Recommendation of Query Terms for Colloquial Texts in Forensic Text Analysis

Jenny Felser[1] Jian Xi[2] Christoph Demus[3] Dirk Labudde[4] Michael Spranger[5]

Abstract: Textual social media content and short messages have gained in importance as evidence in criminal investigations. Yet, the large number of textual data poses a great challenge for investigators. Even though text retrieval systems can assist in finding evidential messages, the success of the search still depends on entering appropriate search terms. However, for colloquial texts these are difficult to determine because one cannot be sure about what terms are used in the texts and might be of interest. Therefore, the aim is to develop a method that recommends keywords and searches phrases based on the underlying data. A particular challenge here is that the appropriate search terms are often non-obvious words that are not expected to be found in the data, but are particularly relevant. In total, four methods were evaluated for extracting and suggesting the most relevant terms and phrases using a real-life dataset. The best results were obtained with topic modeling considering syntagmatic relations.

Keywords: query terms; colloquial texts; forensic text analysis; text retrieval

1 Introduction

In contemporary digital age, the communication via messenger services such as WhatsApp increases significantly. Consequently, the analysis of chat messages of mobile devices also gains in importance for forensic purposes. However, the increasing number of messages develops to a great challenge for the investigators. Although text retrieval systems can assist in finding evidential information in the numerous messages, the success of the search largely depends on entering suitable query terms. Yet, these are difficult to find, especially in the case of colloquial texts. Mostly, it is not known which terms are relevant for the specific case. Furthermore, one cannot be sure if particular words are even used in the chat messages. Therefore, the aim of this paper is to develop a method for recommending search words and phrases based on the underlying dataset. For this purpose, different methods are analysed

[1] University of Applied Sciences Mittweida, Technikumplatz 17, 09648 Mittweida, Germany jfelser@hs-mittweida.de

[2] University of Applied Sciences Mittweida, Technikumplatz 17, 09648 Mittweida, Germany xi@hs-mittweida.de

[3] University of Applied Sciences Mittweida, Technikumplatz 17, 09648 Mittweida, Germany cdemus@hs-mittweida.de

[4] University of Applied Sciences Mittweida, Technikumplatz 17, 09648 Mittweida, Germany labudde@hs-mittweida.de

[5] University of Applied Sciences Mittweida, Technikumplatz 17, 09648 Mittweida, Germany spranger@hs-mittweida.de

doi:10.18420/inf2022_02

and compared, which have in common that they do not require any prior knowledge of the data.

Thereby, a particular challenge is that the most relevant terms for an investigation are often non-obvious and unfamiliar words. For instance, if the case deals with the financial support of a terrorist group, the investigator will assume that the dataset contains words such as “money”, “transfer” or “account”. However, it would be much more interesting if the chats mentioned names of terrorist organisations, brands of weapons that have to be bought for a terrorist attack, or places connected to the terrorist group.

One application of the proposed approaches consists in the development of a Term Tree Recommender (TTR) for the Mobile Network Analyzer (MoNA), a forensic tool for analysing chat messages [1], where the so-called Term Tree identifies messages containing relevant terms for a certain case. The purpose of the TTR is to suggest words that can subsequently be imported to the Term Tree. A time-limited trial version of MoNA including the implemented TTR is available for download [6]. In addition, the presented methods can also be applied in general to find important terms and phrases in texts and particularly in short messages.

The paper is organized as follows: Related work is presented in Sect. 2. After it, Sect. 3 shows the details of the proposed approaches. Afterwards, the experimental results are presented in Sect. 4 and discussed in Sect. 5. Finally, Sect. 6 concludes the paper.

