Microproject 1: Social Business Data Structures and Access Options

As a social media manager at Amazon, you could be tasked with accessing social media from sites such as Facebook, Twitter, Instagram or more, in order to see the opinion of the public about the company's services. This method is called "scraping" and includes several applications such as Octoparse, Socinator, Dexi.io, Scrapinghub or Parsehub. To do so, we applied the Octoparse software on the social network Twitter to collect the views, based on the tweets about the company and the sentiment of the service management. The Octoparse system is a well-known and efficient software that extracts the data from a desired social network under certain steps and creates an output as a CSV-file or Excel table for analysis. We first install the free Octoparse application after signing in, in order to get access of the data on the Social Media Network. After that we can start with the extraction of the content about Amazon, that is generated from the users (customers, suppliers…). The steps to extract the data in a social network using Octoparse are as follows:

Step 1: We first enter a URL (https://twitter.com/Amazon) in the empty space to have access to the Amazon twitter account.

Step 2: We set up the pagination loop to scrape data from multiple posts, for that the Octoparse can move to the next page and thus, generate a complete dataset.

Step 3: We build a loop list to incorporate all the parties containing the data fields. Here, we define the elements that the application should scrape, for example the date, the page title, the text content, the used hashtags, the number of likes, comments, the text content of all comments with the assigned users.

Step 4: We select and extract the data. In this case, we first click on the corner of the first tweet and then, we select manually one by one each element on the tweet as for example the date or number of retweets. After that, we click on "extract text of the selected element" to extract these selected elements. We repeat this action on the other tweets.

Step 5: Finally, we save and start the extraction: we can now run the crawler on our local device to get the data and export it as CSV-, Excel-, Html- or Json-file.

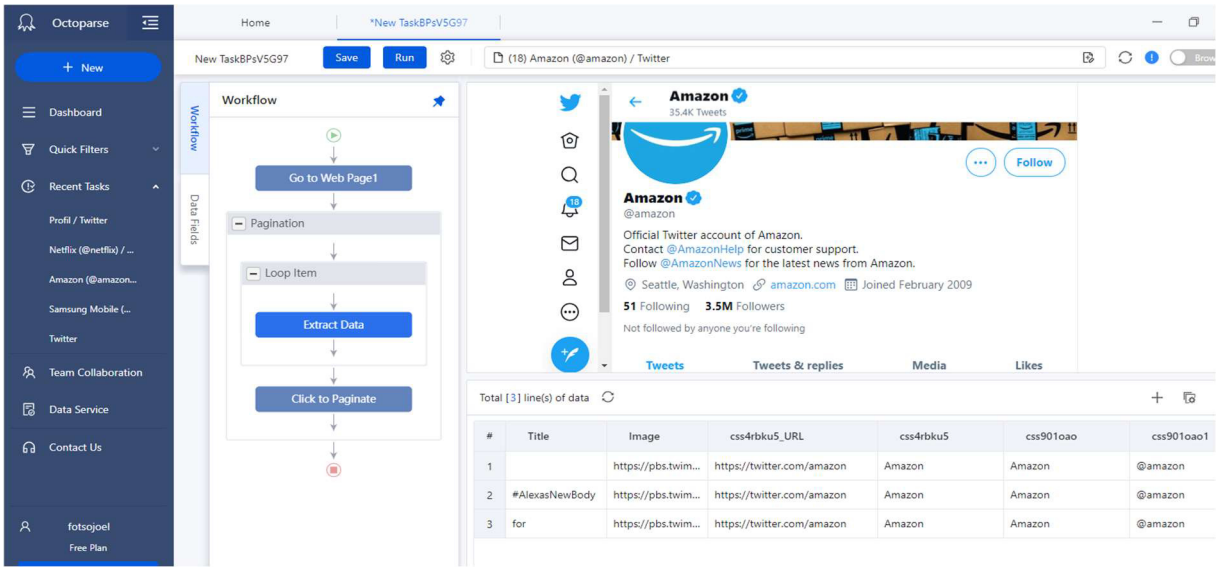

Figure: Process of data scraping from amazon with Octoparse.

After extraction, the collected data is divided onto a spreadsheet with several columns. Therefore, one column is for the date, one another column is for the text content, and the other columns contain the number of likes of comments, of shares, the URL of each collected dataset/observation. The Octoparse system works with a crawler that collects data on a desired target website and outputs it in a CSV-file or Excel table. From this point, further analysis and visualisation can be created and compared. Moreover, it is also possible to categorize these data. On the Octoparse system, Hashtag trend analysis on the social network as Twitter, Facebook or Instagram can be performed and also the analysis of the development of posts or likes on individual post. The Price developments of e-commerce providers can also be analysed, and market niches can be discovered.

Microproject 2: Tools and Dashboards

Nowadays the competition between firms has increased in the market, due to that the companies must be able to design and to create new strategies to improve their performance. One recommendable strategy for any firm is the analysis of the information about the market conditions and their commercial and logistic process. There are several tools that can be used to make these studies, and, in this document, we will compare two of them named Hootsuite and RapidMiner.

RapidMiner is a tool that offers several functions to analyse data sets which can be available in a web page or stored on the computer. The users must pay a quote to use the program, however there is a one-year free trial for students. The program provides several options to analyse the data and to run them. It is not necessary that any code is written by the user. The data can be modified in the program to determine the relevant sample for the analysis. RapidMiner can blend structured with unstructured data and then leverage all the data for predictive analysis. The program can execute tasks in the field of Descriptive Statistics such as: Numerical attributes (mean, median, minimum, maximum, standard deviation, and number of missing values). In addition, it can generate matrices, for instance Covariance, Correlation and Anova-Matrix. It has a wide variety of graphs to show the output such as: Line, Bubble, Parallel, Deviation, Histograms, and others. Furthermore, RapidMiner contains a set of modelling capabilities and machine learning algorithms, it can create Regressions, Decision Trees and Model Ensembles. The program offers a wide variety of statistics resources that are very useful to make for example Customer Relationship Management Analysis or to evaluate the company's providers services.

In the case of Hootsuite, the program focuses its service on managing the content of several profiles in social media at the same time. The users link their social media accounts such as: Facebook, Twitter, Instagram to the program. Hootsuite has 30 days free of payment, however after this period the users must pay a quote. It is very useful to manage all the social media accounts on the same platform, the users can create new content easily, and they can program schedules for each post. In addition, this tool offers analyses about the uploaded content, it generates graphs and tables that express the interaction with the public. It can be used to promote campaigns on the different platforms and to monitor the results. Hootsuite summarizes in its analyses the account performances, it can compare the results with previous periods and detect the main changes in the data related with followers, fans, and posts. Furthermore, it can calculate the real return on your social media investment and demonstrate how these channels and campaigns drive conversions and sales.

If we compare RapidMiner and Hootsuite we can define that both are very useful tools to analyse datasets in the business environment. They contain many functionalities that bring about several advantages to the efficiency in the business operations. However, the strength of each application resides in the area in which they operate. RapidMiner can be used to any type of data, it is not limited to the management of social media, it can be applied in any logistic activity or in the company research. On the other hand, Hootsuite offers the service of scheduling the posts and saves time in this type of activity. According to our point of view, we think that we support more the use of RapidMiner because it offers more statistical tools than Hootsuite, in addition the same data from the social media can be analysed in RapidMiner.

