into a single document. However, as each Tweet is usually just about 1 topic, (Zhao et al., 2011) have developed their own LDA based model which treats each Tweet as a single document. (Chae et al., 2012) use also a probabilistic approach by classifying Tweets using Mallet[1], an open source tool implemented in Java for document classification, topic modeling, information extraction and other machine learning applications. Prior to the topic extraction, the authors use also a stemming algorithm that is applied to

[1] http://mallet.cs.umass.edu/

each word in the Tweet messages. The stemming algorithm reduces the word to its stem, allowing for a more accurate topic extraction.

(Vosecky et al., 2013) develop their own probabilistic topic model called Geographic Twitter Topic Model (GeoTTM). This model extracts user interest topics based on the semantic dimension (words associated with the topic) and the geographic dimensions (regions that were visited or mentioned by the user). GeoTTM model assumes that a document represents all the Tweets of a single user and each Tweet can be assigned to one topic. The topics extracted in the end are a collection of associated words along with the regions mentioned or visited (Figure 3).



Geographic Topics

Topic	Top visited regions	Top mentioned regions
0. busi, compani, social, market, ar, media, thi, thei, facebook	US, California, Los Angeles / San Diego | US, New York ...	US, Florida, Orlando | US, California, San Francisco Area ...
1. food, wine, thi, ar, recip, make, thanksgiv, parti, cook	US, California, Los Angeles / San Diego | US, California ...	US, California, Los Angeles / San Diego | US, North-West ...

Figure 3: GeoTTM topics and associated regions - Screenshot of results in Limosa application (http://webproject2.cse.ust.hk:8031/limosa/main/topic/TopicListAction.list)

(Andrienko et al., 2013) use complementary methods to identify thematic patterns in georeferenced Tweets. The first one, called "Term-Usage Clustered Analysis", consists in finding the most prominent words by counting the frequency of the single words. For each term, a hotspot location is calculated using a clustering algorithm and the terms are plotted to this location. This method for the spatial investigation of Tweet content is extremely useful in discovering relations between places and topics of interest. Up to a certain extent it provides a similar visualization to Trendsmap[2]. A term frequency algorithm will also be used in this thesis as a method for discovering possible topics in the Tweets. However, no association is made between the words and the location where they have been tweeted. The second approach used in the same research is a "keyword based categorization". A list of topics commonly used in Twitter messages is created and each topic is represented by some keywords, that are normally associated to this topic. The Tweets are assigned to these topics based on the topic keywords. This method is really effective and it will be also employed during the classification of the Tweets in

[2] http://trendsmap.com/

this thesis. However, not all Tweets can be classified in this way, therefore a manual classification is also conducted.

Researchers like (Naaman et al., 2010) or (Java et al., 2007) perform a manual classification in order to extract topics from Twitter data. Another option, used by (Alhadi et al., 2011) is to classify the Tweets manually using the Amazon Mechanical Turk service[3], a platform on which people are paid for small tasks such as Tweet classification in this case.

3.2 VISUALIZATION OF TWEETS

The focus in this thesis is on visualizations of Twitter data with a spatial component. Associating a location with a Tweet can add much more value to the analysis of twitter data and help at interpretation. For example trends can be better understood if they are associated to a country or region. Representing these trends on a map rather than showing them in a continuous list gives an added value to the Tweet because they are easier to interpret by putting them in a certain context.

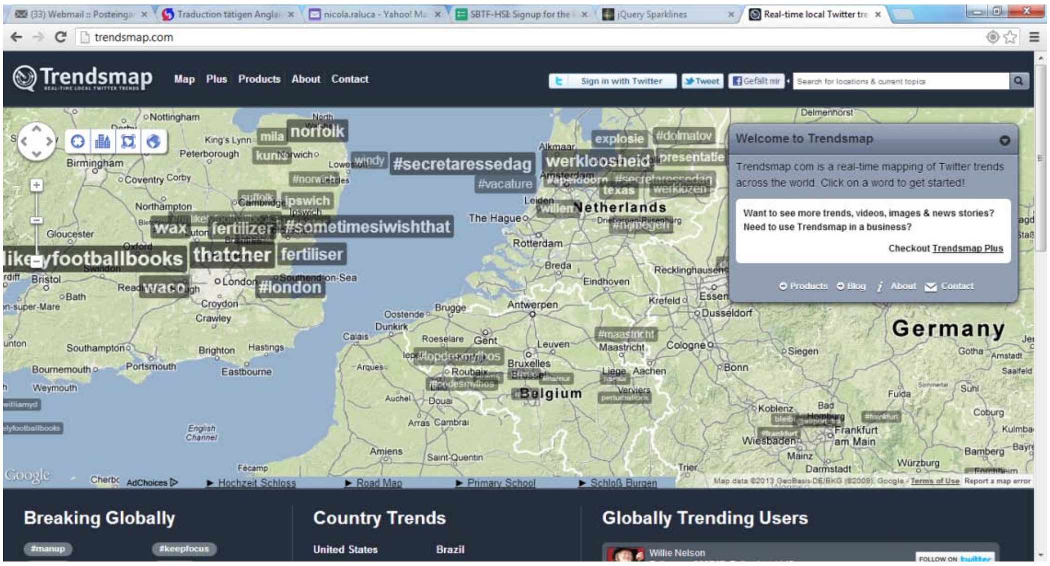
The most common approach of mapping Tweets is to visualize them as markers. Variations in the shape, size and colors of markers and the way they appear on the map are what make each of the visualizations unique. Other methods for visualizing Tweet density include heatmaps or smoky clouds. Some web tools that use these methods are presented in the following paragraphs.

One application that maps real-time Twitter trends around the world is **Trendsmap**[4]. This web application does not map the Tweets directly, but it rather groups them into trends and the trends are visualized on the map as text boxes (Figure 4). The size of the text and the background color represent the importance of the trend, which is calculated by how many people tweet about it. By clicking on one text box, a list of the latest Tweets is shown and also the links and images embedded in the Tweets are extracted and shown separately. Tweets are updated in real time, so if a trend decreases in popularity its text box will be shrunk and its visibility will decrease. New topics can show up and then the other topics will be

[3] https://www.mturk.com/mturk/

[4] http://trendsmap.com

slightly shifted to make place for the new topic. When the application was launched the temporal aspect was not present, as the application maps topics in real time. There were however small timelines with the popularity of the trend in the last 7 days, but they do not provide any user interaction. The new version of the app provides the possibility of scrolling back in time for up to 7 days, but that is only for registered users who pay a tax of 19$/month. Another interesting feature is that the user can select a city or a country and Trendsmap will show the specific trends in that area.



Welcome to Trendsmap

Trendsmap.com is a real-time mapping of Twitter trends across the world. Click on a word to get started!

Want to see more trends, videos, images & news stories?
Need to use Trendsmap in a business?

Checkout Trendsmap Plus

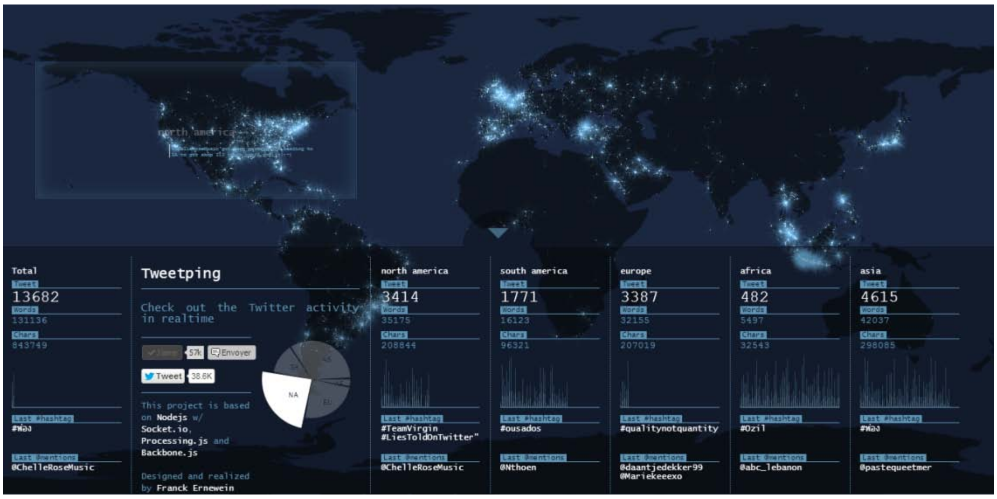
Figure 4: Screenshot of trendsmap.com, a web-application for the spatial visualization of Twitter Trends

Another application that focuses mostly on the visualization is **Tweetping**[5]. The web-application shows real time Tweets as bright pixels on a dark background map, making it look like a night satellite image of the Earth (Figure 5). There is no possibility for zooming in, therefore after letting the application run for a while, the Tweets tend to aggregate into clusters of illuminated areas and it is not possible to access the single Tweets.

It also provides a dashboard with statistics about the number of Tweets divided by continents, hashtags and mentions. The visualization method is unique and impressive. It is also noteworthy that it has no timeline as this is another real

[5] http://tweetping.net/

time application. Although there is a strong focus on the visualization of the Tweets, the application could also be useful to analyze which continent is using Twitter more.



Total	north america	south america	europe	africa	asia
Tweet 13682	Tweet 3414	Tweet 1771	Tweet 3387	Tweet 482	Tweet 4615
Words 131136	Words 35175	Words 16123	Words 32155	Words 5497	Words 42037
Chars 843749	Chars 208844	Chars 96321	Chars 207019	Chars 32543	Chars 298085
Last #hashtag #ห้อง	Last #hashtag #TeamVirgin #LiesToldOnTwitter"	Last #hashtag #ousados	Last #hashtag #qualitynotquantity	Last #hashtag #Ozil	Last #hashtag #ห้อง
Last @mentions @ChelleRoseMusic	Last @mentions @ChelleRoseMusic	Last @mentions @Nthoen	Last @mentions @daantjedekker99 @Mariekeeexo	Last @mentions @abc_lebanon	Last @mentions @pastequeetmer

Tweetping

Check out the Twitter activity in realtime

This project is based on Nodejs w/ Socket.io, Processing.js and Backbone.js

Designed and realized by Franck Ernewein

Figure 5: Screenshot of tweetping.net - visualizes Tweets in real-time all over the world and provides a series of statistics about the Tweets like number of words, of characters or last hashtag

Another spatial visualization of Twitter data is the **Onemilliontweetmap**[6]. This map displays in real-time the last 1 million Tweets. There are 2 visualization methods, one is by displaying the Tweets as clustered markers and another one is as heat map (Figure 6). Besides the visualization, the map provides also some user interaction. The user can specify a keyword or a hashtag and filter the Tweets with them. The displayed Tweets can also be filtered by the most popular hashtags, which are in this case retrieved by the application. A temporal component is also provided, the user being able to filter the Tweets by different time intervals up to the last 6 hours. The map can also be reset for a better visualization of each Tweet.

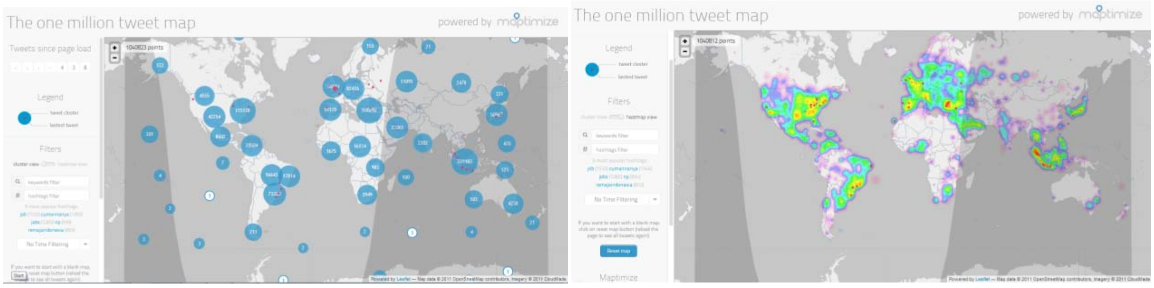
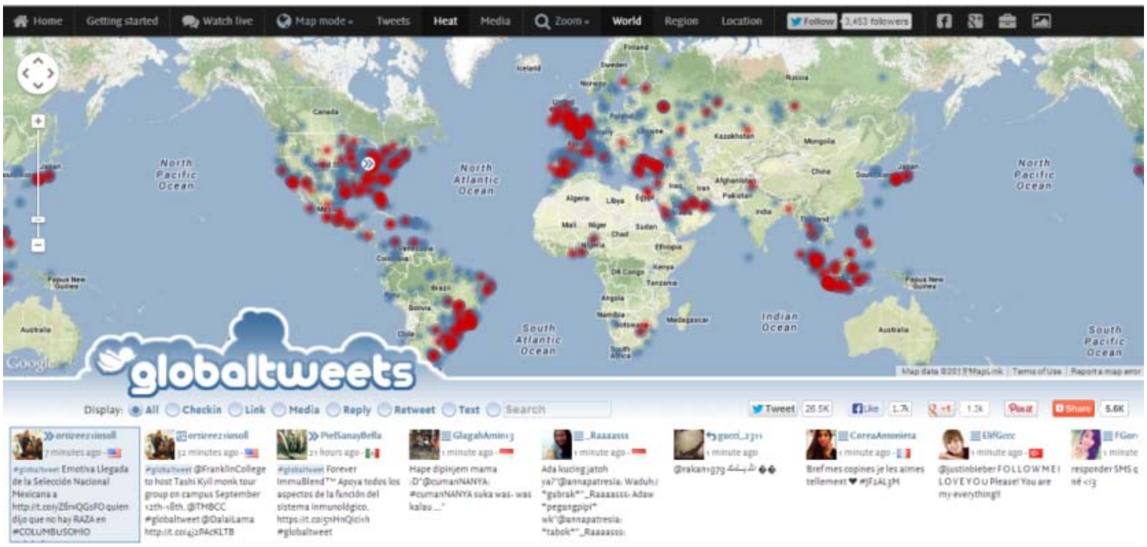


Figure 6: The one million tweet map - cluster visualization (left) and heatmap (right)

[6] http://onemilliontweetmap.com/

A fancy Twitter mapping application is **GlobalTweets**[7]. The principle is the same, on the map the last Tweets are displayed in real time, but this application adds a classification to the Tweets. The user can select between Tweets that show a location status, contain a link to a website, containing a reference to a media like photo or video, Tweets that are replies to other Tweets, Tweets that are retweets, Tweets that contain just text and Tweets filtered by a keyword chosen by user. Below the map there is a timeline that shows the latest Tweets with information like user picture, flag of the country where s/he posts from and Tweet content. The user can click on each timeline element and the map will zoom to the Tweet location and the whole content of the Tweet will be displayed in a bubble. Also the user can choose between several visualization methods: the Tweets can either be seen as markers (Tweets or Media map mode) or as a Heat map (Figure 7).

Another interesting temporal feature is the Live mode, which shows a sequence of latest Tweets in the order that they were posted. This is like a real-time movie of posted Tweets. However, this application is also real time and provides no option to investigate the Tweets in the past on a longer period than the last few hours.



ortizeez1iesoll
7 minutes ago
#globaltweet Emotiva Llegada de la Selección Nacional Mexicana a http://t.co/yZ6nQGsFO quien dijo que no hay RAZA en #COLUMBUSOHIO

ortizeez1iesoll
32 minutes ago
#globaltweet @FranklinCollege to host Tashi Kyil monk tour group on campus September 12th-18th. @TMBCC #globaltweet @DalaiLama http://t.co/4j2PAcKLTB

PielSanayBella
21 hours ago
#globaltweet Forever ImmuBlend™ Apoya todos los aspectos de la función del sistema inmunológico. https://t.co/5nHnQicl1h #globaltweet

GlagahAmin13
1 minute ago
Hape dipinjem mama :D"@cumanNANYA: #cumanNANYA suka was- was kalau ..."