2 Related Work

There have been only few studies to address the issue of recommending search terms in forensic texts. Besides, those studies have focused on the analysis of e-mails and have not considered short messages and chats. For instance, Koven et al. [2] presented a search term recommender for InVEST, a forensic tool for finding evidence in large e-mail datasets, where the proposed approach assumes that the investigator has already searched for a word in the data. Subsequently, the recommender suggests the most important terms in the result set using a variant of TF-IDF algorithm. Then, the investigator can inspect the suggested terms in order to discover more relevant e-mails. The proposed methodology is not applicable to our recommender system, since the search procedure within is fully conducted from scratch with its assistance, whereas the recommender developed by Koven et al. [2] is part of an iterative search process. Furthermore, Teng [3] pointed out that TF-IDF tends to extract words with high frequency and is thus inappropriate to suggest specific and uncommon words.

Addressing the second problem, Joshi and Motwani [4] introduced the recommender TermsNet with the purpose to suggest non-obvious, but nevertheless relevant terms in the field of sponsored search. The authors proposed a graph-based technique, which explores semantic associations between words by considering the syntagmatic relations between

[6] `https://www.hs-mittweida.de/spranger/`

concepts. However, their approach requires the user to specify a few relevant seed words, whereas in this paper, the focus is on the case that no relevant terms of the data are known.

Moreover, a large number of existing studies have examined methods for suggesting query terms in the context of web search. However, most of those approaches rely on query logs [5][6][7] and cannot be used for recommending important words based on a certain underlying dataset. A more relevant work in this field was conducted by Kubek and Unger [8], where the recommender system DocAnalyser suggests the most relevant words of a previously visited web page by a graph-based method that is based on a modified HITS-algorithm. Subsequently, the user can enter those words as query to find similar web documents. Yet, Kubek and Unger [8] emphasized that their approach achieves particularly good results, when the underlying dataset deals with one homogeneous topic. Otherwise, the recommender will mainly suggest words that are related to the topic with the highest coverage in the data [9]. This is particularly problematic for our recommender system because the case-related topic only represents a small portion of the conversations. Furthermore, the proposed method addresses term recommendations merely based on a single web page. Therefore, it is not necessarily appropriate for handling large data.

To overcome these limitations, instead of using HITS-algorithm, a further graph-based algorithm, namely “Rapid automatic keyword extraction” (RAKE) [10], is reported by Wang et al. [9] for suggesting the most important phrases of online news articles. Even though this approach yields promising results, it cannot be guaranteed that it also performs well on messages. In contrast to news articles, short messages and chats contain numerous typos, grammatical mistakes, acronyms [11] and colloquial expressions. Therefore, it constitutes a particular challenge for traditional algorithms and methods.

Keerthana [12] presented a recommender that uses topic modeling for suggesting queries based on conversation fragments. The basic idea is to recommend the terms with the highest probability in the most relevant topic of the dataset. As topic modeling has proven to be successful in revealing latent semantics [13], this method has also potential to recommend specific and unexpected words. However, previous studies [14][15] have shown that conventional topic models do not perform well on short text collections due to their high sparsity. To avoid this problem, we make use of a variant of topic modeling incorporating syntagmatic relations, as proposed by [16].

Consequently, previous studies have either failed to recommend particularly specific terms or have been constrained to the case that the user already knows some relevant words. To fill this gap, the objective of this paper is to develop a new method for term recommendation, which, in addition, is also able to cope with the special linguistic and grammatical characteristics of forensic short messages.

3 Methods

To find the most appropriate approach for recommending search terms, we compared the following four algorithms: differential analysis, document clustering, standard topic modeling and a variant of topic modeling considering syntagmatic relations. All methods were evaluated using a dataset of a real case about the financial support of a terrorist group, which includes approximately 180.000 messages of the messenger service WhatsApp stored on the mobile phone of one case-related person. In order to assess the quality of the different approaches, the recommended terms were compared with the words of an existing Term Tree that had previously been created by an investigator for the analysis of the corresponding case using the software MoNA. The Term Tree contained the following 25 words that the investigator had deemed relevant to the case: *Geld (money), Schwester (sister), Nalan, Hawala, Nierderlande (Netherlands), Buch (book), Bücher (books), scan, Seiten (pages), Kapitel (chapter), Wörter (words), Bieber (beaver), Layth, Verleger (publisher), money, Druck (print), angekommen (arrived), Dolar (dollar), €, Euro (euro), $, Western (western), Union (union), PayPal* and *Dollar (dollar)*. A method was considered as suitable if it recommended many of these or similar terms such as synonyms or thematically related words. The procedure from the preprocessing of the dataset to the evaluation of the applied approaches is shown in Fig. 1.