Microproject 6: Text-Mining in RapidMiner

For this Microproject, we were provided with a set of tweets in a text-file which mention the city of Paderborn. A first glimpse reveals, they are all in English. That is important to know for the stop-word-library later.

Furthermore, the Task provides us with information about the RapidMiner, that we can download in a free version after signing in. Here, we are recommended to download the extension “Text Mining Extension”.

After Downloading and installing, we can start with pre-processing the text. For this we load the TwitPADSet and extract the text from the third column before we parse the data in a so called “Process Document”. Here we start with the pre-processing. To get a Bag-of-Words we need to tokenize the text. Because the cases for the same word are sometimes upper or lower which would lead to two different words, “PADERBORN” and “paderborn” for example, we set all letters to lower cases. This reduces dimensionality later. The next step is the stop-word-filter, that filters out all tokens that are necessary for syntax but give no information. These would make too much noise when analysing the data. For this filtering it is necessary to know the language from the dataset thus it is dictionary-based. The last filter is for single letters, that occur by accident or are meant as abbreviation whose meaning would need context to understand. Subsequently to this, we implemented the tf-idf as a part of pre-processing. At last, we modelled the actual topic-extraction from the documents with Latent-Dirichlet-Allocation.

The result provides two tables. One gives the topics with the top five tokens for each topic. The second gives us a larger table with the processed tweet, the predicted topic, and the probability to belong to each of the allocated topics.

For 4 out of the 5 topics, one of the top words is “paderborn” and we maybe have too many topics. Only topic 5 does not contain “paderborn”. This topic top words are (“girl”, “teen”, “dating”, “sex”, “girls”) which suggests that in this topic there are spam-tweets. In topic 1 the top words contain “low” and “cloudy” and all tweets about the weather are predicted to be in this topic. Unfortunately, many others too, that we can not say this is the weather-topic.

Here one main disadvantage comes to be prominent. Topic modelling is an unsupervised learning method that requires a lot of domain knowledge to interpret the outcome.

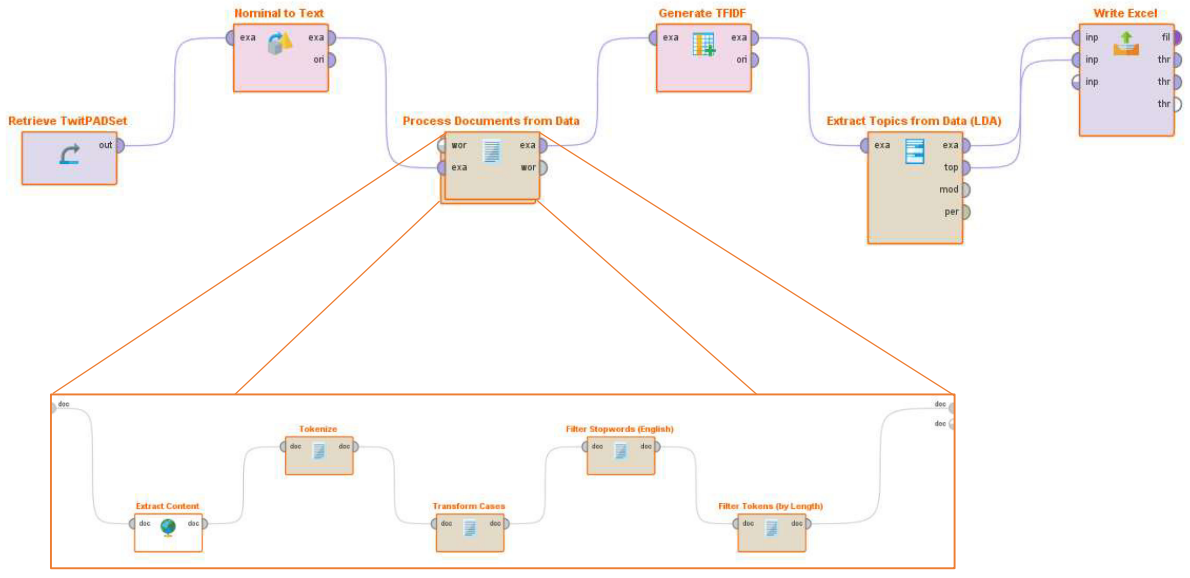

Figure 1: The process for topic-modelling with RapidMiner Studio, based on the provided text-data from Twitter.

Text Mining in Social Networks

with an Application in R

Course

K.184.43241 Social Business Analytics & Management

Winter Term 2020/2021

Seminar Paper

Cruz Gil, Alexandra, student number: 6873141, email: cruzgil@mail.uni-paderborn.de

Fotso Tenku, Joel Cedric, student number: 6810782, email: joelf@mail.uni-paderborn.de

Düx, Patrick, student number: 7110155, email: palbert@mail.uni-paderborn.de

Abstract

The use of platforms such as Facebook, Instagram, Twitter, and others generates a lot of daily information on the Internet. The analysis of this content can be a useful resource to design new strategies that can improve the firm's performance, nevertheless this process tends to be complicated, and it requires previous knowledge about the research topic and skills in the field of Data Science. To develop these studies, tools for Text Mining can be used to analyse the content structure and identify trends and patrons in the dataset, however the procedure to develop this technique correctly is not well defined due to it depends on the data feature and structure. This paper will show the theoretical concepts in the Text Mining Process and it will expose a practical case which use as research topic: Text Mining, Data Mining and Graph Mining. Furthermore, the several advantages that brings the Text Mining application will be discussed and the example will explain the main problems that were detected in the Text Mining process.

Keywords: Text Mining, Keywords, algorithm, Data Mining, Graph Mining.

1 Introduction

Currently the use of Internet is part of the daily routine for billions of people around the world. According to Statista (2021a), the slope of the curve that represents the number of Internet user around the world has showed a positive value. The last October 2020 were 4,66 billion people active in Internet, number that represents the 59 percent of the global population. Every day different activities are developed by individuals on web platforms that generate a large amount of structured and unstructured data which can be used in several fields of analysis about the modern society. Due to the previous situation, companies and institutions have created tools and strategies to analyse the data stored in the world-wide-web to improve their business performance or to achieve specific goals.

The available data on the Internet has several formats, the user's posts can be expressed as videos, audios, photos, texts etc. and getting an understanding of its information is a complex process which require specialized tools and previous knowledge in the research field. In order to process the previous information, the data miners apply a technique called Text Mining, which processes the information in

the post expressed as text and gives trends and patterns, that characterize the individual elements in the dataset and the relationship between them, as output. The steps to develop an efficient and correct text mining analysis is not strictly defined in the literature due to this procedure depend on several elements that cannot be controlled by the researchers. Due to that and the advantages that this procedure offers in the business area, this paper firstly will introduce some theoretical concepts related to the topic Text Mining Analysis in Social Networks. After that, it will expose an example that applies the steps defined in the paper (Lai and To, 2015) to develop a Text Mining analysis in a dataset extracted from the social network Twitter. The program R will be used to filter the information and execute the functions and the criteria that were selected to search for data have been Tweets that contain the terms Text Mining, Data Mining and Graph Mining.