_Raaaasss
1 minute ago
Ada kucing jatoh ya?"@annapatresia: Waduh..! *gabrak*"_Raaaasss: Adaw *pegangpipi* wk"@annapatresia: *tabok*"_Raaaasss:

gucci_z311
1 minute ago
@rakan1979 شـ بـلـلـه ◆◆

CorreaAmoniera
1 minute ago
Bref mes copines je les aimes tellement ♥ #JF1AL3M

ElifGccc
1 minute ago
@justinbieber FOLLOW ME I LOVE YOU Please! You are my everything!!

FGon
1 minute
responder SMS q né <i3

Figure 7: Screenshot of globaltweets.com - visualizes Tweets either as a heat map or as markers. The image presents the heat map mode.

[7] http://globaltweets.com

Tweereal[8] is a real-time Twitter activity map that displays Tweets as fading/appearing colored circles (Figure 8). The Tweets appear on the map as they are created, in real-time. The user can control the size, the opacity of the circles and how fast they disappear from the map. However, no information about the content of the Tweet is displayed.

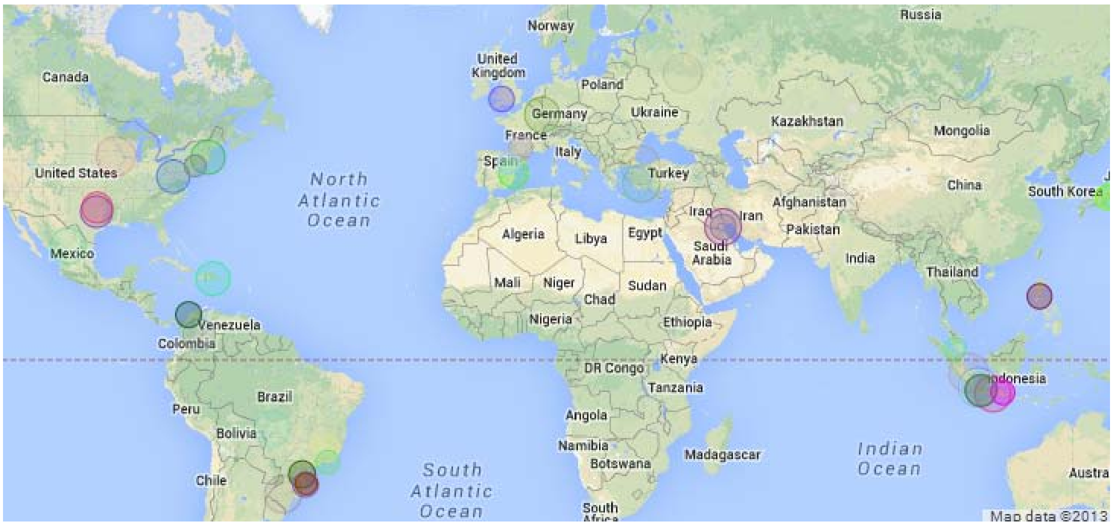


Figure 8: Screenshot of tweereal.com, a web application that maps Tweets as colored circles in real-time

Another Twitter activity map is **A world of Tweets**[9]. The density of Tweets is displayed either as a HeatMap or as a smoky clouds map. Real time Tweets are symbolized with a dot that moves from north to south and stops at the location where the Tweet was created. This is a unique way of visualizing the Tweets. As the application runs for a longer time, more dense areas become distinguishable, being highlighted with red on a heat map (Figure 9) and with black on a smoky cloud map. An interesting aspect of this map is the 3D Stereoview which allows the user to see the map in 3 dimensions with the help of stereoscopic glasses.

[8] http://tweereal.com/
[9] http://aworldoftweets.frogdesign.com/

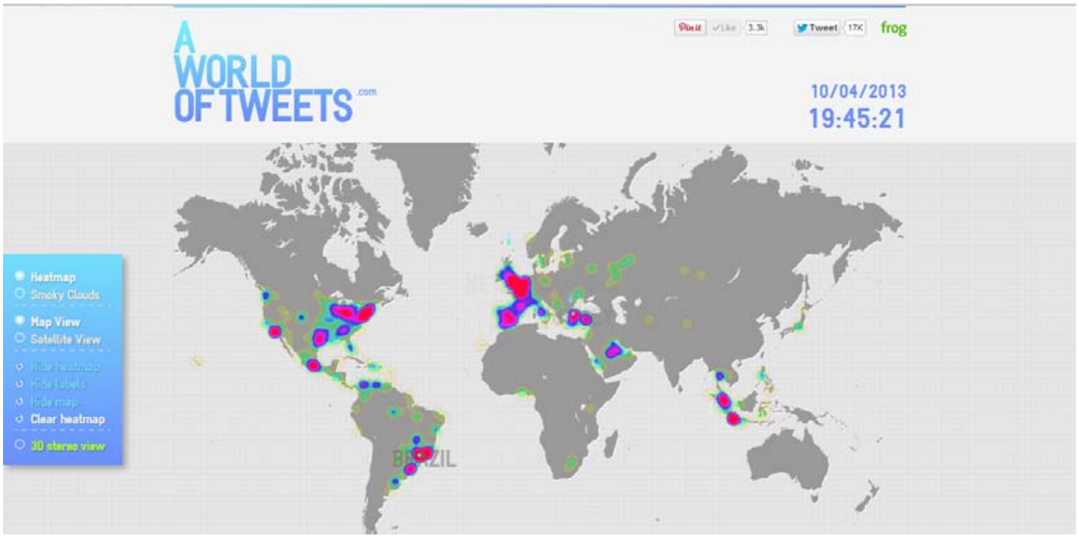
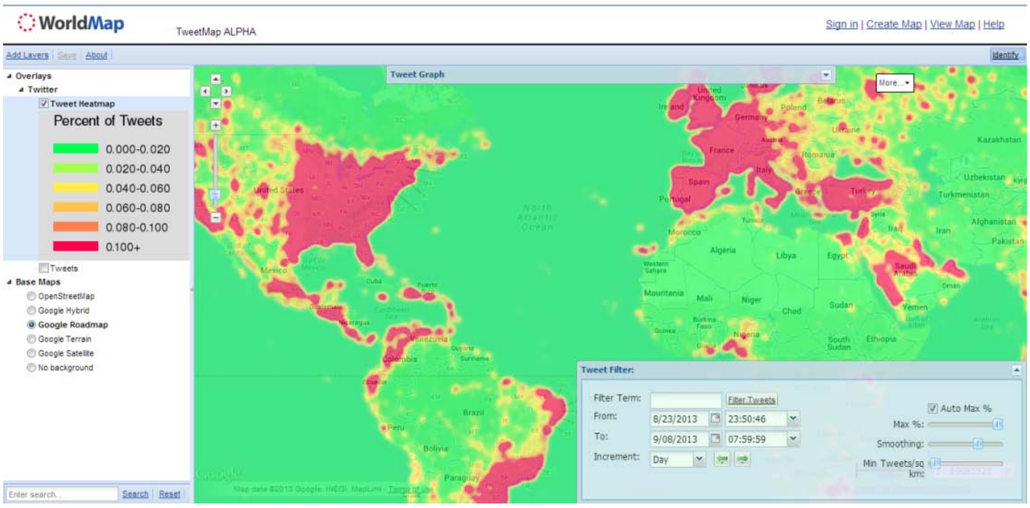


Figure 9: A world of Tweets - shows Tweet density on a world map

A tool that uses a timeline to track for changes is **Tweetmap**[10]. The Tweets can be filtered by a keyword and a period of time of interest. The filtered Tweets are then displayed as a heatmap (Figure 10) or as points and their evolution in time is also shown on a timescale as a percentage of the total number of Tweets. This application is not meant to come with a unique visualization, but to prove the efficiency and speed of MapD, a new database system, in processing large amounts of data (Murphy, 2013).



☑ Tweet Heatmap

Percent of Tweets

🟩	0.000-0.020
🟩	0.020-0.040
🟨	0.040-0.060
🟧	0.060-0.080
🟧	0.080-0.100
🟥	0.100+

□ Tweets

Base Maps

○ OpenStreetMap
○ Google Hybrid
● Google Roadmap
○ Google Terrain
○ Google Satellite
○ No background

Tweet Filter:

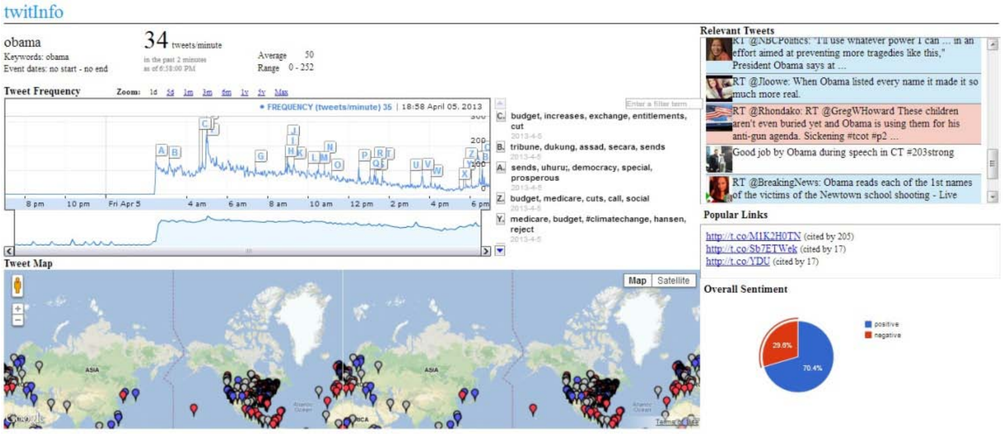
Filter Term: Filter Tweets
From: 8/23/2013 23:50:46
To: 9/08/2013 07:59:59
Increment: Day
☑ Auto Max %
Max %:
Smoothing:
Min Tweets/sq km:

Figure 10: Screenshot of Tweetmap - an application for displaying archived Tweets

[10] http://worldmap.harvard.edu/tweetmap/

Another application that makes use of the temporal perspective is **Twittinfo**[11]. The authors describe it as a system that identifies and labels event peaks, provides a focus and context visualization of long running events, and provides an aggregate view of user sentiment (Marcus et al., 2011).The idea is somehow similar to Tweetmap, the user can filter the Tweets by one or more keywords. Then s/he can seeon a timeline how often these keywords appear and on the map where they appear.

A first panel upper left (Figure 11) shows statistics about the term usage: how many Tweets per minute contain the searched keyword, a timeline of Tweet frequency that can be changed by the user for up to 5 years, peaks of the use of the keyword and a list with the words that the keyword was associated with during those peaks. The lower panel on the left shows the Tweets on a map related to the searched keyword; the Tweets have blue, red or grey pin according to the sentiment displayed towards the key-term. The panel right bellow displays an apple pie chart of the overall feelings associated with the key-term the panel right above shows some relevant Tweets and some popular links that related to the keyword.



obama
Keywords: obama
Event dates: no start - no end

34 tweets/minute
in the past 2 minutes
as of 6:58:00 PM

Average 50
Range 0 - 252

C. budget, increases, exchange, entitlements, cut
2013-4-5

B. tribune, dukung, assad, secara, sends
2013-4-5

A. sends, uhuru;, democracy, special, prosperous
2013-4-5

Z. budget, medicare, cuts, call, social
2013-4-5

Y. medicare, budget, #climatechange, hansen, reject
2013-4-5

Relevant Tweets

RT @NBCPolitics: "I'll use whatever power I can ... in an effort aimed at preventing more tragedies like this," President Obama says at ...

RT @Jloowe: When Obama listed every name it made it so much more real.

RT @Rhondako: RT @GregWHoward These children aren't even buried yet and Obama is using them for his anti-gun agenda. Sickening #tcot #p2 ...

Good job by Obama during speech in CT #203strong

RT @BreakingNews: Obama reads each of the 1st names of the victims of the Newtown school shooting - Live

Popular Links

http://t.co/M1K2H0TN (cited by 205)
http://t.co/Sb7ETWek (cited by 17)
http://t.co/YDU (cited by 17)

Figure 11: Screenshot of TwitInfo - an application for event visualization and sentiment detection

The visualization technique for the mapping part is simple, a Google map is used and on top of it the pins with the relevant Tweets are displayed. The timeline graphic is suggestive, because on hover all the information about the point of the graphic that is hovered over is shown on the upper bar. Also the blue circle that

[11] http://twitinfo.csail.mit.edu/

shows constantly at which point in time the user finds himself. This application uses a timeline to show the evolution of events with different time granularities: the user can choose to see the topic of interest from within the last day up to within the last 5 years. The visualization uses the linking and brushing technique: when the time period is modified, the graphs and the map display changes according to the new time slot.

The **Twitter Stand**[12] application captures Tweets corresponding to breaking news (Samet et al., 2009). It displays on a map the Tweets that contain news information, geolocating them to the position of the news. Tweets that refer to the same news are clustered together. As the Twitter space is "an incredibly noisy medium" (Sankaranarayanan et al., 2011) the retrieval of the news Tweets is made by manually selecting Twitter users whose information can be trustworthy and who are constantly posting news (for example Twitter accounts of news agencies). These users are called the seeders and the posts that they provide are used to define the clusters. Then, the other Tweets are filtered and added to the corresponding cluster based on key terms that they provide.

The visualization method is very simple: the application uses a background map from Google on which the clustered Tweets are overlaid (Figure 12). The classes of the clusters are: general, business, science and technology, entertainment, health and sports and they are shown on the map with different icons according to the topic that they belong to. The application shows recent news, most of them are less than 2 days old. The authors present as an advantage the fact that by pulling out news from Twitter the lag between the time when an event happens and the time when it is posted on Twitter is very little (Sankaranarayanan et al., 2011).

[12] http://twitterstand.umiacs.umd.edu/news/

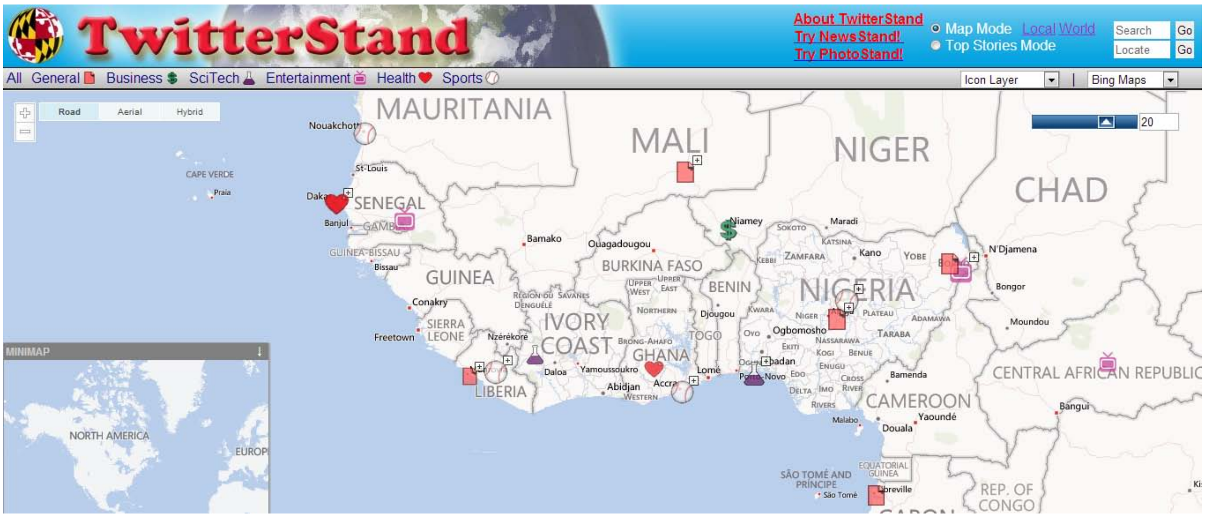


Figure 12: TwitterStand application - Tweets are grouped as clusters on the map according to the news they belong to.