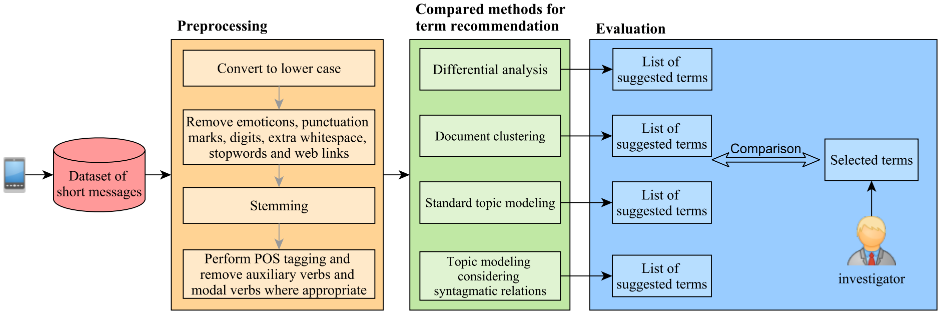
Preprocessing
Compared methods for term recommendation
Evaluation
Dataset of short messages
Convert to lower case
Remove emoticons, punctuation marks, digits, extra whitespace, stopwords and web links
Stemming
Perform POS tagging and remove auxiliary verbs and modal verbs where appropriate
Differential analysis
Document clustering
Standard topic modeling
Topic modeling considering syntagmatic relations
List of suggested terms
List of suggested terms
List of suggested terms
List of suggested terms
Comparison
Selected terms
investigator

Fig. 1: Procedure applied for the comparison of methods for the recommendation of query terms.

3.1 Preprocessing

A number of preprocessing steps were performed before the comparison of the mentioned algorithms. Firstly, the messages were tokenised in terms with lower case. Emoticons, punctuation marks, digits and extra whitespace of the messages were removed. In addition, German and Turkish stopwords were removed using the stopword lists provided by Diaz [17] and all web links were also removed. Stemming was used for all methods in order to reduce the words to their root or stem forms. At recommendation step, the matched words, which are stemmed, were completed in order to recommend the original, correct terms.

Regarding topic modeling considering syntagmatic relations, all parts of speech (POS) were used. Adjectives were additionally included for differential analysis, clustering and standard topic modeling. TreeTagger was used for POS tagging [18]. Moreover, auxiliary verbs and modal verbs were removed for clustering and both variants of topic modeling. Since the dataset consists mainly of German messages, only the German TreeTagger model was used for POS tagging and the removal of auxiliary and modal verbs was also limited to German words.

3.2 Differential analysis

A first idea was to apply differential analysis [19], which is an approach of statistical text analysis [8]. Differential analysis compares the frequencies of the words in the messages with their frequencies in the so-called reference corpus, a text corpus in common German language. In this work, the corpus “Deutscher Wortschatz” [20] of the University of Leipzig was chosen as reference corpus, which predominantly consists of online newspapers. Differential analysis identifies all words that appear in the chat messages significantly more often than in the reference corpus [19]. As these words are considered as discriminating terms of the messages, they are also appropriate search term recommendations. The extracted words were sorted based on the difference of their frequencies to the reference corpus.