2 Text Mining in Social Network-Analysis

Text Mining is a technology that constantly is used for several users to analyse available information on the Internet. Its popularity has increased in the last years due to the several benefits that this strategy offers. This technique brings about the possibility of detecting trends that can be used to improve the business operations or to develop research about relevant topics in the society. The key element that defines Text Mining from other techniques such as Data Mining, is that it provides information that is stored in the text structure and is unknown for the users (Hearst, 2003). The output of an analysis using Text Mining can support some theories or generate new ones. Data Mining can be described as looking for patterns in data (Hand and Adams, 2014). There are two key aspects in the Data Mining field, the first one is model building and the second one pattern detection. Model building in data mining is remarkably like statistical modelling, although new problems arise because of the large sizes of the data sets and the fact that data mining is often secondary data analysis. Pattern detection seeks anomalies or small local structures in data, with the vast mass of the data being irrelevant. Indeed, one view of many large-scale data mining activities is that they primarily constitute filtering and data reduction (Hand and Adams, 2014). As it was mentioned before Data Mining focuses on the patterns in the data. However, in the case of Text Mining, the analysis is focused on the patterns in text (Witten, 2004). Large databases could contain meaningful connections in its structure among the different texts, and this tool can recognise the patterns that characterise the relation between the information, such as citations in the academic literature and Hyperlinks in the Web literature, a task that is not included in the natural language processing. Text Mining covers the field of the automatic natural language processing, but it has in addition, the capacity of analysing the linkage structures (ib., 2004).

The process of mining a dataset compounded by texts is complicated, because it has several steps that are necessary to get an output that reflects patrons with logic information. First, it is necessary to preprocess the texts and they must be saved as structured data (Hotho and Nürnberger, 2005). The first step is named tokenization, the punctuation marks are deleted and substituted by replacing tabs and other non-text characters by single white spaces. The output of this process is as set of different words named dictionaries. The dictionary can contain unnecessary information due to that, there are several methods in order to reduce its size and only let in this location the meaningful words. For example, it is possible to eliminate articles, conjunctions, and prepositions. Other criteria to reduce the dictionary size is the elimination of words that are not used frequently or words that are used with high frequency (ib., 2005). There are many criteria that can be used to reduce the dictionaries size, but each depend on the data miners. This process is complex and, on several occasion, can lead to errors due to the text's structure tends to be messy.

The previous paragraph explains the tokenization process, how the data can be reduced to filter the most relevant words. In addition, there are several methods that can be used to structure the data. This process can simplify the access to the document collection. The existing methods for structuring collections either try to assign keywords to documents based on a given keyword set (classification or categorization

methods) or automatically structure document collections to find groups of similar documents (clustering methods) (Hotho and Nürnberger, 2005). The output of this process can be used to make analytical analysis about the set of available documents using the connection between the elements and the most relevant words in the dictionaries.

Networks also can be defined as social systems where the relationship among the elements in the system, indicated as links, are fundamental. The actors that form the network structure are typically defined in the nodes and they have specific features that distinguish them. In the field of network analysis these characteristics are called attributes. The individual nodes are not the only elements that have attributes, the links (or edges) between the nodes also have attributes (Borgatti, 2018). For instance, an individual node can represent a person with the following attributes (feminine, 25 years) and other node can be defined as (masculine, 24 years), both have different individual features, but they can have in common attributes, for example: both study in the same university or both live in the same city and they maybe are linked together as friend, married or in other relationships. Another element that can be described in a network is the information generated using the social media and all the data created using platforms as Myspace, Facebook, Cyworld, Twitter or Facebook, this information is a useful resource to analyse the human behaviour. An individual's social network is straightforward the aggregate of relationships contracted with others, and social network analysis examines the differing structures and properties of these relationships (Milroy and Llamas, 2013). Social networks sites as web-based services that bring to the user facilities such as the creation of public or semi-public profile within a bounded system, articulate a list of other users with whom they share connection and view and traverse their list of connections and those made by others within the system (Boyd and Ellison, 2007).

Every year the number of users that socialize on platforms such as Facebook, Instagram, Twitter increase significatively. According to the web page Statista (2021b), Facebook the first social network to overcome one billion registered accounts, currently have more than 2,74 billion monthly active users, creating constantly new content in the web. These users are persons who share information on social media about their criteria in several topics or about their routines and interests. Mainly the post structures contain a lot of information expressed in texts. The use of text mining analysis not only focus on the content of the post, in fact the relationship between actors that shared similar content or that have common attributes can be used to develop studies in some scenarios. The relationship between the agents on social media, can be expressed as a network and its content reflect the common an individual attributes that define the nodes in the system.

Text mining is a useful resource to analyse social media content. This procedure examines the differing structures and properties of the user relationships. The data which can be extracted from these sites tend to be unstructured and have several elements such as signs, images, and figures. Due to the previous situation, it is necessary in the tokenization process to filter for the meaningful information. After that, the text miner can define additional criteria to reduce the size in the dictionaries and only take in consideration the keywords that are related with the research topic.

3 Literature overview

3.1 Social Network

According to Aggarwal and Wang (2011) based on their study on the text mining of social networks, they can be considered as a set of nodes, and each node contains a certain amount of the text content or local information and is connected to one another by links. Therefore, they present in their study several algorithms that combine both text content and linkage-structure for using text mining techniques in social networks in the context of a variety of problems such as the keyword search, clustering, and classification. However, these algorithms are more used in the context of the XML domain, but they can also be used in the social network domain.

3.2 Keyword Search

In keyword search, the problem is to determine the set of connected clusters of nodes in social networks containing the specific keywords corresponding to the query. Thus, there are algorithms combining text content and link structure that involve the keyword search over a set of documents as such as XML-documents, as well over a graph (Aggarwal and Wang, 2011). Over the documents, several algorithms based on the SCLA (smallest lowest common ancestor) semantics have been developed as the Indexed Lookup Eager Algorithm, Scan Eager algorithm, the Stack algorithm in order to find the smallest sub-trees that contain the keywords (Xu and Papakonstantinou, 2005). Thus, the set of answers obtained on the base of SCLA according to the query is a subset of answers for the ranking method XRank (Aggarwal and Wang, 2011). Over the graph, many algorithms have been proposed such as BANKS, the bidirectional algorithm, and BLINKS, so the aim is to find a small group of link-connected nodes in the graph which are related to a particular set of keywords. The ObjektRank algorithm on the other hand applies authority-based ranking to keyword search on labelled graphs and returns nodes having high authority with respect to all keywords (ib., 2011).

3.3 Classification

In context of the classification process, the nodes are labelled based on the text only classifier as for example the naive bayes, TFIDF and Probabilistic Indexing (Aggarwal and Wang, 2011). In order to increase the effectiveness of the classification process, many classification algorithms that combine both the text content and link structure have been proposed. Therefore, there exist two methods using both text content and links-structure for classification on social network. (ib., 2011). The first method is the use of the Hypertext classifier as the Bayes classifier that is used when the classes of neighbours of the documents are known, and the iterative relaxation algorithm, that is used when some or all of the neighbour classes are unknown (Chakrabarti et al., 1998). A second method is the utilization of the graph regularization approach for text and link-based classification (Aggarwal and Wang, 2011). Thus, we have the Linear regularized combination and the Linear kernel combination that combine local information with global link structure (Zhang et al., 2006)

3.4 Clustering

In the case of clustering, the set of nodes on the social network with the same content is identified and closely connected to one another. For this, they propose many algorithms that use both the content and the linkage structure in order to perform the clustering process (Aggarwal and Wang, 2011). The graph clustering algorithms as for example the SA-Cluster based on the structural and attribute similarities between the node and the distance measure between them. The goal is to divide a large social network graph associated with attributes into clusters so that each cluster contains a subgraph with uniform attribute values (Zhou et al., 2009). And then, the i-Topic-Modelling algorithm has been developed for the clustering process of the document and consist to calculate the probability for the topics to occur in the document (Sun et al., 2009). On the other hand, the use of the latent position cluster model, in which the probability of the link between the node depends on the distance between them, greatly improves the quality of clustering (Handcock et al., 2007)

In the case of transfer learning Aggarwal and Wang (2011) show how the use of the link-information can perform the classification and the cluster tasks. The link can return to the images, the content information, tags, videos and can used for transfer learning. Therefore, it proposes in its study the aPLSA-algorithm (Annotated Probabilistic Latent Semantic Analysis) that consists to collect image data from those links and uses it in order to construct a text to image mapping.