Limosa[13] is an interactive system for visualization of geographic interests of users in Twitter (Vosecky et al., 2013). The application has 4 interfaces.

1. User visualization - a selection of most popular users (from USA) has been made and Limosa can build a profile of the user: where has the user been, what places is s/he tweeting about and which topics is s/he tweeting more often about.

2. Region visualization - a list of selected regions is shown. The user can select the region of interest and then a map is displayed with the density of twitter messages in that region (marked with red on a Google map - the more intense the color, the more people twitter) - Figure 13. At the bottom of the map several lists are displayed: topics in which the region was mentioned, topics of users in that region, top users that mention the region and nearby regions.

3. Topic visualization - this mode provides a list of topics to be visualized along with the locations where the topics have been mentioned more often. The user can choose one topic and then see the map with locations where the topic was visualized, the users who tweeted about that topic and related topics.

4. Term visualization - the user can search for a term and then see the top 20 topics related to the searched term.

[13] http://webproject2.cse.ust.hk:8031/limosa/index.html

The Tweets are visualized as a heat map, colors ranging from blue (less dense) to red (bigger density). The application does not show any time evolution.

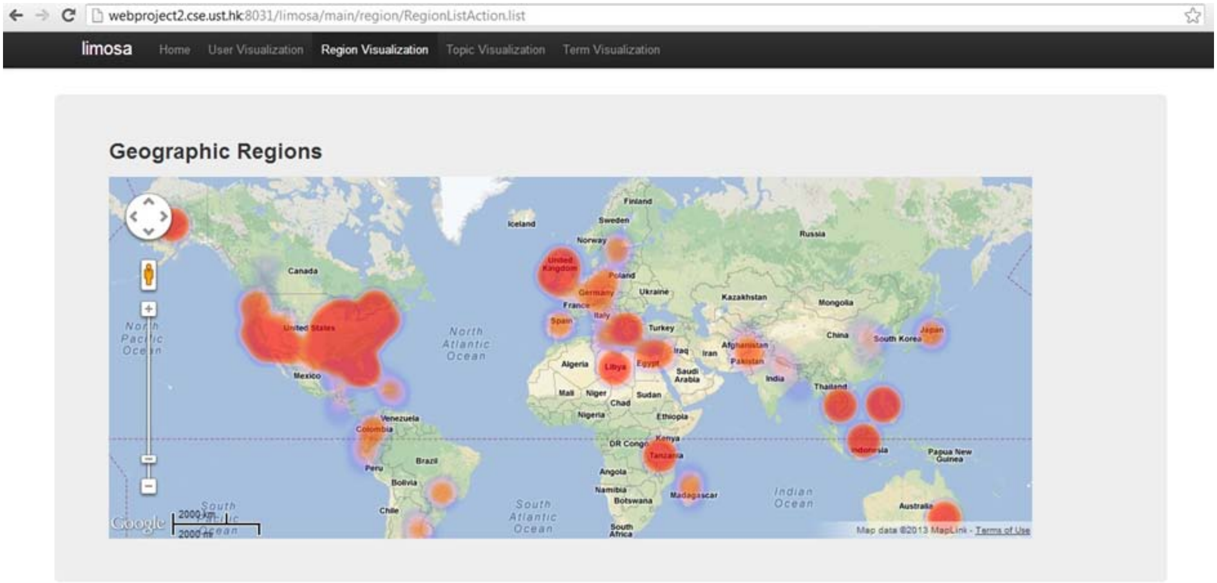


Figure 13: Screenshot of Limosa - an application for visualizing user interests on a map

This overview shows that most of the applications visualize Tweets as markers, as clusters or as heat maps. Most of the applications are real-time and do not have any time component. Therefore in the visualization part of this thesis, the focus will be on developing a web application that places Tweets on the map based on the topics that the Tweets belong to and allows the user to select a time period for which s/he visualized the Tweets.

4 ANALYSIS OF AVAILABLE TOOLS

In order to achieve the goal of classifying the Tweets and visualizing them afterwards several tools were taken into consideration. In this chapter the available tools for classifying the Tweets and visualizing them are presented along with advantages and disadvantages.

4.1 NATURAL LANGUAGE PROCESSING TOOLS FOR TOPIC CLASSIFICATION

Natural language processing (NLP) refers to any kind of computer manipulation of a language used for everyday communication by humans (Bird et al., 2009). Computer manipulation can refer in a simple case to counting word frequencies or in a more complex case to advanced processing of ideas and thoughts up to the point of giving meaningful responses back. From this broad field of NLP, the focus in this thesis will be on topic classification.

Topic classification for Twitter data poses some challenges that make automatic classification of Tweets almost impossible. First of all, the size limit of a Tweet of 140 characters is a major challenge, as most tools for topic classification perform well on longer texts and less well on very short ones. Another problem with the text size is that words tend to be shortened up or abbreviated in order to fit the text size. Then, this makes it harder for an automated classification. Second of all, presence of spam messages or non user generated content can distort the results very rapidly. For example, the left image in Figure 14 shows a word cloud of the most tweeted words in Baden-Württemberg. One could assume that the Tweets in this state are mostly about traffic jam (Stockender), transportation (Verkehr) and weather (Regen, Temperatur).

However, it was noticed that a big amount of Tweets come from non-UGC sources like the local Info Traffic Service and the Weather Service. As these Tweets represented about 45% of all Tweets, this lead to the distorted results. The results after removing the Tweets generated by these 2 users can be seen in the right word cloud in Figure 14.

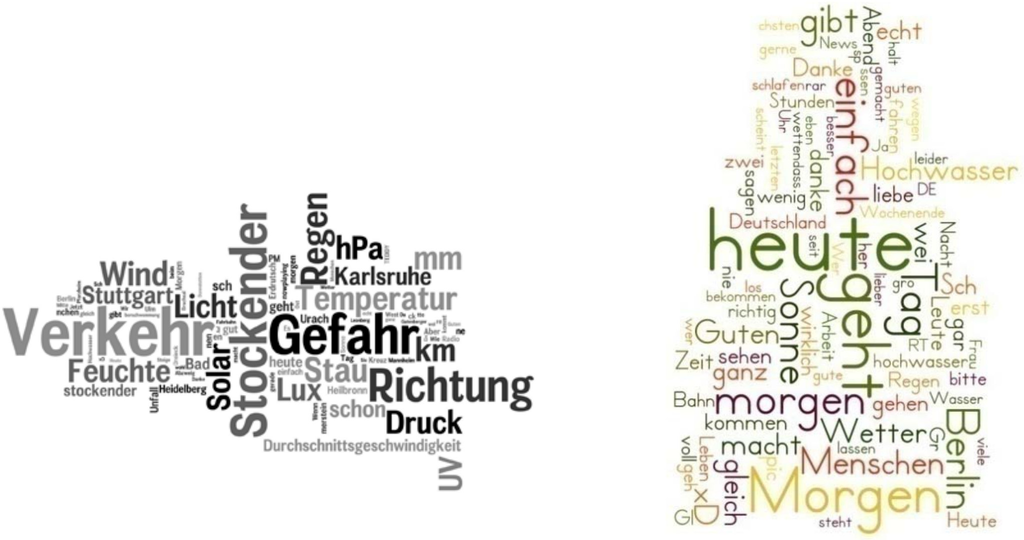


Figure 14: Word frequency using Wordle.com. On the left, before filtering, the results were mostly related to traffic, danger, traffic jam or temperature, rain, humidity. On the right, the word frequency after excluding the noisy Tweets.

The first approach in this part was to investigate if trends in conversations could possibly be identified through word frequency detection. There are many word count tools (even online tools for example WriteWords[14] or Wordcounter[15]), but it is important that the algorithm also includes the possibility of removing the stop words and of customizing the list of stop words. Stop words are the common words that are removed before processing of text. Examples of stop words are "the", "and", "or" etc. A first try was to write a word frequency algorithm in Javascript that would also extract the stop words. The result was that it worked for shorter documents but it was very slow or became irresponsive for more than 200 Tweets. For this reason a Java algorithm written by Fred Swartz[16] was chosen for this task. The algorithm is very efficient and the list of stop words can also be modified.

The second approach was to use a probabilistic topic modeling algorithm to discover possible topics. One of the tools that could be used for topic classification is Weka, developed by the Machine Learning Group at the University of Waikato (Hall et al., 2009). According to the authors, Weka is a tool used for data pre-processing,

[14] http://www.writewords.org.uk/word_count.asp
[15] http://www.wordcounter.com/
[16] http://www.leepoint.net/notes-java/data/collections/maps/ex-wordfreq.html

classification, regression, clustering, association rules and visualization. The datasets for Weka have to be converted to a specific format .arff Attribute-Relation File Format[17].

Another tool used for topic classification is Mallet. *Mallet is a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction and other machine learning applications to text* (McCallum, 2002). According to (Zhao et al., 2011) a topic is a subject discussed in one or more documents and each topic is assumed to be represented by a multinomial distribution of words. This is also what LDA assumes when generating the topics. As Mallet uses the standard implementation of LDA (Underwood, 2012), the topics resulted after running the algorithm are a succession of words with their according probabilities. In the following example, 7 is the rank of the topic and 0.021 is the probability that the topic occurs. The words that belong to the topics are followed by their weights.

```
7   0.021    hochwasser (12) bringen (3) ersatz (2) umwege (1) eisenbahnbrücke (1)
             hässlich (1) immerhin (1) übertreffen (1) neue (1) bietet (1)
             schaulustige (1) meißen (1)
```

With Mallet some parameters can be modified, for example the number of retrieved topics and the number of iterations for the generation of the words in the topics (also called Gibbs Sampling[18]). The number of topics can be chosen depending on the level of detail that is needed and the amount of data. For more detail, more topics are better. The number of iterations of the Gibbs sampling process is recommended to be between 1000 and 2000, as more iterations provide better quality results (McCallum, 2002).

An alternative to Mallet is JGibbLDA, a Java implementation of LDA created by Xuan-Hieu Phan and Cam-Tu Nguyen[19] in 2008. The input data format can be text strings and should be preprocessed beforehand (removing stop words, stemming,

[17] http://www.cs.waikato.ac.nz/ml/weka/arff.html
[18] http://en.wikipedia.org/wiki/Gibbs_sampling
[19] http://jgibblda.sourceforge.net/

etc.). Other topic modeling tools include MatLab Topic Modeling Toolbox[20] and R Topicmodels package[21].

The reason why for this project Mallet was chosen is that this implementation of LDA is quite easy to use, can be run from Eclipse[22] just by importing the project with no further code writing. The tool can also be run from command line. The data can be in .txt format which was very easy to obtain with no further conversion of data. Also Mallet provides the option of removing stop words without previous preprocessing. The documentation for Mallet was also better than for the other tools.

4.2 VISUALIZATION TOOLS

Visualization plays an important role in data analysis. Depending on the visualization method, it can facilitate the interpretation of a big amount of data but it can also present distorted information when visualized in a bad way. So it is important to use the right tools in order to have a correct interpretation of the data. Deciding between the different libraries for data visualization is not easy and it needs a thorough analysis. Different types of visualization libraries are presented separately for a better overview. These are just the types of visualization that were needed in the context of this thesis, but there are many more available. The mapping library would be needed to plot the Tweets on a map, graph visualization tools are needed to show the quantitative side of the Tweets and a timeline is needed to give the user the possibility to visualize Tweets at different moments in time.

4.2.1 Mapping libraries

Nowadays there are many options to build web-maps without the need to program. ArcGISOnline or Geocommons are 2 examples of web mapping platforms where data can be uploaded and the user has the possibility of adding interactivity to the map or changing the design of the map just with some clicks or by dragging and dropping. But for a customization of the applications it is always better to create

[20] http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm

[21] http://cran.r-project.org/web/packages/topicmodels/index.html

[22] http://www.eclipse.org/

mapping applications from scratch using existing libraries. The variety in web-mapping libraries is quite big, as many new ones have been developed for building more attractive web maps.

An important distinction can be made between proprietary and open mapping libraries. From the proprietary branch, **Google Maps API**[23] is the most popular, being used by more than 350000 websites (Wm, 2012). For low traffic use, their services are free, but from more than 25000 map views/day a fee of 0.50$ is charged for each 1000 excess map load[24]. Other proprietary mapping APIs are **Bing Maps API**[25] and **HERE Maps API**[26] developed by Nokia which is at the moment used also by popular mapping services like Yahoo. The good part about proprietary APIs is that usually they provide base maps and advanced features like routing, real-time traffic, geocoding, 3D Buildings or 'bird's eye' view (Wm, 2012).

Open source APIs have the advantage that they can be customized and they allow for mixing different components to create a map. For example, OpenLayers API can be used to fetch map services from Google Maps, OSM or from CloudMade or combine them all together. One can also create his own maps and use his own mapping services.

A pioneering open source web mapping library is **OpenLayers**[27]. In contrast to Google Maps API, OpenLayers is totally free and it is developed by the open source community. Recently, in order to add a friendly user interface to OpenLayers, **GeoExt**[28] was developed. GeoExt combines OpenLayers with ExtJS, another JS library for building rich web mapping applications. Geoext provides customized widgets to view, edit and style geospatial data (GeoExt, n.d.). GeoExt runs entirely on the client-side and therefore no plug-in is needed.

[23] https://developers.google.com/maps/
[24] https://developers.google.com/maps/faq?hl=ro&csw=1#usage_pricing
[25] http://msdn.microsoft.com/en-us/library/gg427610.aspx
[26] http://developer.here.com/javascript-apis
[27] http://openlayers.org/
[28] http://www.geoext.org/

The initial idea of the application was to create a choropleth map, so **GeoChart**[29] Visualization from Google was also taken into consideration. Geochart is a map with 2 modes: region mode, which creates a choropleth map based on the values of each region and a marker mode where each marker is represented by bubbles scaled according to the marker's value. The rendering is based on SVG or VML[30].

Another Javascript mapping framework that creates SVG maps without using any mapping services is **Kartograph**[31]. According to its developer (Aisch, 2012), the big advantage of this framework over other mapping libraries is that the map projection of the displayed data can be chosen. Another advantage is that the data needed to be used within the library can be in .shp, .kml or .geojson formats. With this library, maps can be built in a very artistic way as for example La Bella Italia in Figure 15.