3.3 Document clustering

A further approach was to cluster the messages using an agglomerative hierarchical cluster analysis as explained by Murtagh and Contreras [21]. However, the high sparsity of short messages is a well-known problem with hierarchical clustering [22]. One way to tackle this problem is to form pseudo-documents by aggregating several messages [22], whereby in this work, a pseudo-document consisted of all messages written in one chat at one day. Initially, four clusters were created, of which the investigator must select the case-related one. Subsequently, only the messages in the respective cluster were considered for extracting relevant words, whereas irrelevant small talk conversations were excluded in order to reduce noise in the data. The words that occured most frequently in the considered messages were recommended, which is similar to the approach described by Jones, Doane, and Jones [23].

3.4 Standard topic modeling

Another method for explorative data analysis is topic modeling, for which Latent Dirichlet Allocation (LDA) described by Blei, Ng, and Jordan [24] was chosen in this study. The input of the LDA is usually a document-term matrix (DTM) [25]. As with the previous clustering, a document was defined as the amount of all messages that were written in one chat at one day. The number of topics was set at four in accordance with clustering. The output of the LDA needed to recommend terms is the probability distribution over words that characterise each topic [24]. Similar to document clustering, the investigator is presented with all the topics from which he has to select the case-related one. Afterwards, the most probable words of that topic are suggested.

3.5 Topic modeling considering syntagmatic relations

As mentioned before, standard LDA can frequently be improved for short messages by taking syntagmatic relations into account [16]. To do so, the LDA was repeated with a term co-occurrence matrix (TCM) as input, which captures the local context of words by using a skip-gram model [23]. In contrast to the previously described approach, the most appropriate number of topics was estimated based on probabilistic coherence as proposed by Jones, Doane, and Jones [23]. Analogous to standard topic modeling, recommendations of query terms are obtained by suggesting the most probable terms of the topic selected by an investigator.

4 Results

4.1 Differential analysis

In general, differential analysis did not lead to the recommendation of relevant terms. As an example, the first nine terms recommended by difference analysis and their frequencies in the messages are shown in Tab. 1. As can be seen, none of the words of the Term Tree were suggested, but mostly Arabic phrases and colloquial terms.

term	**term frequency**
Allah	1871
Alhamdulillah	939
Shaa	872
Inshallah	844
Subhanallah	792
haha	748
hahaha	666
nochmal	621
amin	547

Tab. 1: The first nine terms suggested by differential analysis.

4.2 Document clustering

It was not possible to identify the case-related cluster because none of the most frequent terms in the respective documents also appeared in the Term Tree. Furthermore, it is remarkable that the clusters vary widely regarding the number of assigned documents. As an example, the five most frequent words and the number of documents respectively pseudo-documents in the four clusters are given in Tab. 2.

cluster	# documents	most frequent terms
1	3831	ok, diggi, digger, moin, hayirli
2	2	yorum, resimlere, çekiyor, yapmis, deyince
3	2	einzelne (individual), Preise (prices), verschiedenen (different), digga, schau (look)
4	1	gucks (look), Bunker (bunker), digga, ersma (first), frohes (happy)

Tab. 2: The five most frequent words of four clusters using hierarchical document clustering.

4.3 Standard topic modeling

The first of four topics calculated by LDA could be identified as the case-related topic, because it is the only one for which terms in the Term Tree appear among the 80 most probable words, namely "Druck" and "PayPal". Moreover, there are seven further terms among these 80 words that are thematically similar to the terms in the Term Tree, for example "Sponsoren" and "fees". Those words are shown with their topic probabilities and their position among the 80 most probable words in Tab. 3.

position	term	probability in topic 1
7	spenden (donate)	0.0016043884
11	Rechnung (bill)	0.0014620923
23	Druck (print)	0.0012842222
43	fees	0.0011419261
54	PayPal	0.0010707781
58	Preise	0.0010352041
64	Belege (receipts)	0.0009996300
71	Gebühren (fees)	0.0009640560
75	Sponsoren (sponsors)	0.0009284820

Tab. 3: Relevant topic calculated by traditional topic modeling.