4 Conceptual Framework

As a concept to manage social media content analysis, we suggest the approach from Lai and To (2015), where the analysis is following a four-step-model, as stated in figure 1. Our suggested framework applies the algorithms, that combine text-mining with network-analysis to the model. In the following section

we discuss the steps in particular. For the second and third step we use a little different attempt that suit our topic better.

4.1 The Scope of the Analysis

Based on the specified scope there are different tools, algorithms or methods that provide the appropriate outcome that leads to different data-collections needed and make different assumptions on the method chosen, which makes this step the most critical phase in the concept (Lai and To, 2015). The Scope to alter the company's web-page visibility, for example, needs other approaches, namely some keyword-search-analysis, than the customer segmentation from the online-shop the company is running or the classification of ratings or Emails.

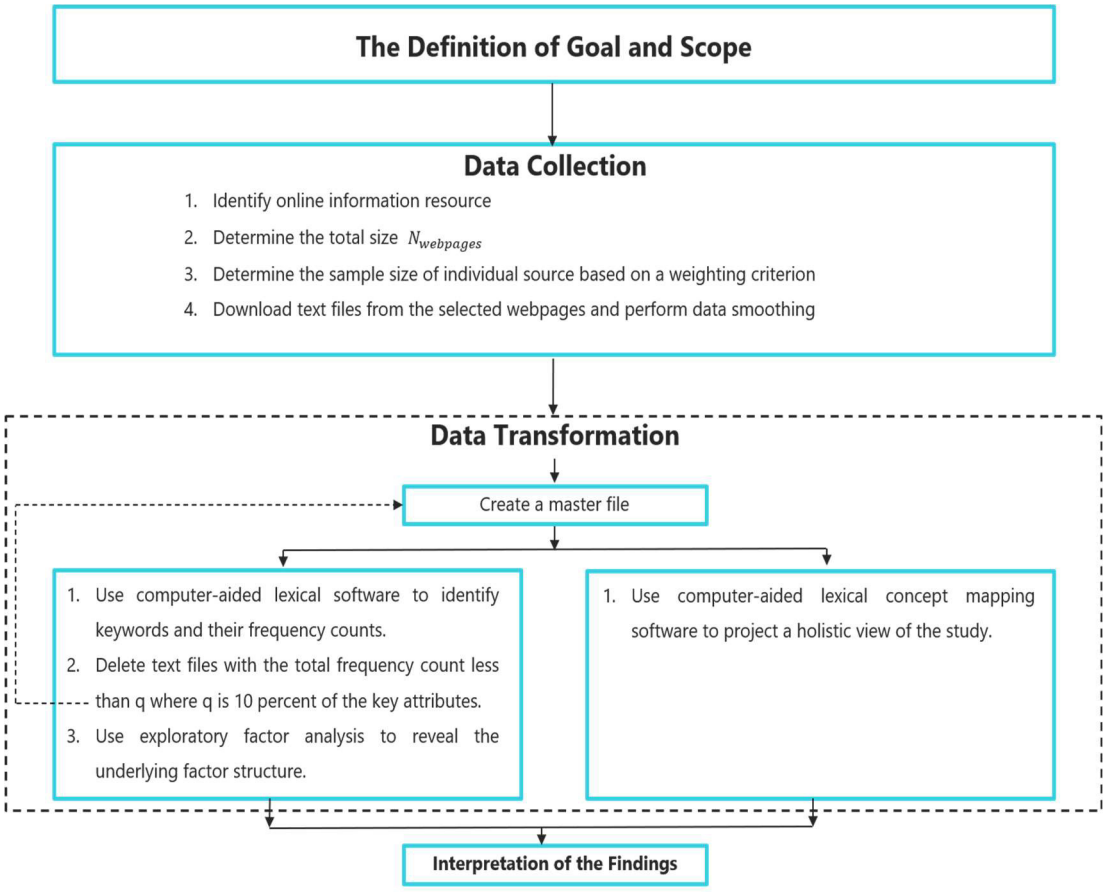

Figure 1: Four Phase of the ''social-media-to-concepts'' approach borrowed from Lai and To, (2015)

4.2 Data

The first challenge after the definition of the scope is to get the data. Here the three different sources Application Programming Interfaces (API), a web-crawler or a downloaded file are normally available (Debortolli et al., 2016) and can be applied on the Company's social media accounts. The requirements on the Dataset and -structure is incrementally dependent on the task and therefore based on the findings in step one.

After the data is gathered, the pre-processing step is required to extract and select the features because text-data is normally unstructured and loosely (Ifran et al., 2015). At first features from the text are extracted. Here the steps tokenization, stop-word-removal, stemming, part-of-speech- (POS-) tagging,

parsing, keyword-spotting are the state-of-the-art tasks with one limitation for the last one which needs a large database in form of dictionaries (ib.). The tokenization divides the text into sequences of words, which can have the length of n, a number between 1 and the number of words in the text, so called n-grams (ib.). With that tokenized text, the stop-words are removed to reduce the number of words in the document and improve the effectiveness of the processing (ib.). Stemming the words is the technique to reduce the words to their root form, which also reduces the number of different words with the same root because appended suffixes, which are grammatically required, are removed from the words (ib.). In case of POS-tagging each word gets an attribute, dependent on its part of speech (ib.). With these attributes one can for example filter for nouns. This requires dictionaries for example (ib.). Parsing is, next to POS-tagging, another approach to use the syntactical information of the word. Here the whole sentence is dissembled in a tree-like structure, which can be used to analyse it for the correct syntactical and grammatical structure (ib.). The keyword-spotting technique helps to decide, whether a text has useful content or not by seeking for pre-categorized words (ib., Strapparava and Ozbal, 2010, Ling et al., 2006). The second step is the feature selection from the vectorized documents where each record represents a word or set of n words (Ifran et al., 2015). The extracted features are still full of irrelevant information thus filtering for the relevant information is needed (ib.). The most common methods to get the most important features is the term-frequency inverse document frequency (tf-idf) (ib.). The term-frequency is defined as how often a single term t occurs in a document while the idf is the ratio between the total number of documents to the number of documents that contain the term t. The tf-idf is the product of these both measurements.