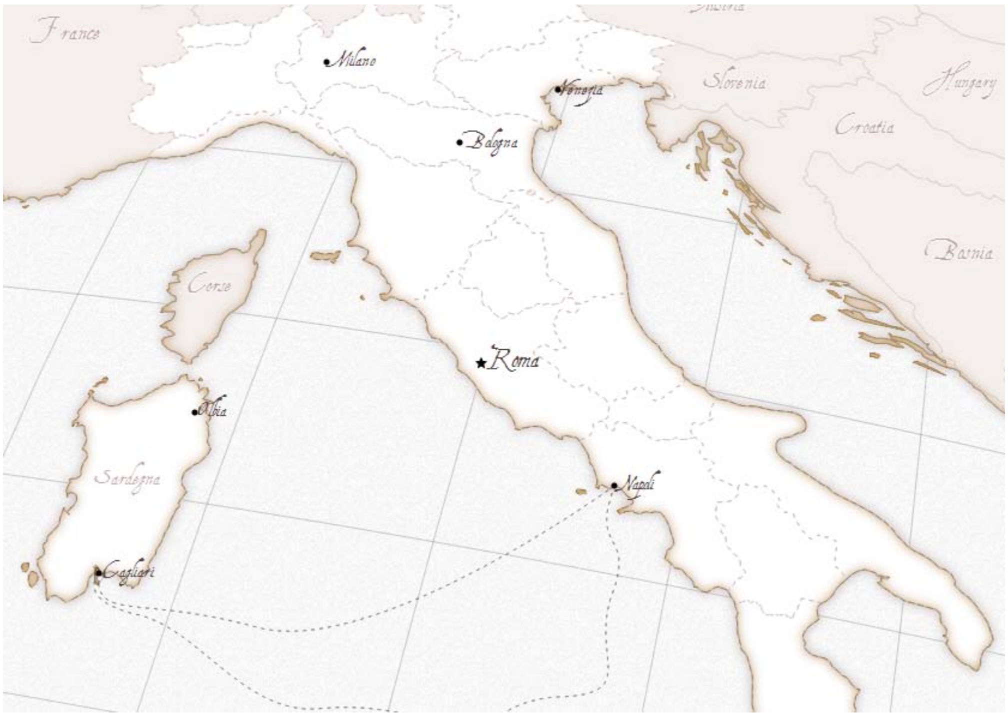


Figure 15: La Bella Italia, map produced with Kartograph library (Source: http://kartograph.org/showcase/italia/)

[29] https://developers.google.com/chart/interactive/docs/gallery/geochart#Overview
[30] https://developers.google.com/chart/interactive/docs/gallery/geochart#Data_Format
[31] http://kartograph.org/

Polymaps[32] is a lightweight mapping framework that uses SVG for vector data rendering and therefore the maps look very appealing. It can also use image-based web maps from OpenStreetMap, CloudMade or Bing. One mention is that Polymaps has not had any new releases since 2011.

Leaflet[33] is another open source mapping API. It is a lightweight Javascript library that can be combined with user customized map tiles created with CloudMade or MapBox to produce very appealing map applications. The features that this library offers are quite limited compared to OpenLayers but they can be extended with further plug-ins for example like setting labels or creating heatmaps or advanced visualizations.

For the final application Leaflet was used for the following reasons: a mapping library with a marker-clustering plug-in was needed and the map had to look appealing. The Tweets were clustered depending on topic and distance and for that Leaflet Marker Cluster plug-in was used. Leaflet also supports data in .geojson format. Another important aspect is that Leaflet is rich in tutorials and examples.

For the base map Cloudmade mapping services were used because their styling editor is very flexible when designing how the tiles should look like. Therefore the base map can be customized up to a big extent using Cloudmade.

4.2.2 Chart visualization tools

Under chart visualization tools are meant pie, scatter, bar or bubble charts, tree maps, stack bars etc. These are mainly used to display some quantitative information. The options are numerous, there are many libraries which can be used for this kind of representation. Most of them are written in Javascript and use SVG to display the graphical elements on the canvas.

Google Charts[34] is a Javascript library that provides ready-to-use charts easy to embed in web pages. The data can be loaded in the charts as .json format. The charts can also be customized by adding event listeners to different chart

[32] http://polymaps.org/
[33] http://leafletjs.com/
[34] https://developers.google.com/chart/interactive/docs/reference?hl=en

components, by changing the axes or by adding tooltips. However creating a new chart from scratch is quite difficult to implement and time consuming.

Another library for visualization is **Highcharts** also written in HTML5/Javascript. The advantage of Highcharts is that the charts are highly interactive and have a very appealing design. The library also supports a big variety of chart types.

gRaphaël[35] is a project of SenchaLabs and it provides charts like pie, bar or line graphs. Infovis is another toolkit that allows to create charts that vary from the simple bar/pie and area charts to highly interactive and innovative charts like sun burst, treemaps or force directed graphs.

D3 (Data-driven documents)[36] is a visualization library that was released in 2011 by PhD. student Mike Bostock. The library is the successor of Protovis, a library that generates SVG graphics from data. D3 has the advantage of allowing great control over the final result with relatively easy to understand code (Viau, 2012). D3 is not just for charts, but it allows a great flexibility in building many types of visualization. D3 can be used in building timelines but also as a mapping library as it can interpret .geojson files.

For the final visualization of the graph charts D3 was used mainly because of the freedom it gives in designing graph charts, but also because of the numerous tutorials that can be found online.

4.2.3 Timeline visualization tools

Envision.js[37] is a library that can be used for creating fast, interactive HTML5 timeline visualizations. The data can be visualized also in real-time so that the graphic is updated whenever new data comes in. An advantage about Envision is that it allows the user to select a specific time period that s/he wants to examine more closely (Figure 16).

[35] http://g.raphaeljs.com/

[36] http://d3js.org/

[37] http://www.humblesoftware.com/envision

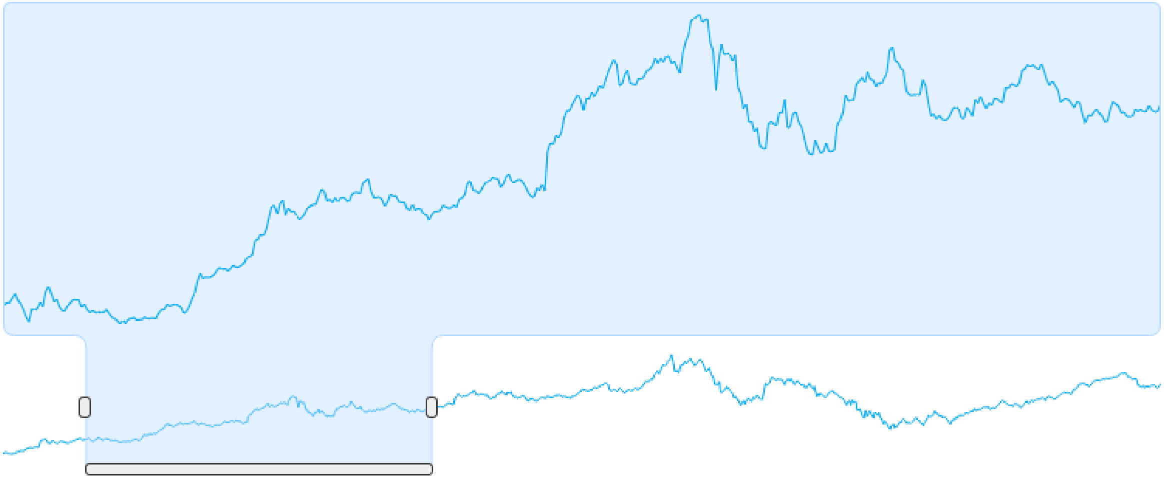


Figure 16: Envision Time Series - allows the user to zoom into a period of interest (Source: http://www.humblesoftware.com/envision/demos/timeseries)

Cubism.js[38] is a plug-in based on D3 for visualizing time series. With this tool multiple time series can be visualized which is interesting for visualization of topic evolution in time for example. However, for more than 5 topics the graph might become hard to read (Figure 17). Data in .json format can only be used if the plug-in is extended to support such functionality. Moreover, the user cannot interact with the timeline to choose a specific time period.

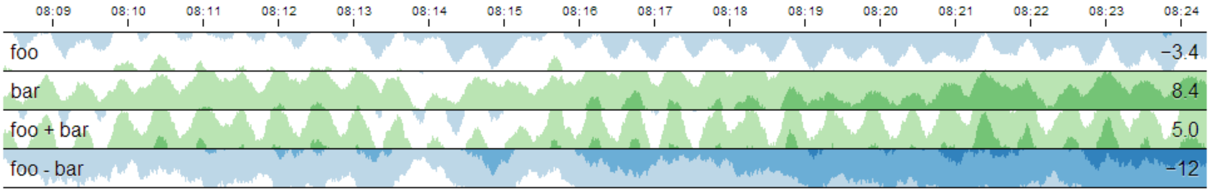


Figure 17: Time series visualized with Cubism.js. No user interaction to select a time period is possible.

TimelineJS[39] is an open-source tool that can be used to create visually rich-interactive timelines. It is oriented more toward event visualization rather than statistical data visualization. This could be useful in case the timeline would have been populated with events that took place in the period of time of the Tweets. Then the topics of discussion could be compared with the events that were happening at the moment.

[38] http://square.github.io/cubism/
[39] http://timeline.verite.co/

In the end for the timeline no special timeline library was used. Instead, a D3 bar chart was built with a brush on top of it that allows for selection of different periods of time. The data was filtered using a library called **Crossfilter**[40] which allows for very fast filtering of data according to any dimension in the dataset (not only time). Crossfilter works with .json data and provides extremely fast interaction between components that visualize the same data filtered by different dimensions. These visualizations that are linked through some interaction are called coordinated views (McCrocklin, n.d.).

[40] http://square.github.io/crossfilter/

5 IMPLEMENTATION

5.1 CLASSIFICATION OF TWEETS

Extracting topics from Twitter data is a very difficult task and it has proved to be the most challenging one during this project. The approach used to deal with this issue was to identify trends in the discussion using word frequency algorithm and to identify topics using a probability model based on co-occurrence of words. However, after the first tests of topic extraction with Mallet it became clear that the results could not be visualized in the form given by this software. Some of the words associated together in topics had no connection to each other and displaying the topics in that form would have made no sense for the final user. With this in mind, it was preferred that the word frequency and Mallet topic classification were used for a first analysis and the creation of general categories. This process will be described in chapter 5.1.1.. The final classification of Tweets, which actually means assigning a general topic to each Tweet, is described in chapter 5.1.2.

5.1.1 Discovering topics using word frequency algorithm and probabilistic topic modeling tool Mallet

In this subchapter it is described how the 11 general categories were created in the end. In the following subchapter it is described how the categories were assigned to each Tweet.

The chosen data was for Saxony, during 31st of May - 6th of June 2013. This selection was made because of the flood period in this area. It started on the 2nd of June and the 6th of June was the day when the level of the Elbe started stagnating and volunteers were not needed so much anymore. A reason to choose this period was to investigate if the flood event is reflected in the Tweets and how useful the posts about the flood would be for disaster relief organizations. As the initial data were at Germany level, I imported the data in ArcGIS and performed a spatial selection to retrieve just the Tweets from Saxony. A second selection was performed at date level. The Tweets were selected for each day and saved in separate files so that analysis could be performed on each one of them. Tweets were also "cleaned", which means that all the Tweets that were not in German language were deleted, just to make the processing easier. Tweets that were not generated by a user were also deleted. To this category belong for example, Wetter Dresden Tweets, with information about the weather in Dresden. In the end the number of Tweets was reduced to 1577.

At this point, a problem related to character encoding was encountered. As all data are encoded in UTF-8 in ArcGIS, special characters could not be displayed. To solve this problem the data were saved in Latin1 (ISO-88591) encoding using QGIS.

The first step in investigating the data was to see what the most frequent words were for each day of the selected period. For this, a Java algorithm[41] written by Fred Swartz was used as mentioned in the previous chapter. A comprehensive list of German stop words was found on Snowball project website[42]. More stop words were added to this list in the beginning, while testing the word frequency algorithm. The results of algorithm were used to create word clouds for an easier visualization of the word frequency. The word clouds were generated using Wordle[43], a great tool for creating and styling word clouds. From Figure 18 one can see how flood related Tweets (marked by the word "hochwasser") are more frequent as the flood level increases. The word clouds were used in the end to discover what are the most discussed topics judging by word frequency.

[41] http://www.leepoint.net/notes-java/data/collections/maps/ex-wordfreq.html
[42] http://snowball.tartarus.org/algorithms/german/stop.txt
[43] http://www.wordle.net/

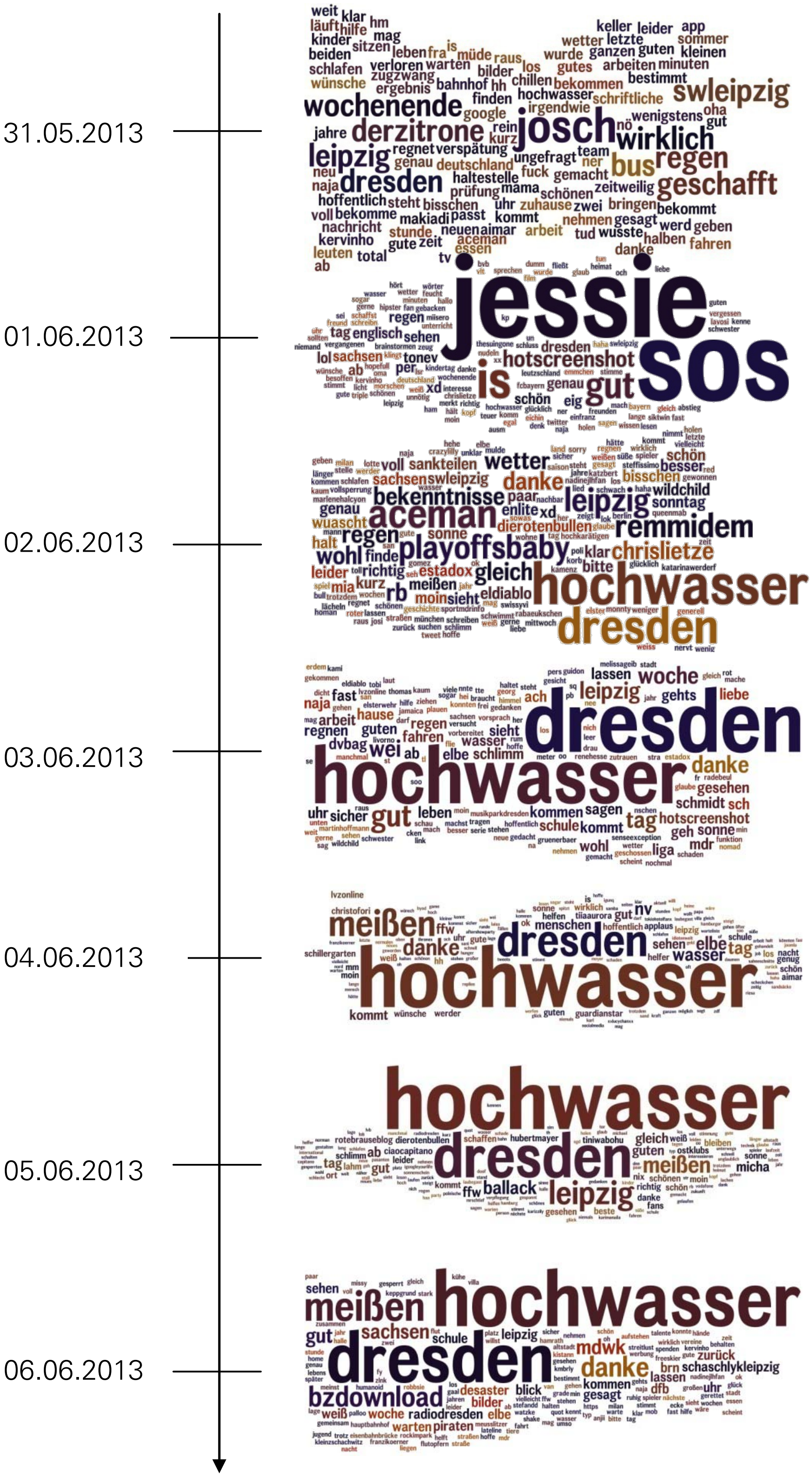


Figure 18: Word clouds generated with http://www.wordle.net/from the most frequent tweeted words during 31.05 - 06.06.2013 in Saxony.