4.4 Topic modeling considering syntagmatic relations

Estimating the appropriate number of topics by probabilistic coherence resulted in 13 topics. The ninth topic was considered the relevant topic for the case as shown in Tab. 4 because the word “Geld” of the Term Tree and further three thematically related terms, namely “Fees”, “Konto” and “Rechnung”, are among the 50 most probable words of this topic. In addition, we also regarded words such as “Twitter”, “Event” and ”Tipeee” as important for the case because it was apparent from the context that these words were related to the financial support of the terrorist organisation. As outlined in Tab. 4, a total of 15 terms were considered as relevant. Those terms are shown with their position among the 50 most probable terms and their topic probability.

position	term	probability in topic 9
1	Twitter (twitter)	0.017992353
2	Team (team)	0.015728373
4	Event (event)	0.014847936
5	Statement (statement)	0.014093276
8	Tipeee	0.012835509
9	Twitch	0.011955073
15	Discord	0.009439539
18	SWH	0.008433326
19	Fees	0.008181772
24	Partner	0.007678666
25	Geld (money)	0.006546676
26	Konto (account)	0.006420899
39	Betterplace	0.004534249
40	Projekt (project)	0.004408472
45	Subs	0.003779589
49	Rechnung (bill)	0.003402259

Tab. 4: Relevant topic calculated by topic modeling incorporating syntagmatic relations.

4.5 Comparison with state-of-the-art approaches

The results of the proposed approaches were compared with two commonly used methods for the recommendation of query terms, namely TF-IDF and RAKE. The first ten terms suggested by these methods are listed in Tab. 5. As RAKE is only applicable for finding relevant words in single documents and not in the whole dataset, the presented term recommendations refer to each pseudo-document consisting of all messages in one chat. As an example, the recommended terms for the first three chats are presented. As outlined in Tab. 5, none of the terms of Term Tree or similar words are among the words suggested by these methods.

position	suggested query terms by TF-IDF	suggested query terms by RAKE		
		chat 1	**chat 2**	**chat 3**
1	voll	ewig draußen rum	fufu danke fürs	guten appetit danke
2	enişte	absoluter ruhemodus	btw	biste glitch theater
3	güzel	guten hunger	relativ	cool lässt
4	klingt	halle geworfen	sascha	haha geilo
5	digga	leute direkt	wach	iwie ggn
6	nen	liebe messen		paar fotos
7	ding	nähe entspannt		paar vorträge
8	wach	voll happy		schmecken naja
9	hause	nacht eistüte		würd sagen
10		draußen		

Tab. 5: The first ten terms suggested by two common methods for recommending query terms.

5 Discussion

Both standard topic modeling and topic modeling considering syntagmatic relations suggested terms that were indeed considered relevant by an investigator. In contrast, other current approaches such as TF-IDF or RAKE were not able to identify the words that were actually important, but mainly recommended slang terms and empty phrases. This result highlights the suitability of the proposed approaches for chat messages.

However, standard topic modeling fails to recommend specific and non-obvious words. For instance, it is not surprising that the terms “fees” and “Gebühren”, as shown in Tab. 3, occur in the context of the topic “money”. In contrast, by incorporating syntagmatic relations, topic modeling is able to recommend specific and unfamiliar words such as the crowdfunding platform “Tipeee”, the association “Streaming with Heart” (SWH) and the donation platform “Betterplace”. Although these words have not been previously added by the investigator to the Term Tree, it could be concluded from their context in the chat messages that those words are indeed case-related. Consequently, this approach succeeds in revealing terms that the investigator would not otherwise have considered as search terms. Therefore, this method is also considered the most useful approach for recommending search terms.

Moreover, the results demonstrate that differential analysis is unsuitable for suggesting appropriate query terms based on the chat messages. A possible explanation for this result is that differential analysis does not only reveal differences of content between the messages and the reference corpus, but also stylistic differences [26]. Whereas the reference corpus is written in German standard language, the messages in this study are colloquial and contain expressions of youth language and Turkish phrases. As a consequence, the words that exclusively occur in the messages, thus being identified as characteristic terms, are particularly slang terms, anglicisms and Turkish filler words. In addition, these results demonstrate that term frequency is not sufficient as sole criterion for extracting relevant words, since the high-frequency terms, which can be seen in Tab. 1, are not necessarily representative for the analysed text.