Afterwards the two most common analysis-techniques, the classification and clustering, can be implemented (Ifran et al., 2015). For text classification in networks first labelled data is usually required and second, the algorithms need a text-corpus and a linkage (Aggarwal and Wang, 2011). In most algorithms they are embedded in hypertext and connected by hyperlinks (Chakrabarti et al., 1998; Zhang et al. 2006). Other algorithms like the Email-classification from Cohen et al. (2004) or from Carvalho and Cohen (2005) require also a labelled training and test-set with specific labels based on the data. As described the purpose of clustering can be the enhancement of text-mining or the graph-clustering. In both cases and in contrast to classification, no labels are needed here. There are different approaches to recommend if the Data is in a dynamic stream or static, if the edges have attributes or the documents have connections (Aggarwal and Wang, 2011).

4.3 The Algorithms

In this section, Lai and To (2015) follow with the data-transformation like in figure 1. Because we have no specific task to transform the data for, we want to make suggestions for some specific tasks and recommend the suitable algorithm.

4.3.1 Keyword-Search

For keyword search in social networks, algorithms for graph-data may be the most suitable. But not just so, they also suit for relational- and tree-structured databases as these can be seen as a special case for graphical structured databases (Aggarwal and Wang, 2011). The BANKS-algorithm is easy to use and can be applied without knowledge about the moderately large databases (Bhalotia et al., 2002). But we want to recommend the BLINKS-algorithm as an optimization of BANKS and taking bidirectional search to account as well (He et al., 2007). BLINKS is an algorithm that uses an index structure to search over the data. It has two fundamental steps, the first one is to define a list which contain the Keywords. After the that, the algorithm computes the shortest distance from every node to the keyword and the elements from this list are organised according to the shortest distance. The second step is to calculate the shortest graph distance between the query and the nodes. All this information is organised in a hash table and can be used to avoid unnecessary research over the data set. This algorithm can detect the roots efficiently because it has a big picture of all the connections and structure of the data set, element that is the main advantage that this algorithm offers.

4.3.2 Classification

The approach of classification is fairly different from keyword search before. In many cases, there is a prediction to make, whether the observation belongs to a group or not (is this email urgent or not, is it maybe spam?), the outcome is somehow qualitative (James et al., 2013). The algorithms we suggest are the Email-Classifier from Carvallho and Cohen (2005) that can be used to structure big number of Emails. Another algorithm we can suggest is the Bayes classifier. It is based on the classification of the content of the document given that the neighbours are known. Thus, it explores all classes of the neighbours and then classifies the content in predefined classes on the basic of the probability distribution. Therefore, the category so the probability distribution is maximum receives the document (Chakrabarti et al., 1998). We also have the iterative relaxation algorithm that used when some or none of the neighbours are unknown (ib., 1998). It first classifies all neighbours and then iteratively classifies the content document from the classes of the neighbours.

4.3.3 Clustering

Clustering is the approach to subdivide a group of observations that are similar in each cluster and sufficient different among the clusters (James et al., 2013). Here the choice to make is, if the text mining shall be enhanced by the information from the structure or the network mining enhanced by the attributes which also can be textual data. In the first case, we suggest the iTopic-Modelling form Sun et al. (2009). Here the classical topic modelling is combined with the probability of being connected from the structural data (ib.). In the case that the text is treated as set of attributes to get a more accurate clustering we can recommend the model-based clustering where the attributes help to transfer the observations to a Euclidean space and then clusters them by their distance.

groupvar	textvar
The Guardian	3:29 am Man Steps Onto the Moon
New York Times	Men Walk on Moon - Austronauts Land on Plain, Collect Rocks, Plant Flag
Boston Globe	Man Walks on Moon
Houston Chronicle	Armstrong and Aldrich Take One Small Step for Man on the Moon
Washington Post	The Eagle Has Landed Two Men Walk on the Moon
Chicago Tribune	Giant Leap for Mankind
Los Angeles Times	Walk on Moon That's One Small Step for Man, One Giant Leap for Mankind

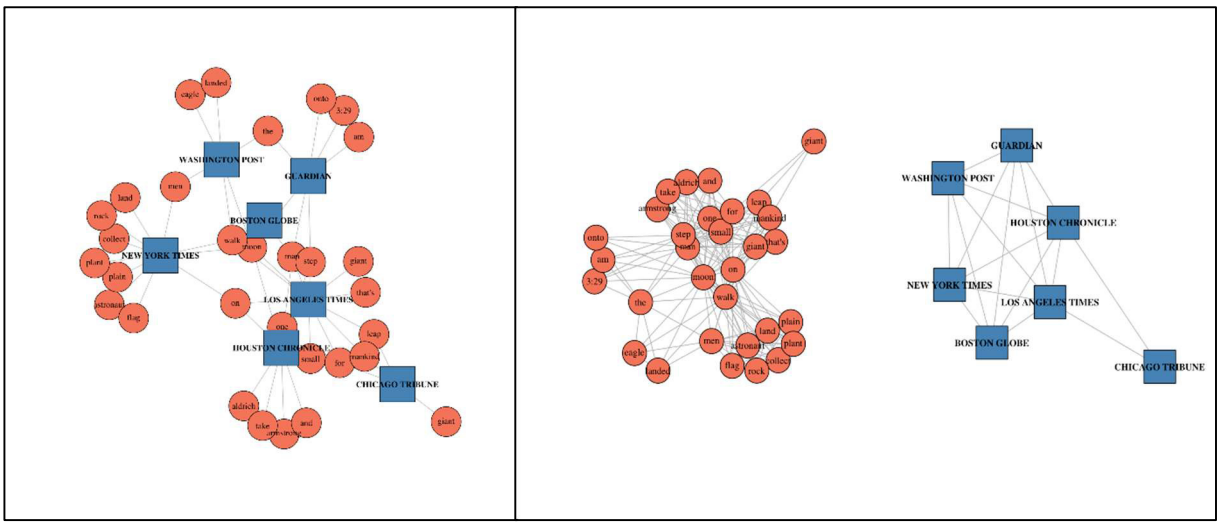

Figure 2: Assembled pictures from Chris Bails Moon landing headlines example for the textnets-package. Source: Chris Bail (2017)

4.3.4 The Outcome

For the Outcome we come back to the Framework from Lai and To (2015). Here the domain-knowledge from the investigator is needed to summarize the results and identify implications and limitations (ib.). The need of domain-knowledge is especially needed for unsupervised learning techniques like clustering (James et al., 2013). When doing a keyword search the result maybe is needed to enhance the company's visibility in the net. For clustering one will probably get segmented customer.

5 Application

To apply the previously discussed Framework to a real topic. We first searched for tools, that take both to account, graph- and textual-information. Unfortunately, there is no such a tool yet despite to the encountered advantages. Therefore, we had to adopt an algorithm in R-Studio and had to do some manually calculations. In particular we used the textnets-package form Bail (2017) on GitHub and Excel because none of the Packages we found so far worked in an appropriate way.

Because the topic of Data-, Text- and Graph-mining is of especial interest we wanted to see, which Keywords are closely related to them in the Twitter-community. When doing our research for the topic, we encountered the problem that there is a wide field already, with work in different directions that are more or less related to Text-mining in Graph-data. The Question is whether there are related fields we need to study or take in account when working on the field.

To achieve the scope, we take the Twitter-network as data source for the relatively easy access via the API and with the rtweets-package in R. Other resources could be for instance the Blog Stack-Overflow, where mostly practical issues are discussed, and the users help each other. Nevertheless, we used Twitter for its easy access. About the sample size as suggested in the original framework we do not worry that it is too big. Even if we can get up to 18,000 tweets in one session with the rtweets-package search_tweets()-function we only got approximately 4500 due to the sparse communication on twitter about the Keywords "Datamining", "Textmining" and "Graphmining".