The second step was to use a probabilistic topic modeling tool and for that, Mallet was chosen. The same stop words list was used as for word frequency. Some parameters can be changed in Mallet in order to make the topic modeling more accurate. One of these parameters is the number of topics. The size of the document collection and the overview level determine the number of topics (Chae et al., 2012). For a broad overview 10 topics can be enough but for a deep insight 100 topics might be a good number. In this case 10 topics were used as this was the approximate number of topics that were expected to arise in the Twitter data for 1st day. The number of words per topic can be increased from the default of 5 to 10 or more. In the model that was used to classify the Tweets a document represents the collection of Tweets for a day and each topic is a distribution of 10 words. The number of iterations was 1000 although the algorithm was also tested with 2000 iterations and the difference between the topics was not very significant. Table 2 and Table 3 show the results of Mallet Topic Model algorithm using 1000 iterations, respectively 2000.

0	0.105	josi (2) genau (2) gut (2) lassen (2) stählst (1) aufsetz (1) fake (1) wohn (1) konsequenzen (1) früher (1)
1	0.102	dresden (4) hochwasser (3) liga (2) hoffentlich (2) fahre (2) verdenken (1) klappt (1) heller (1) hunger (1) tagen (1)
2	0.126	gut (4) schiedsrichter (2) spiel (2) dresden (2) hochwasser (2) regen (2) film (1) glühen (1) min (1) teilen (1)
3	0.095	katastrophenhilfe (2) minuten (2) sehen (2) gut (2) hochwasser (2) zerdrücke (1) frau (1) möglichen (1) aaron (1) letzten (1)
4	0.096	dresden (2) jahr (2) hätte (2) amujan (1) pracht (1) vielen (1) spree (1) erstwunsch (1) käme (1) bleibt (1)
5	0.094	gleich (2) appelt (2) burgern (1) lass (1) hochwasserwarnung (1) geo (1) traut (1) richtigen (1) falschen (1) umwege (1)
6	0.080	erstmal (2) anzumachen (1) davor (1) einschlafen (1) geschafft (1) haut (1) werder.de (1) suchen (1) vielleicht (1) kommt (1)
7	0.091	bringen (3) playoffsbaby (2) gewonnen (2) wohl (2) las (1) einstellung (1) argument (1) ausspricht (1) lipsync (1) geistert (1)
8	0.103	dresden (3) wetter (3) recht (2) träume (2) gut (2) schlafen (1) imbrunst (1) meldung (1) dachte (1) milan-trainer (1)
9	0.108	ersatz (2) derzeit (2) fehlt (2) wirklich (2) meißen (2) alan_berlin (2) gut (2) sonne (2) hand (1) narben (1)

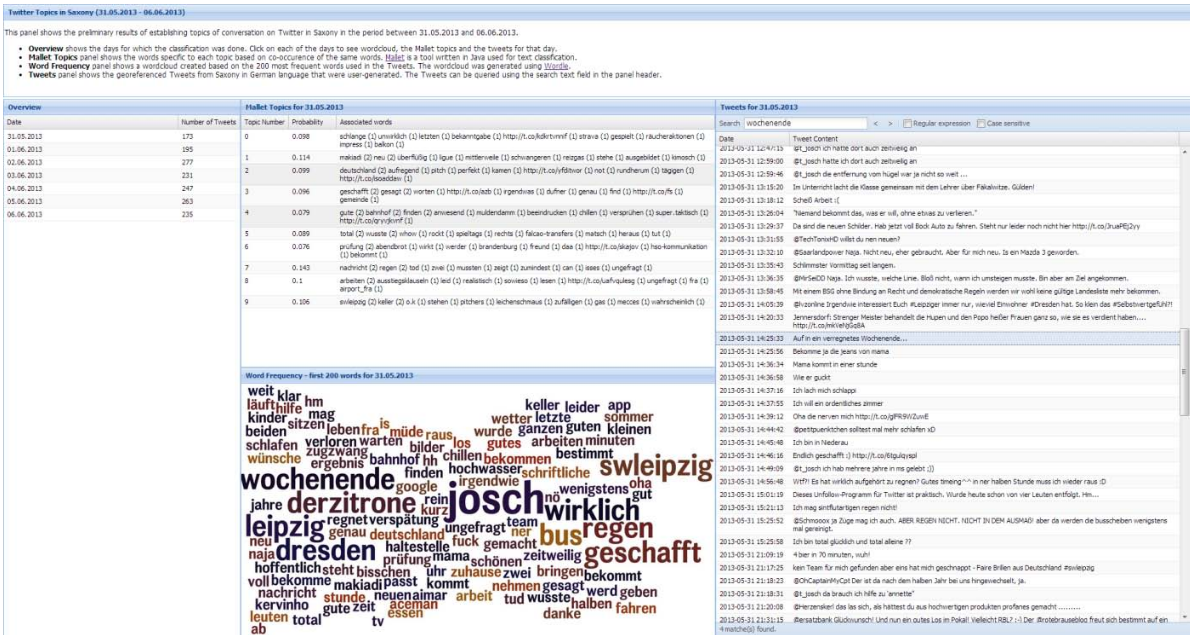
Table 2: Results of topic modeling with Mallet for Tweets of 2. June 2013 in Saxony with 1000 iterations.

0	0.086	dresden (3) direkt (2) hochwasser (2) lass (1) tagen (1) neuen (1) stark (1) verloren (1) applaus (1) neapel (1)
1	0.078	sankteilen (2) jahr (2) gut (2) schreiben (2) geschafft (1) reuter (1) bekomms (1) dachte (1) bild (1) zeitpunkt (1)
2	0.097	sachsen (3) langsam (2) leipzig (2) hätte (2) gut (2) hochwasser (2) glühen (1) hochwasserwarnung (1) bestimmen (1) las (1)
3	0.142	wetter (3) gut (3) hochwasser (3) hast (2) fertig (2) dresden (2) liebe (2) sonne (2) davor (1) hand (1)
4	0.089	playoffsbaby (2) dresden (2) steht (2) paar (2) stählst (1) verhält (1) frau (1) aufgelöst (1) manko (1) weiteres (1)
5	0.100	korb (2) richtig (2) gleich (2) besser (2) einschlafen (1) vielen (1) werder.de (1) spricht (1) empfiehlt (1) paul (1)
6	0.087	dresden (2) verdenken (1) imbrunst (1) heller (1) aufsetz (1) geo (1) auflösung (1) argument (1) käme (1) aaron (1)
7	0.087	swleipzig (2) danke (2) bringen (2) gewonnen (2) amujan (1) zerdrücke (1) mandau (1) fake (1) allegri (1) konstanz (1)
8	0.116	hochwasser (3) derzeit (2) geben (2) zeigt (2) dresden (2) sehen (2) danke (2) monnty (2) gut (2) regen (2)
9	0.119	raus (3) sankteilen (2) mach (2) wasser (2) gut (2) leider (2) schlafen (1) klappt (1) anzumachen (1) min (1)

Table 3: Results of topic modeling with Mallet for Tweets of 2. June 2013 in Saxony with 2000 iterations.

As Table 2 and Table 3 show, the results produced by Mallet are not very intuitive and quite fuzzy. I wanted the topics to be more clear, so in the end I decided to create more broad categories, having as starting points the results from word frequency, from Mallet and the classification framework developed by Stephan Dann, containing 5 high level categories and 28 single subcategories (Dann, 2011).

Topics that were easily detected by using both word frequency and Mallet were natural disasters (later entitled floods as this was the only natural disaster that occurred during this period) and weather. Next, specific words were assigned to each category, to be able to identify Tweets that belong to it afterwards. For example, floods is associated with the words: "hochwasser", "flut", "katastrophe", "ehrenamt", "evakuierung", "pegel", "überschwemmung" or "sperrung". Weather was identified with words such as "sonne", "regen", "feucht", "wetter". Categories were built like this in order to be able to assign a category to each Tweet based on the words it contains and using SQL commands. For building the categories and their associated words, a web tool was created to help having a better overview of the results from word frequency and Mallet. The results from Mallet and the Tweets were converted to XML format. The tool also allows for search through Tweets to see in which context a word was used (Figure 19).



Twitter Topics in Saxony (31.05.2013 - 06.06.2013)

This panel shows the preliminary results of establishing topics of conversation on Twitter in Saxony in the period between 31.05.2013 and 06.06.2013.

- **Overview** shows the days for which the classification was done. Click on each of the days to see wordcloud, the Mallet topics and the tweets for that day.
- **Mallet Topics** panel shows the words specific to each topic based on co-occurence of the same words. Mallet is a tool written in Java used for text classification.
- **Word Frequency** panel shows a wordcloud created based on the 200 most frequent words used in the Tweets. The wordcloud was generated using Wordle.
- **Tweets** panel shows the georeferenced Tweets from Saxony in German language that were user-generated. The Tweets can be queried using the search text field in the panel header.

Overview

Date	Number of Tweets
31.05.2013	173
01.06.2013	195
02.06.2013	277
03.06.2013	231
04.06.2013	247
05.06.2013	263
06.06.2013	235

Mallet Topics for 31.05.2013

Topic Number	Probability	Associated words
0	0.098	schlange (1) unwirklich (1) letzten (1) bekanntgabe (1) http://t.co/kdkrtvnnif (1) strava (1) gespielt (1) räucheraktionen (1) impress (1) balkon (1)
1	0.114	makiadi (2) neu (2) überflüßig (1) ligue (1) mittlerweile (1) schwangeren (1) reizgas (1) stehe (1) ausgebildet (1) kimosch (1)
2	0.099	deutschland (2) aufregend (1) pitch (1) perfekt (1) kamen (1) http://t.co/yfditwor (1) not (1) rundherum (1) tägigen (1) http://t.co/isoaddaw (1)
3	0.096	geschafft (2) gesagt (2) worten (1) http://t.co/azb (1) irgendwas (1) dufner (1) genau (1) find (1) http://t.co/fs (1) gemeinde (1)
4	0.079	gute (2) bahnhof (2) finden (2) anwesend (1) muldendamm (1) beeindrucken (1) chillen (1) versprühen (1) super.taktisch (1) http://t.co/qryvjkvnf (1)
5	0.089	total (2) wusste (2) whow (1) rockt (1) spieltags (1) rechts (1) falcao-transfers (1) matsch (1) heraus (1) tut (1)
6	0.076	prüfung (2) abendbrot (1) wirkt (1) werder (1) brandenburg (1) freund (1) daa (1) http://t.co/skajov (1) hso-kommunikation (1) bekommt (1)
7	0.143	nachricht (2) regen (2) tod (1) zwei (1) mussten (1) zeigt (1) zumindest (1) can (1) isses (1) ungefragt (1)
8	0.1	arbeiten (2) ausstiegsklauseln (1) leid (1) realistisch (1) sowieso (1) lesen (1) http://t.co/uafvqulesg (1) ungefragt (1) fra (1) airport_fra (1)
9	0.106	swleipzig (2) keller (2) o.k (1) stehen (1) pitchers (1) leichenschmaus (1) zufälligen (1) gas (1) mecces (1) wahrscheinlich (1)

Word Frequency - first 200 words for 31.05.2013

weit klar hm
läuft hilfe mag
kinder sitzen
beiden leben fra is müde raus
schlafen verloren warten bilder los
wünsche zugzwang ergebnis bahnhof hh chillen bekommen
keller leider app
wetter letzte sommer
wurde ganzen guten kleinen
gutes arbeiten minuten
bestimmt
wochenende finden hochwasser schriftliche swleipzig
google irgendwie
jahre derzitrone rein kurz josch nö wenigstens oha gut wirklich
leipzig regnet verspätung team
genau deutschland ungefragt ner bus regen
neu haltestelle fuck gemacht
naja dresden prüfung mama schönen zeitweilig geschafft
hoffentlich steht bisschen uhr zuhause zwei bringen bekommt
voll bekomme makiadi passt kommt nehmen gesagt werd geben
nachricht stunde neuen aimar arbeit tud wusste halben fahren
kervinho gute zeit aceman essen danke
leuten total tv
ab

Tweets for 31.05.2013

Search wochenende < > □ Regular expression □ Case sensitive

Date	Tweet Content
2013-05-31 12:47:15	@t_josch ich hatte dort auch zeitweilig an
2013-05-31 12:59:00	@t_josch hatte ich dort auch zeitweilig an
2013-05-31 12:59:46	@t_josch die entfernung vom hügel war ja nicht so weit ...
2013-05-31 13:15:20	Im Unterricht lacht die Klasse gemeinsam mit dem Lehrer über Fäkalwitze. Gülden!
2013-05-31 13:18:12	Scheiß Arbeit :(
2013-05-31 13:26:04	"Niemand bekommt das, was er will, ohne etwas zu verlieren."
2013-05-31 13:29:37	Da sind die neuen Schilder. Hab jetzt voll Bock Auto zu fahren. Steht nur leider noch nicht hier http://t.co/JruaPEj2yy
2013-05-31 13:31:55	@TechTonixHD willst du nen neuen?
2013-05-31 13:32:10	@Saarlandpower Naja. Nicht neu, eher gebraucht. Aber für mich neu. Is ein Mazda 3 geworden.
2013-05-31 13:35:43	Schlimmster Vormittag seit langem.
2013-05-31 13:36:35	@MrSeiDD Naja. Ich wusste, welche Linie. Bloß nicht, wann ich umsteigen musste. Bin aber am Ziel angekommen.
2013-05-31 13:58:45	Mit einem BSG ohne Bindung an Recht und demokratische Regeln werden wir wohl keine gültige Landesliste mehr bekommen.
2013-05-31 14:05:39	@lvzonline Irgendwie interessiert Euch #Leipziger immer nur, wieviel Einwohner #Dresden hat. So klein das #Selbstwertgefühl?!
2013-05-31 14:20:33	Jennersdorf: Strenger Meister behandelt die Hupen und den Popo heißer Frauen ganz so, wie sie es verdient haben.... http://t.co/mkVeNjGq8A
2013-05-31 14:25:33	Auf in ein verregnetes Wochenende...
2013-05-31 14:25:56	Bekomme ja die jeans von mama
2013-05-31 14:36:34	Mama kommt in einer stunde
2013-05-31 14:36:58	Wie er guckt
2013-05-31 14:37:16	Ich lach mich schlappi
2013-05-31 14:37:55	Ich will ein ordentliches zimmer
2013-05-31 14:39:12	Oha die nerven mich http://t.co/glFR9WZuwE
2013-05-31 14:44:42	@petitpuenktchen solltest mal mehr schlafen xD
2013-05-31 14:45:48	Ich bin in Niederau
2013-05-31 14:46:16	Endlich geschafft :) http://t.co/6tgulqyspl
2013-05-31 14:49:09	@t_josch ich hab mehrere jahre in ms gelebt ;))
2013-05-31 14:56:48	Wtf?! Es hat wirklich aufgehört zu regnen? Gutes timeing^^ in ner halben Stunde muss ich wieder raus :D
2013-05-31 15:01:19	Dieses Unfollow-Programm für Twitter ist praktisch. Wurde heute schon von vier Leuten entfolgt. Hm...
2013-05-31 15:21:13	Ich mag sintflutartigen regen nicht!
2013-05-31 15:25:52	@Schmooox ja Züge mag ich auch. ABER REGEN NICHT. NICHT IN DEM AUSMAß! aber da werden die busscheiben wenigstens mal gereinigt.
2013-05-31 15:25:58	Ich bin total glücklich und total alleine ??
2013-05-31 21:09:19	4 bier in 70 minuten, wuh!
2013-05-31 21:17:25	kein Team für mich gefunden aber eins hat mich geschnappt - Faire Brillen aus Deutschland #swleipzig
2013-05-31 21:18:23	@OhCaptainMyCpt Der ist da nach dem halben Jahr bei uns hingewechselt, ja.
2013-05-31 21:18:31	@t_josch da brauch ich hilfe zu 'annette'
2013-05-31 21:20:08	@Herzenskerl das las sich, als hättest du aus hochwertigen produkten profanes gemacht ....
2013-05-31 21:31:15	@ersatzbank Glückwunsch! Und nun ein gutes Los im Pokal! Vielleicht RBL? :-) Der @rotebrauseblog freut sich bestimmt auf ein

4 matche(s) found.