With regard to clustering, the inadequate results can be explained by the fact that the outcomes of clustering, especially of hierarchical clustering algorithms, strongly depend on the underlying data [27]. Algorithms for document clustering are based on documents that are linguistically correct [28]. The chat messages differ significantly of such documents.

The main limitation of our study so far is the lack of a more thorough evaluation of the compared approaches. Therefore, a more comprehensive user study is required, in which the relevance of terms to the messages is assessed by multiple annotators rather than just a single investigator. In addition, the methods should be evaluated on more forensic datasets covering different offense domains and multiple languages. Besides, it would be interesting to investigate if topic modeling considering syntagmatic relations also yields promising results for other forensically relevant document types, especially for longer text types such as case records and interrogation transcripts.

Furthermore, the approaches that have been investigated so far are completely unsupervised and assume that no important terms of the data are known. However, the investigator might already know a few relevant words, for example from police interrogations or from the case record. Hence, further studies should examine methods considering those words for recommending query terms. Possibilities include to analyze paradigmatic relations to case-related terms and to apply semantic similarity measures such as Latent Semantic Analysis and Pointwise Mutual Information.

6 Conclusion

Finding evidential information in numerous chat messages that need to be analysed in the investigative process is a challenging task. Therefore, in this paper we presented different approaches to develop a method for recommending query terms for colloquial texts. Overall, topic modeling incorporating syntagmatic relations was the method of the four investigated approaches that was capable of recommending the most terms that had previously been considered by an investigator relevant. A particular advantage of this method is that it is able to suggest non-obvious and specific terms that the investigator would not otherwise have thought to search for in the dataset. Entering these appropriate query terms in text retrieval systems prevents important clues contained in the messages from being overlooked. Furthermore, the use of specific search terms, as opposed to general terms, can also reduce the number of messages returned by those systems, resulting in significant time savings for the investigator.

Nevertheless, further experiments involving more real-life datasets and evaluation of the results by more annotators need to be conducted. Furthermore, a possible improvement is to apply methods that take into account a few known relevant words.

References

[1] Michael Spranger et al. “MoNA: Automated Identification of Evidence in Forensic Short Messages”. In: *International Journal On Advances in Security* 9.1&2 (2016), pp. 14–23.

[2] Jay Koven et al. “InVEST: Intelligent Visual Email Search and Triage”. In: *Digital Investigation* 18.16 (Apr. 2016), pp. 138–148.

[3] Miao Teng. “Using the Ship-Gram Model for Japanese Keyword Extraction Based on News Reports”. In: *Complexity* 21.4 (2021), pp. 1–9.

[4] Amruta Joshi and Rajeev Motwani. “Keyword Generation for Search Engine Advertising”. In: *IEEE International Conference on Data Mining-Workshops*. Vol. 6. IEEE. 2006, pp. 490–496.

[5] Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza. “Query Recommendation using Query Logs in Search Engines”. In: *International conference on extending database technology*. Springer. 2004, pp. 588–596.

[6] Qi He et al. “Web Query Recommendation via Sequential Query Prediction”. In: *IEEE International Conference on Data Engineering*. Vol. 25. IEEE. 2009, pp. 1443–1454.

[7] Qiaozhu Mei, Dengyong Zhou, and Kenneth Church. “Query Suggestion Using Hitting Time”. In: *Proceedings of the ACM conference on Information and knowledge management*. Vol. 8. 17. 2008, pp. 469–478.

[8] Mario Kubek and Herwig Unger. “Search Word Extraction Using Extended PageRank Calculations”. In: *Autonomous Systems: Developments and Trends*. Springer, 2012, pp. 325–337.