The Dataset we gathered had 90 variables we did not need at all up to the user_id and the raw text of the tweets. Here we applied some of the steps explained in section 4.2. with the PrepText()-function. These steps were the tokenization, stop-word-removal, and POS-tagging. Because of computational issues we saved this document in Excel and made an English dictionary with the frequency of the single words over all documents. With this dictionary we filtered out the tokens that occur less than eight times, which seemed to be suitable. The term-frequency would not be appropriate because of the short length of

Top tokens by Categories

	Dictionary	Betweeness	Closeness
1	machinelearning	new	abdsc
2	iot	like	versus
3	deeplearn	python	know
4	serverless	datascience statistic	bigdata
5	iiot	yelp yellowpage	datascience statistic
6	tensorflow	free	accelerant
7	python rstat	know	iiot
8	reactj	amp	top
9	nlp	neuralnetwork	python
10	nosql	machinelearning	new

Table 1: Top tokens by the three categories occurrence (dictionary), betweenness-centrality and closeness-centrality. Source: downloaded Twitter-dataset, own calculations in R.

tweets. As you may know the Tweet as a microblog is only 140 characters long and we interpret the occurrence of a word more than one time rather to be spam than meaningful content. After we had pruned the set of tokens, we also had some words like “datamin” instead of “datamining” and “machin-learn” instead of “machinelearning” because the pre-processor lacks a proper stemmer which forced us to make this by hand. Following Hotho and Nürnberger (2005) we also filtered out the tokens from the search-query and the token “via” because it tags when you get a tweet that has its origin at another user.

With the pre-processed text we now made the network with the CreateTextnet()-function. This allows us to get the connections between the words and we give a brief description here. As you can see in figure 2 in the upper table there are the headlines from some newspapers the day after the moon landing. These can be also assembled as a graph as you can see in the lower left graph. Here the quadratic nodes are the newspapers, and the round nodes are the words occurred in the headlines. In the two graphs on lower right side of figure 2 you can see the assembling, when in the right case the cooccurred words are the edges or, as we used it, the author makes the edges. In our next step we graph the data and give us the relationships visualized as in figure 3. With the assembling of the data as graph, there are also opportunities to measure the importance of words with betweenness- and closeness-centrality. The three highest betweenness-centralities have the tokens “new”, “like” and “python” (table 1). The highest closeness-centrality have the tokens “abdsc”, “versus” and “know” (table 1). The closeness-centrality is

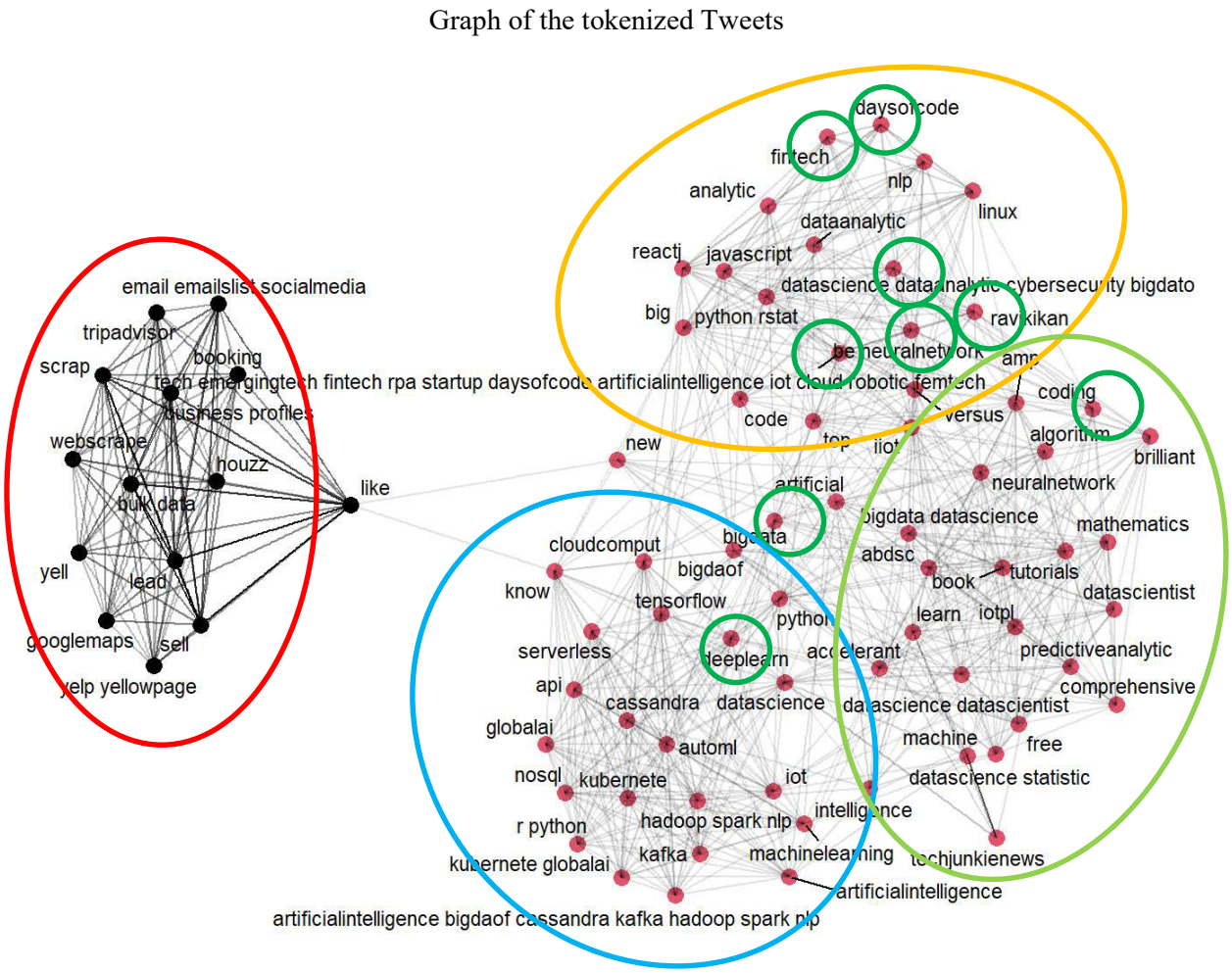

Figure 3: *Graph of the occurred tokens connected by the users. Dataset from Twitter, search query: Datamining or Textmining or Graphmining. Own calculations in R-studio. The-Figure without coloured circles in the appendix.*

a measurement for the efficiency of a vertex as it is larger when the average distance from this node is the shorter to all other nodes (Okamoto et al., 2008). If nodes were to be the users, centrality-measurements would indicate the ones with influence, here the nodes are words and so, higher influence maybe is not the best definition but maybe these words are used together in most contexts and so give us the most requested topics. So maybe the users want to "know" about the "abdsc" or have a comparison ("versus"), "datascience statistic" or "bigdata". Here further analysis would be needed, and one problem comes up. Because it is a microblogging network there occur many abbreviations, we do not know the meaning. This can be problematic as for example "abdsc" seems to be a popular abbreviation, but we didn't find a conclusive meaning.