Figure 19: A self-developed Web tool displaying Mallet Topic results (center top), word frequency (center bottom), Tweets with a search function (right) for each day of the analyzed period

A first sketch of the categories and their associated words was thus built, but the list was also filled during the manual classification. As the classification was performed separately for each day, words discovered to belong to a topic for a first day were added to that topic and used in the classification for the second day.

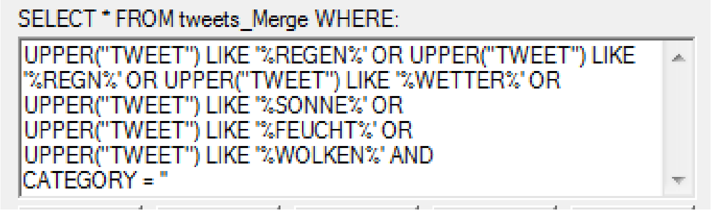
5.1.2 Semi-Automatic/Manual classification

An entirely automatic classification of Tweets was not possible within this project. Building the categories and associating words that would allow for a semi-automatic classification of Tweets using SQL commands to assign the topics only worked for about 50% of the Tweets. The rest of them had to be manually classified. Manual classification brought its benefits in the way that the sequence of words associated to each category grew bigger as new words were associated to the corresponding category. The categories resulted in the end along with their associated words, hashtags and symbols are represented in Table 4.

Category	Associated words, hashtags or symbols
Floods	hochwasser, sperrung, gesperrt, sandsack, pegel, katastrophe, ehrenamt, helfer, einsatz, flut, opfer
Weather	regen, regn, wetter, sonne, feucht, wolken
Sport	Ballack, destro, Insigne, sport, fußball, DFB, #u21em, #dierotebullen, fc, lotte, #playoffsbaby, meisterschaft, pokal, club, liga, korb, spiel
Events	Rock im park, mm13, mwfestival, swleipzig, forum
Education	Prüfung, lehrer, schule
Transport	Bus, Straßenbahn
Technology	Version, testen, ios
Personal Conversations	"@" sign in the beginning of the Tweet

Table 4: Categories along with associated words, hashtags or symbols

The Tweets were queried for the associated words exactly in the order presented in the table and no more than one topic was assigned to each Tweet. For example, the condition for a Tweet to belong to the weather category would be similar to the one in Figure 20. So if a Tweet were associated to a category, it would remain in that category even if it contains words from other categories that follow after its category. This is just a simplification because many times a Tweet could belong to several categories. The assignment of several topics to one Tweet could be achieved in further research in this field. For now, this would complicate both the classification and the visualization tasks.



```
SELECT * FROM tweets_Merge WHERE:
UPPER("TWEET") LIKE "%REGEN%' OR UPPER("TWEET") LIKE
"%REGN%' OR UPPER("TWEET") LIKE "%WETTER%' OR
UPPER("TWEET") LIKE "%SONNE%' OR
UPPER("TWEET") LIKE "%FEUCHT%' OR
UPPER("TWEET") LIKE "%WOLKEN%' AND
CATEGORY = "
```

Figure 20: SQL query to identify Tweets that belong to the weather category

During the manual classification other classes were created, that could not be identified previously. One of them is "status", which represents all the Tweets about the current state of the user, where s/he is, her/his feelings or quotations. Another class that was hard to identify is entitled opinions, but it is actually reduced just to opinions about products. By products we understand any sort of commercial material, ranging from movies, tv-shows, cars, electronic devices or any other commercial devices. Some Tweets could not be assigned to any category but that was less than 10% of the total amount of Tweets. The results of the category classification can be visualized in Figure 21.

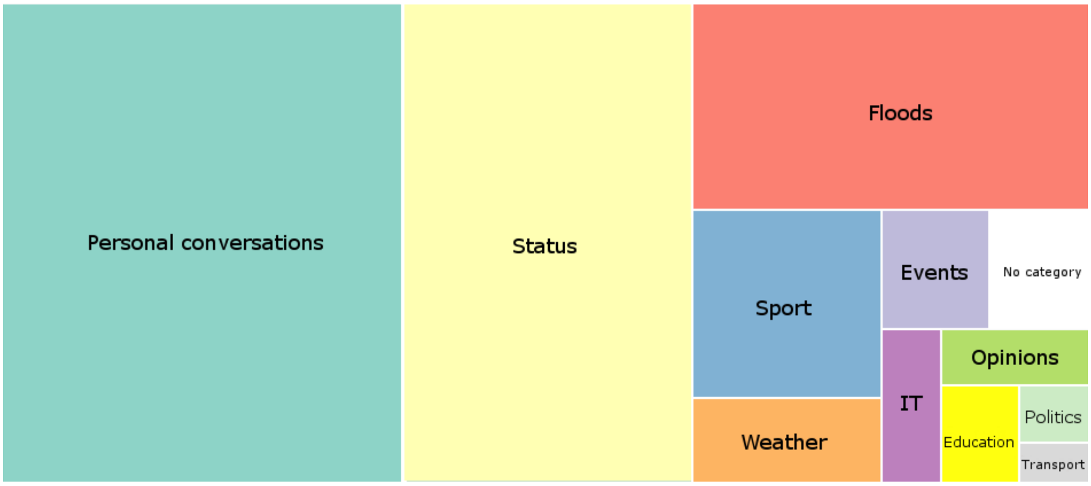


Figure 21: Treemap of the results of categorization of Tweets in the period of 31.05.-06.06.2013 in Saxony

5.2 CREATION OF A WEB APPLICATION FOR TWITTER TOPIC VISUALIZATION

In the beginning of this work it was not exactly defined which results should be visualized and how. Having as starting point the fact that the Tweets should be pre-processed to show the trends in conversations, the first idea was to extract and visualize the most used words and topics in each state of Germany on a monthly basis. The spatial component would be a choropleth map that would show how many Tweets are in each state. When the user hovers over a state, the most used words are displayed as a tag cloud and the topics are shown in a treemap visualization. A prototype (with fake data) can be seen in Figure 22.

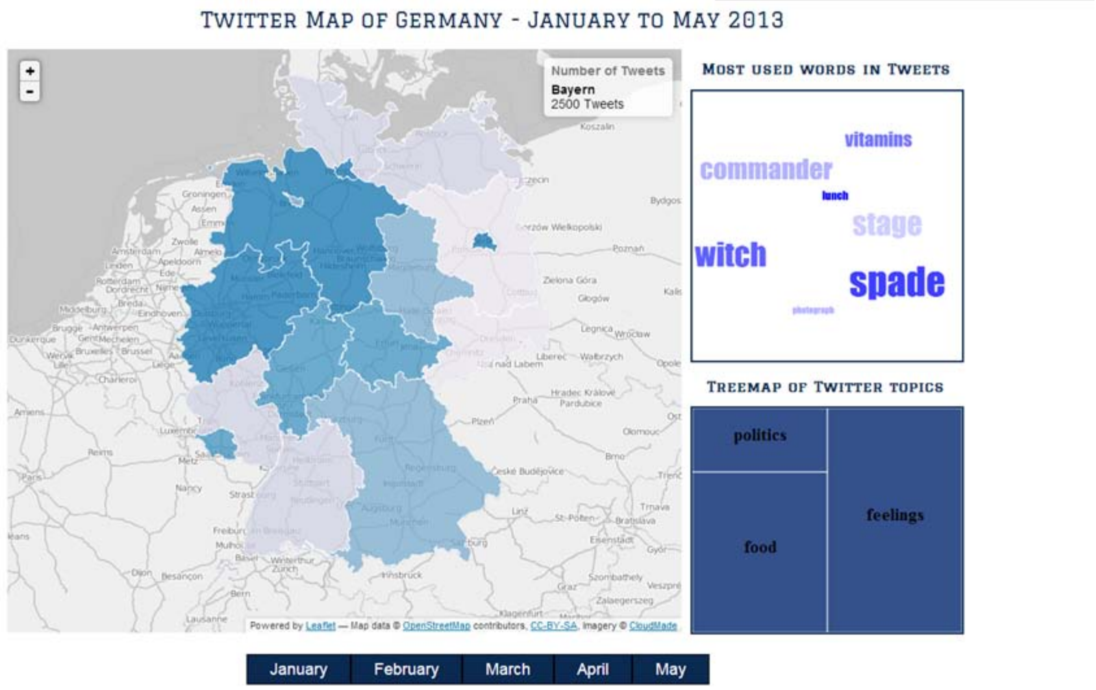

TWITTER MAP OF GERMANY - JANUARY TO MAY 2013
Number of Tweets
Bayern
2500 Tweets
MOST USED WORDS IN TWEETS
vitamins
commander
lunch
stage
witch
spade
photograph
TREEMAP OF TWITTER TOPICS
politics
feelings
food
January
February
March
April
May
Powered by Leaflet — Map data © OpenStreetMap contributors, CC-BY-SA, Imagery © CloudMade

Figure 22: Twitter Map of Germany - first ideas of a map to show topics discussed on Twitter in Germany

However, this visualization turned out to have some deficiencies. First of all it was observed that the number of Tweets does not change that much for each state, so the map in combination with the timeline would not bring too much added value to the user. Second of all the tag cloud and the treemap would change every time the user would hover over a different state which would not allow for comparisons between the states because one can only see the current state.

Other problems were raised also at the level of the amount of data that should be processed. The timeline at a monthly level makes it easier for the processing of data,but it is not recommended for detection of events or topics of discussion that happen during a shorter amount of time. Regarding the timeline, it was also decided that it is not optimal to constrain the user to a fixed period of time (like in the example above - the user can only choose one specific month for which s/he can visualize the results), but to give her/him the option to select her/his own period of interest.

5.2.1 Topic visualization

With these ideas in mind, the initial concept was changed. The first goal was to visualize the topics not only in graphics outside the map, but also on the map. The Tweets are classified into categories and then grouped together on the map depending on the distance between them and the category that they belong to. The Tweets were converted to .geoJSON format in order to be visualized on the map. Different colors were used to distinguish between the different topics. The colors of the topics were chosen using the ColorBrewer[44]. A qualitative color scheme of 11 colors (Table 5) was chosen to make a good contrast with the grey background map.

Personal conversations	🟩
Status, feelings, current activity	🟨
Events: concerts, conferences	🟪
Floods	🟥
Sport	🟦
Weather conditions	🟧
Opinions about products	🟩
Transportation	⬜
IT & Tech	🟪
Politics	🟩
Education	🟨

Table 5: Twitter topics represented with different colors

The map was implemented using Leaflet library[45]. For the background the map services from Cloudmade were used because they also allow for personalized styling with the Cloudmade Editor[46]. The clustering was done by using the already created MarkerCluster plug-in for Leaflet[47]. Small adjustments were made so that the colors

[44] http://colorbrewer2.org/
[45] http://leafletjs.com/
[46] http://maps.cloudmade.com/editor
[47] https://github.com/Leaflet/Leaflet.markercluster

of the clusters would match the colors of the topics. The size of the cluster circles is directly correlated with the number of Tweets in that cluster. One problem with the clusters is that sometimes Tweets come from the exact same location. So the cluster circles are overlaid and therefore it is impossible to access the lower cluster. This problem could eventually be solved in the future by adding a user control for the layers and their rendering order. The user could then change how the layers overlay. A similar problem occurred when a user tweeted many times from the same location and then the Tweets become inaccessible (Figure 23).

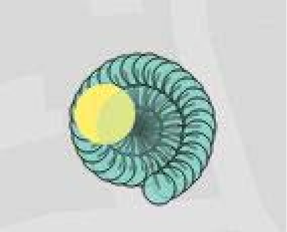


Figure 23: Tweets from the same location cannot be accessed

The MarkerCluster has a great option in this case to show the Tweets as a spider net. By increasing the distance multiplier of this net all the Tweets can be visualized (Figure 24).

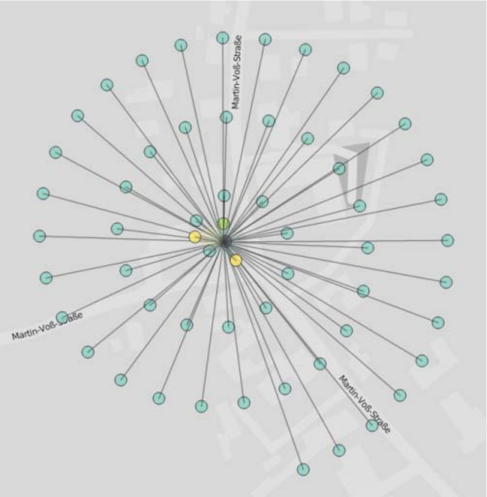


Figure 24: Spider net of Tweets with the same location

Figure 25 shows an area of the map with clustered markers. The clustering algorithm displays with a gray polygon the area on which the clustering is performed.

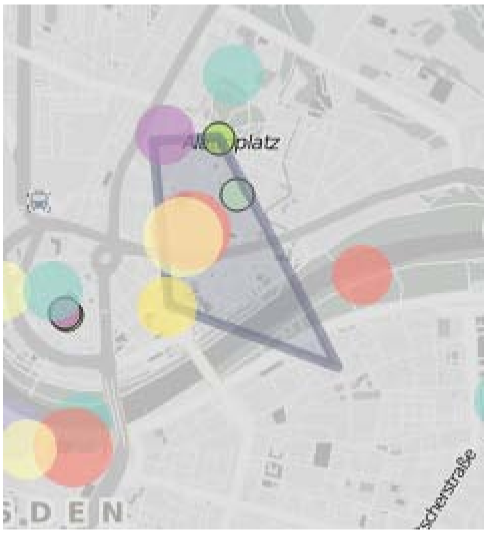


Figure 25: Marker clustering according to topics and distance; visualization of the clustered area with the help of gray polygons.

Although this approach is much better than the previous one because it shows on the map where the topics are stemming from, it still has some deficiencies. One of them is that in the big cities the clusters are very close to each other, creating difficulties when comparing between the amount of Tweets for each topic. For example, in Figure 26 the clusters are overlapping in Dresden and Leipzig for a zoom level of 9.