[9] Zihuan Wang et al. “A news-topic recommender system based on keywords extraction”. In: *Multimedia Tools and Applications* 77.4 (2018), pp. 4339–4353.

[10] Stuart Rose et al. “Text Mining: Applications and Theory”. In: Citeseer, 2010. Chap. Automatic Keyword Extraction from Individual Documents, pp. 1–20.

[11] Ville H. Tuulos and Henry Tirri. “Combining Topic Models and Social Networks for Chat Data Mining”. In: *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI’04)*. IEEE. 2004, pp. 206–213.

[12] S. Keerthana. “Recommended Search of Documents from Conversation with Relevant Keywords Using TextSimilarity”. In: *Journal of Network Communications and Emerging Technologies (JNCET)* 7.2 (Feb. 2017). ISSN: 2395-5317.

[13] Yuzhong Chen et al. “Popular topic detection in Chinese micro-blog based on the modified LDA model”. In: *Web Information System and Application Conference (WISA)*. Vol. 12. IEEE. 2015, pp. 37–42.

[14] Ou Jin et al. “Transferring Topical Knowledge from Auxiliary Long Texts for Short Text Clustering”. In: *Proceedings of the ACM international conference on Information and knowledge management*. Vol. 20. 11. Oct. 2011, pp. 775–784.

[15] Yuheng Hu et al. “What Were the Tweets About? Topical Associations between Public Events and Twitter Feeds”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 6. 1. 2012, pp. 154–161.

[16] Xiaohui Yan et al. “A Biterm Topic Model for Short Texts”. In: *Proceedings of the 22nd international conference on World Wide Web*. Vol. 13. 22. 2013, pp. 1445–1456.

[17] Gene Diaz. *Stopwords ISO*. `https://github.com/stopwords-iso/stopwords-iso`. Sept. 2020.

[18] Helmut Schmid. *TreeTagger-a part-of-speech tagger for many languages*. Tech. rep. Ludwig-Maximilians-Universität Munich, 1994.

[19] Gerhard Heyer, Uwe Quasthoff, and Thomas Wittig. *Text Mining: Wissensrohstoff Text - Konzepte, Algorithmen, Ergebnisse*. Bochum: W3L-Verlag, 2006. ISBN: 978-3-937-13730-8.

[20] Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. “Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages”. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*. Vol. 12. 8. 2012, pp. 759–765.

[21] Fionn Murtagh and Pedro Contreras. “Algorithms for hierarchical clustering: an overview”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2.1 (Jan. 2012), pp. 86–97.

[22] Oren Tsur, Adi Littman, and Ari Rappoport. “Efficient Clustering of Short Messages into General Domains”. In: *Proceedings of the International AAAI Conference on Weblogs and Social Media*. Vol. 7. 2013, pp. 621–630.

[23] Thomas Jones, William Doane, and Thomas Jones. *Package textmineR*. `https://cran.microsoft.com/snapshot/2022-04-10/web/packages/textmineR/textmineR.pdf`. June 2021.

[24] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. “Latent Dirichlet Allocation”. In: *Journal of Machine Learning Research* 3.1 (Jan. 2003), pp. 993–1022.

[25] Kurt Hornik and Bettina Grün. “topicmodels: An R package for fitting topic models”. In: *Journal of statistical software* 40.13 (2011), pp. 1–30.

[26] Uwe Quasthoff. “Korpusbasierte Wörterbucharbeit mit den Daten des Projekts Deutscher Wortschatz”. In: *Linguistik online* 39.3 (June 2009), pp. 151–162.

[27] Moses Charikar, Vaggos Chatziafratis, and Rad Niazadeh. "Hierarchical Clustering better than Average-linkage". In: *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM. 2019, pp. 2291–2304.

[28] Kevin Dela Rosa et al. "Topical Clustering of Tweets". In: *Proceedings of the ACM SIGIR: SWSM* 63.10 (2011).