After all these pre-processing we look for similarities in the words to cluster them for. In figure 3 you see a red and black cluster. The bigger red cluster contains most tokens, and it is looking like one can divide this cluster in three subcluster by sight. The bottom left subcluster with keywords like "Cassandra", "Kafka", "automl", "iot", "kubernete", "api" seem to address the practical computer science. The bottom right subcluster however seems to address the mathematical, statistical data science and theoretical computer science and background with the corresponding keywords. The upper subcluster, however has a special issue, because its closely related to the right twitter-post from figure 4. All green circled keywords are part of the tweet from Ravi Kikan, who was mentioned so often that his name occurres in the graph ("ravikikan"). The post about his new book appears about 120 times in the dataset.

The black cluster seems not to be related to the topic very well. Here tokens like "tripadvisor", "yelp yellowpage" and "booking" seem very common and it is probably a spam-tweet. And by looking at the tweets, there is really such a tweet that's hashtags and content all occurs in this cluster as you can see in figure 4 on the left side and the red circled tokens in figure 3 which are all part of that tweet, that also occurs 147 times in our dataset. And one thing is also salient. While in the red cluster all edges are quite transparent, they appear opaquer in the black cluster. Maybe here lies a pattern to identify such unwanted SPAM-Tweets.

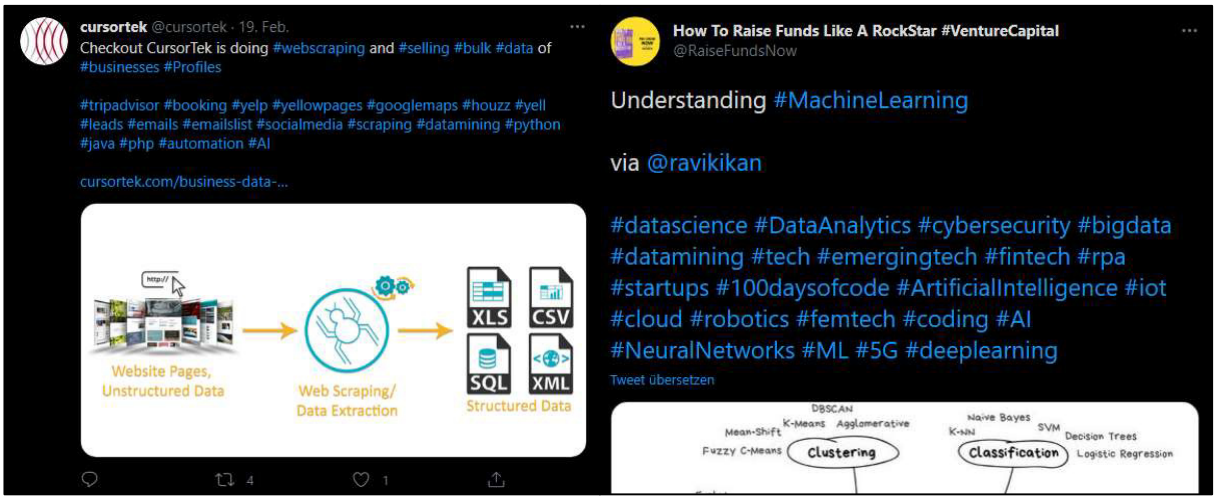

Figure 4: Left: supposed SPAM-Tweet, source: Twitter, 2021a. Right: supposedly from the community shared tweet, source: Twitter, 2021b

6 Conclusion

The expansion of the web and social network technologies has led to a tremendous amount of unstructured data, which may also be expressed in the form of linked networks. Thus, text mining is an important tool for the exploration and analysis of its data. Therefore, we have presented in this paper different algorithms in text mining that combine both text and links structure in order to greatly improve the effectiveness of the application about these data for a wide variety of problems such as keyword

search, classification, and clustering. Based on the approach from Lai and To (2015), where the analysis is follows a four-step-model namely the definition of goal, the data collection, the data transformation and algorithm application, and interpretation of results, we have applied the framework from the social network twitter data. From the search query “Data Mining”, “Text Mining” or “Graph Mining”, we have collected about 4500 tweets. Thus, the purpose our framework was to see, which Keywords are closely related to them in the Twitter-community, so we used the connection of the words in a graph to cluster for topics. After pre-processing and tokenization process, we obtained a network with two types of cluster, so each contains the nodes of word in the same context, and which are connected through the links (figure 3 and Appendix A). As result of our research the main problems that we identified were related to the process of cleaning the data set, due to the heterogenous and dirty language, the occurrence of abbreviations which we do not know the meaning and keywords that have the same meaning but are unrelated as spam messages. However, the cleaning dictionary seemed to be incomplete, and despite the theoretical advances taking graph information and text information together we found no tool that considers them both. On the other side, we already had quite high computational time for the small graph. Our approach seems not suitable because we could not really detect rally new knowledge. But maybe the found pattern for SPAM-Tweets initiates future research.

Reference

((2002). Proceedings 18th International Conference on Data Engineering, 18th International Conference on Data Engineering, San Jose, CA, USA, 26 Feb.-1 March 2002. IEEE Comput. Soc.

(2006). 2006 International Conference on Computing and Informatics. Piscataway, I E E E.

(2009 - 2009). 2009 Ninth IEEE International Conference on Data Mining, 2009 Ninth IEEE International Conference on Data Mining (ICDM), Miami Beach, FL, USA, 06.12.2009 - 09.12.2009. IEEE.

(2011). Social Network Data Analytics. Springer, Boston, MA.

Aggarwal, Charu C. (Ed.) (2011). Social Network Data Analytics. Boston, MA, Springer US.

Aggarwal, Charu C./Wang, Haixun (2011). Text Mining in Social Networks. In: Charu C. Aggarwal (Ed.). Social Network Data Analytics. Boston, MA, Springer US, 353–378.

Bail, C. (2017). cbail/textnets. Available online at https://github.com/cbail/textnets (accessed 2/26/2021).

Bhalotia, G./Hulgeri, A./Nakhe, C./Chakrabarti, S./Sudarshan, S. (2002). Keyword searching and browsing in databases using BANKS. In: Proceedings 18th International Conference on Data Engineering, 18th International Conference on Data Engineering, San Jose, CA, USA, 26 Feb.-1 March 2002. IEEE Comput. Soc, 431–440.

Borgatti, Stephen P./Everett, Martin G./Johnson, Jeffrey C. (2018). Analyzing social networks. 2nd ed. New York, SAGE.

Boyd, Danah M./Ellison, Nicole B. (2007). Social Network Sites: Definition, History, and Scholarship. Journal of Computer-Mediated Communication 13 (1), 210–230. https://doi.org/10.1111/j.1083-6101.2007.00393.x.

Carvalho, Vitor R./Cohen, William W. (2005). On the collective classification of email "speech acts". In: Gary Marchionini (Ed.). SIGIR 2005. Proceedings of the Twenty-Eighth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 2005, Salvador, Brazil, the 28th annual international ACM SIGIR conference, Salvador, Brazil, 8/15/2005 - 8/19/2005. New York, N.Y., Association for Computing Machinery, 345.

Chakrabarti, Soumen/Dom, Byron/Indyk, Piotr (1998). Enhanced hypertext categorization using hyperlinks. ACM SIGMOD Record 27 (2), 307–318. https://doi.org/10.1145/276305.276332.