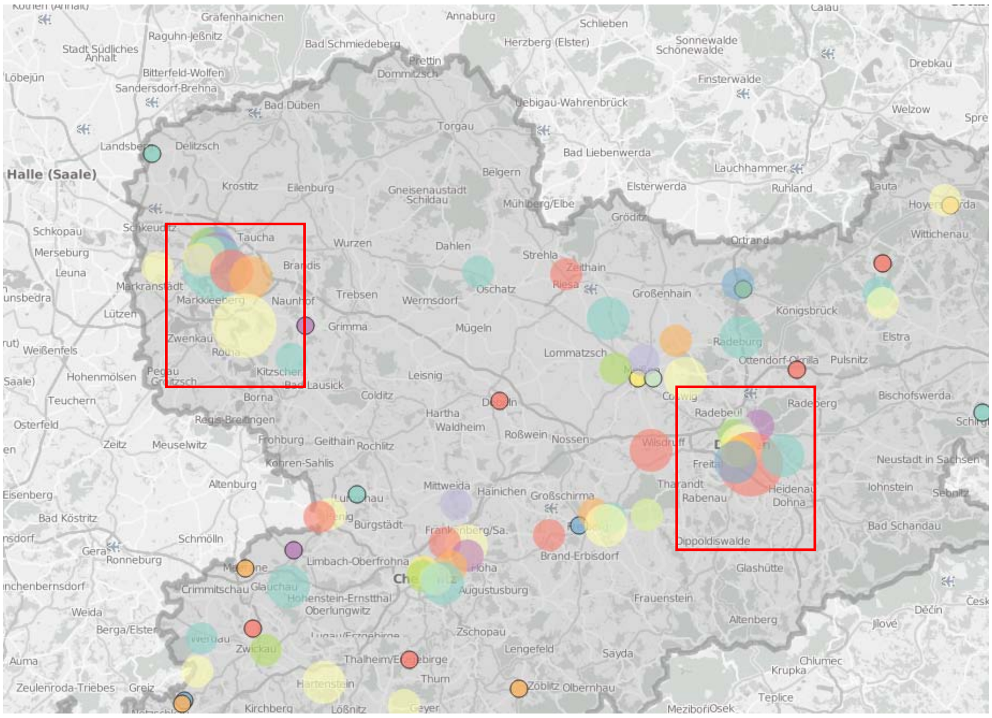


Figure 26: Overcrowded marker clusters in Dresden and Leipzig

The overall quantity of Tweets for each topic can be visualized in a bar chart (Figure 27). Basically the chart replaces the treemap in the old visualization. There is a slight difference in the perception of the information when visualized with a bar chart as opposed to a treemap visualization. And that is, when a topic does not have too many Tweets, the treemap would display an extremely small area that would be barely perceivable. The bar chart will just leave the corresponding bar empty (for example, Transportation in Figure 27. Therefore a bar chart has the advantage that the user can see also the topics that people do not talk about at all.

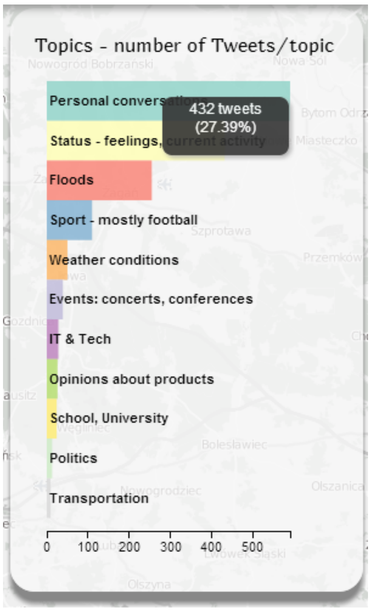


Figure 27: Bar chart visualizing topics and tooltips for showing more quantitative information

The topic bar chart was also built in D3. For the bar chart the data was filtered using the "category" dimension. The category of unclassified Tweets was excluded from the visualization and in the end 11 topics are represented. The bar chart visualization has the advantage of showing clearly what is the most discussed topic for the selected period. Tooltips were added to show also exact percentages of Tweets in the topics. Other possible visualizations could have been a treemap (with the already mentioned disadvantage of not showing at all the topics which are not discussed) or a theme river. The last one could have been probably used also as a timeline as it shows the evolution in time of different topics.

5.2.2 Timeline

The time visualization uses a bar chart built also with D3 that allows the user to select the period of time s/he is interested in. Therefore in the end the Tweets and topics can be visualized on a daily basis, but also on a longer time period if the user chooses so. The analyzed period was reduced to the week of floods in Germany and the region of interest was reduced to Saxony. Tweets are filtered with Crossfilter based on the time component at day level and the results are reflected in the map visualization and the topic bar chart.

Basically the user control lies in the timeline. This allows to see how the topics are changing over time. The user can also zoom in to discover the content of each Tweet. For future development it would be beneficial to allow more interaction by being able to select the topics that s/he sees on a map, i.e. selecting topics by layers or being able to query Tweets by words and displaying only Tweets that match the query. The final application is presented in Figure 28.

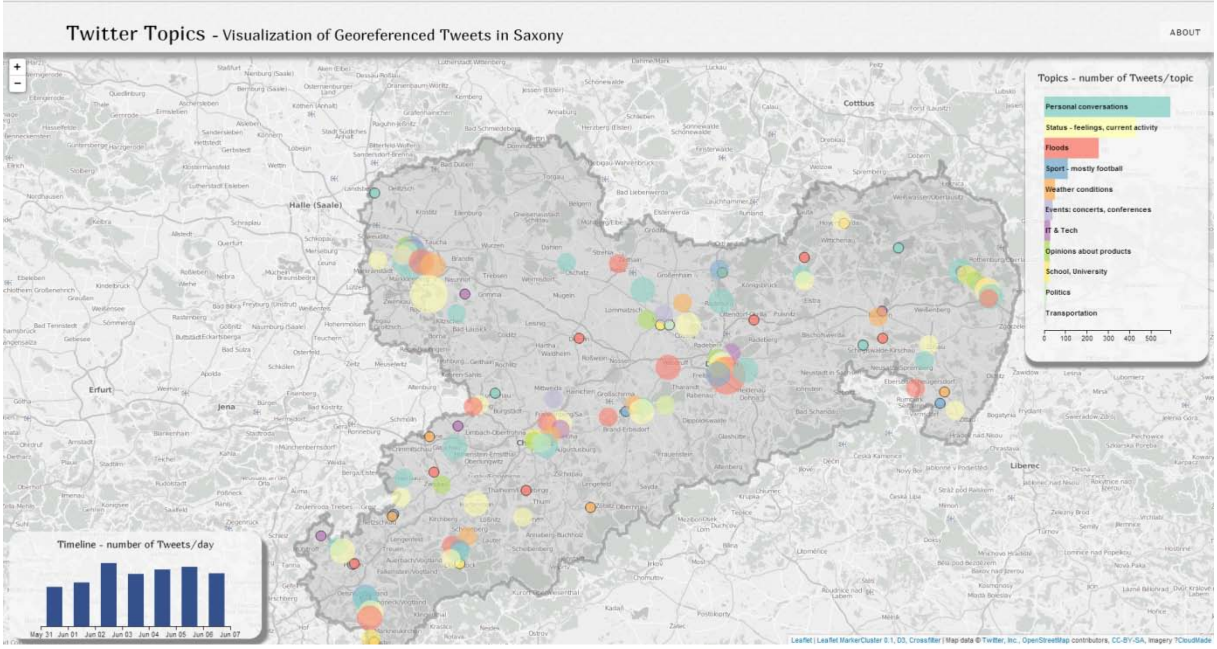


Figure 28: Final web application: Twitter Topics - Visualization of Georeferenced Tweets in Saxony (the final application can be found at: http://wwwpub.zih.tu-dresden.de/~s4080765/TopicVisualization/)

For the practical part of this thesis I would like to mention some of the tutorials that have been extremely helpful in creating the application. For the coordination between the map and timeline views using crossfilter, an extremely clear and helpful tutorial is provided by Brendan Kenny: Maps Garage: Exploring Map Data with Crossfilter. For building the topic bar chart great tutorials about D3 are provided by Scott Murray on his blog. For learning the basics of Crossfilter and the simple principles behind it the blog of Rusty Klophaus was extremely helpful. Other interesting posts about D3 can be found on this blog.

6 EVALUATION OF THE RESULTS

6.1.1 Evaluation of the classification

The classification was done partly manually, partly automatically by assigning a category to a Tweet based on keywords. A Tweet could belong to several categories, but to ease the categorization and the visualization processes only 1 topic was assigned to each Tweet function of its importance. The importance of the categories is judged by their order in Table 4. However ambiguities can arrive when classifying Tweets. For example a Tweet like "This weather makes me tired." was automatically classified in the Weather topic but it actually would belong to a status update, therefore would fit much better the topic status. A manual verification was done on 100 Tweets. Out of these 100, about 5 were discovered to belong to different topics than the ones they were assigned to, therefore an error of about 5% is expected within the currently classified Tweets.

6.1.2 Evaluation of the application

The web application was evaluated in terms of design and usability. The issues that were considered to possibly raise problems were:

1. General purpose of the application: the overall message of the application might not be clear enough for a first time user. To make sure the application delivers its message, an "About” section was created in which the purpose of the application is clearly stated. The evaluation checks to see if this About section was actually useful for the user or not.

2. Design: displaying 11 different categories in an efficient and easy to understand way is not an easy task. The color method was chosen and then it was tested if the users agree with the design of the application when using so many colors.

3. Usability: users might find it difficult to interact with the timeline, it might not be very intuitive for them how to use it. To overcome this, a description was added in the “About” section on how to use the timeline and a tooltip was also added on the title of the Timeline bar chart that explains the user that s/he has to click and drag.

Further on, a survey was made to see if the users indeed encountered the mentioned problems and if some of the solutions presented have helped them to overcome the problems. A screenshot of the questionnaire used for the survey is presented in Figure 29. The survey was made online using SurveyMonkey[48].

For the evaluation of the application a questionnaire was sent to 16 users. Users with no background in cartography or design/usability, aged between 20 and 30 were preferred.

To evaluate the overall design of the application the users were asked to rate it keeping in mind that design refers to colors, contrast, text fonts or positioning of windows on the screen. 5 choice answers were offered ranging from very good to very bad. All the participants marked the design as being either very good (56.25%), or good (43.75%). Therefore it seems the multitude of colors did not pose a big problem for the overall aspect of the web map.

To evaluate how well the web map delivers the information to its users, the participants had to answer a second question: "Did you understand that the purpose of the application is to show the topics discussed on Twitter in Saxony?". The majority of the answers here were "Yes, it was easily understandable" (75%) and some (25%) replied "Yes, but it was not so easy to understand". No participant marked that they did not understand the purpose of the application.

The About section seemed also to be useful as 93.75% marked that this section has helped them in better understanding the web map. 1 answer was negative, saying that the participant found the About section useless.

The third problem, the timeline seemed to be challenging as half of the users confessed to have needed some time to understand how the timeline works. The other half mentioned that they found it easy to understand. In the comments one user wrote that s/he did not know that the time period could be selected until reading the questionnaire. So it seems that this needs some more improvement in terms of usability. One possible solution could be to add tooltips straight on the timeline, not just on the title, but this would imply that the tooltip shows up

[48] https://de.surveymonkey.com/home/

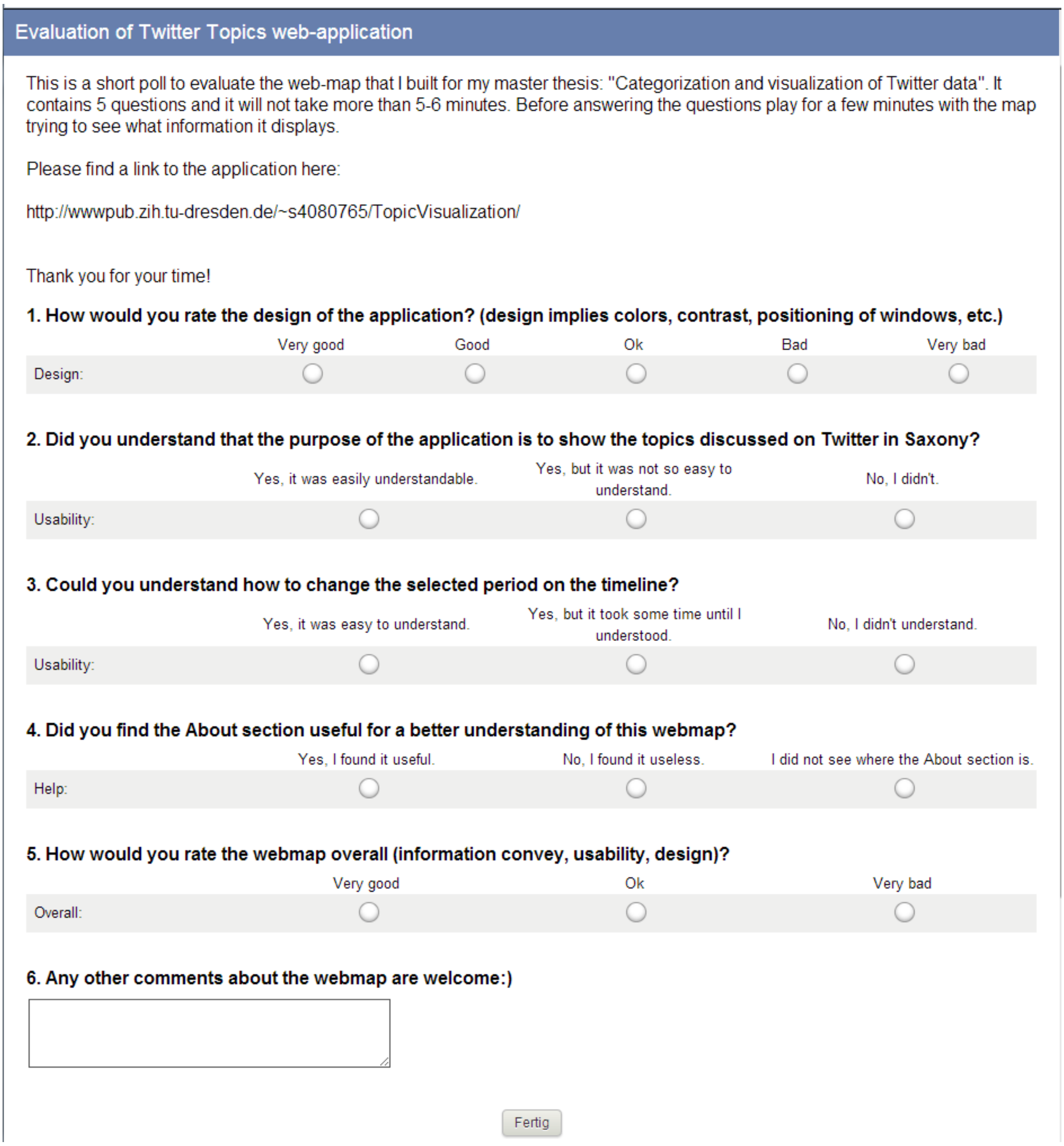
whenever the mouse is on the timeline. For the user, this might be rather annoying, instead of useful.

Some comments mentioned that the gray polygons that mark the clustered areas are confusing for the user and it is difficult to understand what they stand for. An improvement in this case could be changing the colors of the polygons to the color of the cluster they represent.

The whole results of the survey can be seen in Table 6.

How would you rate the design of the application? (design implies colors, contrast, positioning of windows, etc.)	Very good	Good	OK	Bad	Very bad
	56.25%(9)	43.75%(7)	0%(0)	0%(0)	0%(0)
Did you understand that the purpose of the application is to show the topics discussed on Twitter in Saxony?	Yes, it was easy understandable.	Yes, but it was not so easy to understand.	No, I did not.		
	75% (12)	25%(4)	0%(0)		
Could you understand how to change the selected period on the timeline?	Yes, it was easy to understand.	Yes, but it took some time until I understood.	No, I did not understand.		
	50%(8)	50%(8)	0%(0)		
Did you find the About section useful for a better understanding of this web map?	Yes, I found it useful.	No, I found it useless.	I did not see where the About section is.		
	93.75%(15)	6.25%(1)	0%(0)		
How would you rate the overall map design?	Very good.	OK	Very bad		
	68.75%(11)	31.25%(5)	0%(0)		

Table 6: Results of the survey on design and usability of Twitter Topics web application



Evaluation of Twitter Topics web-application

This is a short poll to evaluate the web-map that I built for my master thesis: "Categorization and visualization of Twitter data". It contains 5 questions and it will not take more than 5-6 minutes. Before answering the questions play for a few minutes with the map trying to see what information it displays.

Please find a link to the application here:

http://wwwpub.zih.tu-dresden.de/~s4080765/TopicVisualization/

Thank you for your time!

1. How would you rate the design of the application? (design implies colors, contrast, positioning of windows, etc.)

	Very good	Good	Ok	Bad	Very bad
Design:	○	○	○	○	○

2. Did you understand that the purpose of the application is to show the topics discussed on Twitter in Saxony?

	Yes, it was easily understandable.	Yes, but it was not so easy to understand.	No, I didn't.
Usability:	○	○	○

3. Could you understand how to change the selected period on the timeline?

	Yes, it was easy to understand.	Yes, but it took some time until I understood.	No, I didn't understand.
Usability:	○	○	○

4. Did you find the About section useful for a better understanding of this webmap?

	Yes, I found it useful.	No, I found it useless.	I did not see where the About section is.
Help:	○	○	○

5. How would you rate the webmap overall (information convey, usability, design)?

	Very good	Ok	Very bad
Overall:	○	○	○

6. Any other comments about the webmap are welcome:)

Figure 29: Questionnaire used to evaluate the design and usability of the Twitter Topics Web application.

7 CONCLUSIONS

7.1 ANSWERS TO RESEARCH QUESTIONS

One of the research questions of this thesis was to find a suitable classification method for georeferenced Twitter data, with the main goal of identifying general topics from the Tweets. In the end the topic extraction was carried out using word frequency algorithms, probabilistic topic modeling and manual classification. The first 2 methods, word frequency and topic modeling were employed to create some preliminary topics such as *floods, weather* or *events*. Some words were assigned to them based on the results of the word frequency algorithm and Mallet tool. Further on, starting the manual classification some other topics were discovered, as for example *status* or *opinions*. Probably Tweets from the status category would be impossible to classify automatically because of the big diversity of words used for this category. In the end an amount of 12 topics were discovered (including the Tweets which could not be classified and which were assigned to "no category") and Tweets were assigned to these topics either using SQL queries or manually. The results of the classification were satisfying as most of the tweets were classified correctly. The errors in classification that might have occurred are also due to the ambiguity of some Tweets and to the fact that some Tweets may belong to several topics. A personal curiosity and the reason why the period of floods was selected, was to see if this event would be reflected in the Tweets. It turned out that shortly before the flood (31.05.-01.06.2013) there were very few Tweets that anticipated the hazard and then on the 2nd of June, when the water level began to grow significantly, the number of Tweets began to grow as well. Figure 30 shows how the event took proportions also in the micro-blogging community. However, most of the Tweets were not really useful and would have not been worthy of consideration by relief organizations.

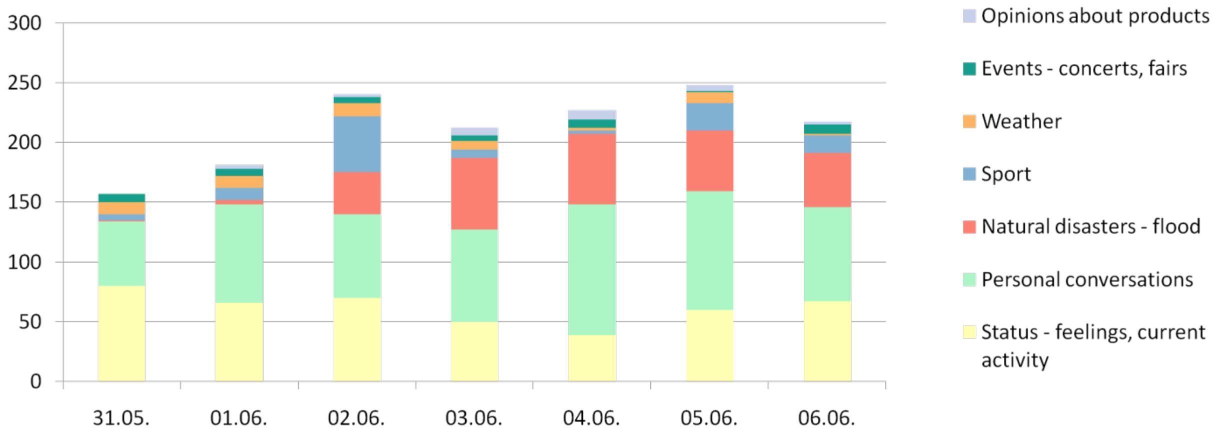


Figure 30: Twitter topic evolution in number of Tweets/day

The second goal of this thesis was to find an appropriate visualization method for the topics both spatially and temporally. After analyzing the available visualization tools that are mandatory for the visualization, it was concluded that a good spatial representation of the topics could be achieved through a clusterization of the Tweets based on the distance but also on the topic they belong to. The disadvantage of this representation shows up when too many clusters form around one point and become overcrowded. In this case, the user cannot distinguish anymore between the different topics on the map.

The timeline is important in discovering the evolution of topics in time and it is necessary that the user has control over the timeline and that s/he can select the time period that s/he is interested in. The timeline used in the web map was the standard one used in the Crossfilter examples.

In the end the current web map allows for visualization of the topics as clustered markers on a map and a quantitative perception of them through a topic bar chart. The user control lies in the timeline bar chart where the user can click to select the time period of interest.

7.2 FURTHER RESEARCH

Although the classification of Tweets proved to be relatively accurate, this method of classification is definitely not optimal and needs to be further automated. For a further research it should also be considered assigning several categories to a single Tweet and the categories could be at their turn divided into sub-categories that would be directly associated with the event or the topic of discussion. For example category "Politics" would have a sub-category "National elections in Germany".

For the visualization part, in the future a more sophisticated time component could be built that also keeps track of real life events. This could be useful to compare the topics discussed in the social media and the events that happen in real life. It has been seen that social media tends to reflect the real life, but it would be interesting to see up to which extent this happens.

Another interesting visualization would be to see how the topics evolve spatially and temporally. The web map built for this thesis shows the topics at the state set by the user through the timeline. It would be an added value to the map to see an evolution of the topics in time. Where were the topics discussed the day before and where will they be one day after? A proper visualization method should be investigated in this case.

As a general conclusion, Twitter data classification and visualization are not easy tasks, but they can lead to very interesting results. Micro-blogging communities tend to be a mirror of the society and data extraction and visualization could help in better understanding it. This master thesis is meant to be a small step in this direction and many improvements still need to be done for achieving significant and accurate results.

8 BIBLIOGRAPHY

8.1 BOOKS, JOURNAL ARTICLES AND CONFERENCE PROCEEDINGS

Alhadi, A.C., Gottron, T. & Staab, S., 2011. Exploring User Purpose Writing Single Tweets. Koblenz, 2011. WebSci '11.

Andrienko, G. et al., 2013. Thematic Patterns in Georeferenced Tweets through Space-Time Visual Analytics. *Computing in Science and Engineering*, 15(3), pp.72-82.

Bird, S., Klein, E. & Loper, E., 2009. *Natural Languages Processing with Python*. Sebastopol: O'Reilly Media.

Blei, D.M., 2012. Probabilistic Topic Models. *Communications of the ACM*, 55(4).

Blei, D.M., Ng, A.Y. & Jordan, M.I., 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* , (3), pp.993-1022.

Cao, N. et al., 2012. Whisper: Tracing the Spatiotemporal Process of information Diffusion in Real Time. *IEEE Transactions on Visualization and Computer Graphics*, 18(12).

Chae, J. et al., 2012. Spatiotemporal Social Media Analytics for Abnormal Event Detection and Examination using Seasonal-Trend Decomposition. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference*. Seattle, WA, 2012.

Dann, S., 2011. 28 Boxes: A content Classification Framework for Twitter. In MacCarthy, M., ed. *Australian and New Zealand Marketing Academy Conference*. Perth, 2011. School of Marketing, Faculty of Business and Law Edith Cowan University.

Dörk, M., Carpendale, S., Collins, C. & Williamson, C., 2008. VisGets: Coordinated visualizations for web-based information exploration and discovery., 2008. Proceedings Information Visualization and Computer Graphics.

Earle, P. et al., 2010. OMG Earthquake! Can Twitter Improve Earthquake Response? *Electronic Seisomologist*.

Goodchild, M.F., 2007. Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. 2.

Guy, M. et al., 2010. *Lecture Notes in Computer Science*, (6065), pp.42-53.

Hahmann, S. & Burghardt, D., 2013. On the influence of Geospatial Context on Mobile Microblogging Contents. In *Proceedings of the 26th International Cartographic Conference*. Dresden, Germany, 2013.

Hall, M. et al., 2009. The WEKA Data Mining Software: An update. *SIGKDD Explorations*, 11(1).

Hao, M. et al., 2011. Visual Sentiment Analysis on Twitter Data Streams. In IEEE, ed. *IEEE: Symposium on Visual Science Analytics and Technology*. Providence, USA, 2011.

Hong, L. & Davidson, B.D., 2010. Empirical Study of Topic Modeling in Twitter. In *1st Workshop on Social Media Analytics*. Washington DC, 2010.

Horn, C., 2010. *Analysis and Classification of Twitter messages*. Graz: Graz University of Technology.

Jackoway, A., Samet, H. & Sankaranarayanan, J., 2011. Identification of Live News Events Using Twitter. In *Location Based Social Networks*. Chicago, 2011. ACM.

Java, A., Song, X., Finin, T. & Tseng, B., 2007. Why we Twitter: Understanding Microblogging Usage and Communities. San Jose, 2007. ACM.

Krumm, J., Davies, N. & Chandra, N., 2008. User generated content. (8).

Lee, K. et al., 2011. Twitter Trending Topic Classification. In IEEE, ed. *International Conference on Data Mining Workshops*. Vancouver, 2011.

Leetaru, K.H. et al., 2013. Mapping the global Twitter heartbeat: The geography of Twitter. *first monday - online journal*, 18(5-6).

MacEachren, A.M. et al., 2011. SensePlace2: GeoTwitter Analytics Support for Situational Awareness. Providence, 2011. IEEE Symposium on Visual Analytics Science and Technology.

Marcus, A. et al., 2011. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 annual conference on Human factors in computing systems (CHI '11)*. New York, 2011. ACM.

Naaman, M., Boase, J. & Lai, C.-H., 2010. Is it Really About Me? Message Content in Social Awareness Streams. Savannah, 2010. ACM.

Rajaraman, A., Leskovec, J. & Ullman, J.D., 2012. Data Mining. In *Mining of Massive Datasets*. Palo Alto: http://i.stanford.edu/~ullman/mmds/ch1.pdf. pp.7-8.

Sakaki, T., Okazaki, M. & Matsuo, Y., 2010. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. Raleigh, North Carolina, 2010. WWW2010.

Samet, H. et al., 2009. TwitterStand: News in Tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. Seattle, 2009. ACM New York.

Sankaranarayanan, J., Hanan, S. & Jackoway, A., 2011. Identification of live news events using Twitter. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*. Chicago, 2011. ACM New York.

Vosecky, J., Jiang, D. & Ng, W., 2013. Limosa: A System for Geographic User Interest Analysis in Twitter. Genoa, Italy, 18-22 March 2013.

Zhao, W.X. et al., 2011. Comparing Twitter and Traditional Media using Topic Models. *Research Collection School Of Information Systems*, (http://ink.library.smu.edu.sg/sis_research/1375/), pp.338-49.

8.2 ONLINE RESOURCES

Aisch, G., 2012. *Why We Need Another Mapping Framwork*. [Online] Available at: http://vis4.net/blog/posts/introducing-kartograph/ [Accessed 7 September 2013].

GeoExt, n.d. *GeoExt*. [Online] Available at: http://opengeo.org/technology/geoext/ [Accessed 6 September 2013].

Holt, R., 2013. *Twitter in Numbers*. [Online] Available at: http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html [Accessed 6 May 2013].

Knoesis, 2012. *Knoesis.org*. [Online] Available at: http://twitris.knoesis.org/election/network/ [Accessed 21 May 2013].

McCallum, A.K., 2002. *MALLET: A Machine Learning for Languages Toolkit*. [http://mallet.cs.umass.edu].

McCrocklin, M., n.d. *Coordinated Visualizations*. [Online] Available at: http://bl.ocks.org/milroc/raw/6316349/#0 [Accessed 6 September 2013].

Mitchell, L. et al., 2013. *The Geography of Happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place*. [Online] Available at: http://arxiv.org/abs/1302.3299v2 [Accessed 15 May 2013].

Murphy, I.B., 2013. *Fast Database Emerges from MIT Class, GPUs and Student's Invention*. [Online] Available at: http://data-informed.com/fast-database-emerges-from-mit-class-gpus-and-students-invention/ [Accessed 2 May 2013].

Stephens, M., 2013. *The Geography of Hate*. [Online] Available at: http://www.floatingsheep.org/2013/05/hatemap.html [Accessed 16 May 2013].

Tsukayama, H., 2013. *Twitter turns 7: Users send over 400 million tweets per day*. [Online] Available at: http://articles.washingtonpost.com/2013-03-21/business/37889387_1_tweets-jack-dorsey-twitter [Accessed 02 September 2013].

Twitter, 2013a. *The Twitter Glossary*. [Online] Available at: https://support.twitter.com/articles/166337-the-twitter-glossary [Accessed 9 September 2013].

Twitter, 2013b. *About public and protected Tweets*. [Online] Available at: https://support.twitter.com/articles/14016-about-public-and-protected-tweets [Accessed 18 September 2013].

Twitter, 2013c. *Twitter Support*. [Online] Available at: https://support.twitter.com/articles/78525-about-the-tweet-location-feature [Accessed 9 September 2013].

Twitter, 2013d. *Twitter Libraries*. [Online] Available at: https://dev.twitter.com/docs/twitter-libraries [Accessed 2 September 2013].

Twitter., 2013e. *Using the Twitter Search API*. [Online] Available at: https://dev.twitter.com/docs/using-search [Accessed 26 August 2013].

Underwood, T., 2012. *Topic Modeling made just simple enough*. [Online] Available at: http://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/ [Accessed 14 September 2013].

Viau, C., 2012. *What's behind our Business Infographics Designer? D3.js of course.* [Online] Available at: http://www.datameer.com/blog/uncategorized/whats-behind-our-business-infographics-designer-d3-js-of-course-2.html [Accessed 5 September 2013].

Wm, L., 2012. *The top seven alternatives to the Google Maps API*. [Online] Available at: http://www.netmagazine.com/features/top-seven-alternatives-google-maps-api [Accessed 7 September 2013].

Wolf, L., 2011. *Twitter Statistics - 2008, 2009, 2010, 2011*. [Online] Available at: http://womeninbusiness.about.com/od/twittertips/a/twitter-statistics.htm [Accessed 6 May 2013].