Cohen, W., Carvalho, V., & Mitchell, T. (2004). Learning to classify email into “speech acts”. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 309–316. Available online at https://www.aclweb.org/anthology/w04-3240.pdf.

Debortoli, Stefan/Müller, Oliver/Junglas, Iris/vom Brocke, Jan (2016). Text Mining for Information Systems Researchers: An Annotated Topic Modeling Tutorial. Communications of the Association for Information Systems 39, 110–135. https://doi.org/10.17705/1CAIS.03907.

Hand, David J./Adams, Niall M. (2014). Data Mining. In: N. Balakrishnan/Theodore Colton/Brian Everitt et al. (Eds.). Wiley StatsRef: Statistics Reference Online. Chichester, UK, John Wiley & Sons, Ltd, 1–7.

Handcock, Mark S./Raftery, Adrian E./Tantrum, Jeremy M. (2007). Model-based clustering for social networks. Journal of the Royal Statistical Society: Series A (Statistics in Society) 170 (2), 301–354. https://doi.org/10.1111/j.1467-985X.2007.00471.x.

He, Hao/Wang, Haixun/Yang, Jun/Yu, Philip S. (2007). BLINKS. In: Lizhu Zhou/Tok Wang Ling/Beng Chin Ooi (Eds.). Proceedings of the 2007 ACM SIGMOD international conference on Management of data - SIGMOD '07, the 2007 ACM SIGMOD international conference, Beijing, China,

11.06.2007 - 14.06.2007. New York, New York, USA, ACM Press, 305.

Hearst, Marti (2003). What is Text Mining? SIMS UC Berkley. https://www.jaist.ac.jp/~bao/MOT-Ishikawa/FurtherReadingNo1.pdf. Available online at https://www.jaist.ac.jp/~bao/MOT-Ishikawa/FurtherReadingNo1.pdf (accessed 2/27/2021).

Hotho, Andreas, Andreas Nürnberger (2005). A brief survey of text mining. LdV Forum 2005 (20), 19–62. Available online at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.447.4161&rep=rep1&type=pdf.

Irfan, Rizwana/King, Christine K./Grages, Daniel/Ewen, Sam/Khan, Samee U./Madani, Sajjad A./Kolodziej, Joanna/Wang, Lizhe/Chen, Dan/Rayes, Ammar/Tziritas, Nikolaos/Xu, Cheng-Zhong/Zomaya, Albert Y./Alzahrani, Ahmed Saeed/Li, Hongxiang (2015). A survey on text mining in social networks. The Knowledge Engineering Review 30 (2), 157–170. https://doi.org/10.1017/S0269888914000277.

James, Gareth/Witten, Daniela/Hastie, Trevor/Tibshirani, Robert (2013). An Introduction to Statistical Learning. New York, NY, Springer New York.

Lai, Linda S.L. /W.M. To (2015). Content analysis of social media: A grounded theory approach. Journal of Electronic Commerce Research 16 (2), 138–152. Available online at https://www.researchgate.net/profile/wai_ming_to/publication/276304592_content_analysis_of_social_media_a_grounded_theory_approach.

Ling, Huan Su/Bali, Ranaivo/Salam, Rosalina Abdul (2006). Emotion detection using keywords spotting and semantic network IEEE ICOCI 2006. In: 2006 International Conference on Computing and Informatics. Piscataway, I E E E.

Milroy, Lesley/Llamas, Carmen (2013). Social Networks. In: J. K. Chambers/Natalie Schilling (Eds.). The Handbook of Language Variation and Change. Oxford, UK, John Wiley & Sons, Inc, 407–427.

Okamoto, Kazuya/Chen, Wei/Li, Xiang-Yang (2008). Ranking of Closeness Centrality for Large-Scale Social Networks. In: Franco P. Preparata/Xiaodong Wu/Jianping Yin (Eds.). Frontiers in Algorithmics. Second Annual International Workshop, FAW 2008, Changsha, China, June 19-21, 2008, Proceeedings, Berlin, Heidelberg, 2008. Berlin, Heidelberg, Springer-Verlag Berlin Heidelberg, 186–195.

Statista (2021a). Internet users in the world 2020 | Statista. Available online at https://www.statista.com/statistics/617136/digital-population-worldwide/ (accessed 2/27/2021).

Statista (2021b). Most used social media 2020 | Statista. Available online at https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (accessed 2/27/2021).

Strapparava, Carlo/Gözde Özbal (2010). The color of emotions in texts. Proceedings of the 2nd Workshop on the Cognitive Aspects of the Lexicon 2010, 28–32. Available online at https://www.aclweb.org/anthology/w10-3405.pdf.

Sun, Yizhou/Han, Jiawei/Gao, Jing/Yu, Yintao (2009). iTopicModel: Information Network-Integrated Topic Modeling. In: 2009 Ninth IEEE International Conference on Data Mining, 2009 Ninth IEEE International Conference on Data Mining (ICDM), Miami Beach, FL, USA, 06.12.2009 - 09.12.2009. IEEE, 493–502.

Twitter (2021a). cursortek auf Twitter: "Checkout CursorTek is doing #webscraping and #selling #bulk #data of #businesses #Profiles #tripadvisor #booking #yelp #yellowpages #googlemaps #houzz #yell #leads #emails #emailslist #socialmedia #scraping #datamining #python #java #php #automation #AI https://t.co/L8HY1PV7AZ https://t.co/6uaLsWMLob" / Twitter. Available online at https://twitter.com/cursortek/status/1362884673072939020 (accessed 2/27/2021).

Twitter (2021b). How To Raise Funds Like A RockStar #VentureCapital auf Twitter: "Understanding #MachineLearning via @ravikikan #datascience #DataAnalytics #cybersecurity #bigdata #datamining #tech #emergingtech #fintech #rpa #startups #100daysofcode #ArtificialIntelligence #iot #cloud #robotics #femtech #coding #AI #NeuralNetworks #ML #5G #deeplearning https://t.co/OF7xaDyWgl" / Twitter. Available online at https://twitter.com/RaiseFundsNow/status/1363329741017681920 (accessed 2/27/2021).

Witten, Ian H. (2004). Text Mining. University of Waikato, Hamilton New Zealand. Available online at http://www.cs.waikato.ac.nz/~ihw/papers/04-ihw-textmining.pdf.

Xu, Yu/Papakonstantinou, Yannis (2005). Efficient keyword search for smallest LCAs in XML databases. In: Fatma Ozcan (Ed.). Proceedings of the 2005 ACM SIGMOD international conference on Management of data - SIGMOD '05, the 2005 ACM SIGMOD international conference, Baltimore, Maryland, 14.06.2005 - 16.06.2005. New York, New York, USA, ACM Press, 527.

Zhang, Tong/Popescul, Alexandrin/Dom, Byron (2006). Linear prediction models with graph regularization for web-page categorization. In: Lyle Ungar (Ed.). KDD 2006. Proceedings of the twelfth ACM SIGKDD international conference on knowledge discovery and data mining, New York, New York, USA. New York, New York, USA, ACM.

Zhou, Yang/Cheng, Hong/Yu, Jeffrey Xu (2009). Graph clustering based on structural/attribute similarities. Proceedings of the VLDB Endowment 2 (1), 718–729. https://doi.org/10.14778/1687627.1687709.

Appendix

Appendix A – The clean word-network graph