http://doi.org/10.35668/2520-6524-2022-1-05
UDC 930/004.6

A. BAKLĀNE, Master of Philosophy
V. SAULESPURĒNS, Master of Computer Science

THE APPLICATION OF LATENT DIRICHLET ALLOCATION FOR THE ANALYSIS OF LATVIAN HISTORICAL NEWSPAPERS: OSKARS KALPAKS' CASE STUDY

Abstract. *In the last 20 years, topic modeling and the application of LDA (latent Dirichlet allocation) model in particular has become one of the most commonly used techniques for exploratory analysis and information retrieval from textual sources. Although topic modeling has been used to conduct research in a large number of projects, the technology has not yet become a part of the common standard functionalities of digital historical collections that are curated by the libraries, archives and other memory institutions. Moreover, many common and well researched natural language processing techniques, including topic modeling, have not been sufficiently applied to working with sources of small or low-resource languages, including Latvian. The paper reports the results of the first case study where the LDA methodology has been used to analyze a data set of historical newspapers in Latvian. The corpus of the newspaper Latvian Soldier is used to conduct the analysis, focusing on the performance of the topics related to the first commander of Latvian army Oskars Kalpaks as an example. In the research of digital humanities, the results of the topic modeling have been used and interpreted in several distinct ways depending on the type and genre of the text, e.g., to acquire semantically coherent, trustworthy lists of keywords, or to extract lexical features that do not aid thematic analysis but provide other insights about the usage of language instead. The authors of this paper propose applications that could be most suitable for the analysis of historical newspapers in large digital collections of memory institutions, as well as recount the challenges related to working with textual sources that contain optical recognition errors, problematic segmentation of articles and other issues pertaining to digitized noncontemporary data.*

Keywords: *topic modeling, latent Dirichlet allocation, topic coherence, historical newspapers, natural language processing of Latvian, digital humanities, Oskars Kalpaks.*

INTRODUCTION

The study documented in this paper is a part of broader set of initiatives that are aimed at researching and developing natural language processing and machine learning tasks that could be implemented to create new digital services in the platforms of the National Digital Library of Latvia (NLL) [1]. Among other areas of text mining and information retrieval, the functionalities based on the probabilistic topic modeling are being considered as promising candidates for the new services. The main goal of this case study was testing of the latent Dirichlet allocation (LDA) technique for automatic topic detection to evaluate and determine its usability for analyzing historical textual sources in Latvian. The LDA methodology was selected as one of the most well understood and extensively tested approaches of topic modelling. The method was applied to analyzing a subset of Latvian historical newspapers, which is currently the largest and most popular digital full text collection curated by the NLL.

NLL has been digitizing historical collections of newspapers, books, images, audio and video collections since 1999 [2; 3]. Textual collections in particular have been the focus of mass digitization projects: it is estimated that the collection of digitized newspapers entails more than 80 % of all periodicals published until 1990 and substantial sections of contemporary digitized and digitally born newspapers and magazines [4]. The materials have undergone the process of segmentation and optical recognition; hence, the users have been able to make use of the opportunities provided by the full text search. Following the development of language technologies and current trends in digital humanities research, there is a demand for developing new services that would enable further in-depth analysis of digital documentary sources [5; 6].

The development of digital services in the memory institutions is hindered by specific methodological limitations that accompany the processing and analysis of historical documents. Many challenges are avoided by computational systems which focus on textual data that are created recently and consist of very large balanced datasets

that are stylistically homogenous and standardized; historical data sets, on the contrary, often entail relatively small corpora that lack uniformity and balance [7]. Although processing of most ancient materials may prove to be most difficult, also the extensive and increasingly homogenous textual resources created in the 19th century and in the beginning of 20th century present distinct challenges that need to be accounted for when developing text mining services for digital collections. Spelling variation is one of the main challenges that need to be tackled when analyzing texts from diverse time periods [8].

In addition to the slow implementation of language technologies in the digital libraries of memory institutions, the approbation of technologies is even more lagging in small languages with poor digital language resources. Currently, the greatest success in the field of natural language processing has been achieved in working with large, high-resource languages. It has been often pointed out in the last decade that natural language processing is much more challenging for low-resource or resource-poor languages that have not developed massive annotated corpora and such resources as treebanks and wordnets [9]. According to META-NET estimation in 2012, Latvian was evaluated as providing weak or no support for all text and speech processing tasks [10]. Although many language resources have been developed since, the trend of has been continuing in the current era; for instance, word embeddings based natural language processing is less developed for low-resource languages [11]. From 24 official European-Union languages, 15 languages, including Latvian, could be considered under-resourced [12].

Consequently, the relevancy of this study can be summarized as follows: 1) it is the first application of the LDA methodology to the Latvian historical newspaper sources and one of the first applications of LDA for Latvian texts in general; 2) it is the first exploration aimed at implementation of topic modelling techniques in Latvian heritage collections; it follows also an international trend in this respect.

The structure of the article: the section "Review of previous work" explains the concept of LDA, references some notable examples of usage in the field of digital humanities, and discusses the limitations; the section "Data set" provides information on the parameters of the data and the rationale for selecting this data for the case study; the following section "Results" reports the outcomes of the LDA training of the Oskars Kalpaks' data set; the article is completed with the "Conclusions" section that summarizes the insights acquired in the case study and points towards next steps.

REVIEW OF PREVIOUS WORK

The methodologies of topic modelling entail a variety natural language processing and machine learning techniques aimed at discovering implicit thematic structures, i.e., topics in large collections of documents. A topic model is a probability distribution over the vocabulary of words that occur in the corpus of documents and it is typically expressed as probability-weighted lists of words. The lists are expected to be semantically coherent from the point of view of a human reader, however, some types of text, e.g., poetry or unbalanced collections of documents do not provide basis for thematically sound lists of keywords (nevertheless, can be still used for exploration of the corpus). The results of topic modelling algorithms can be used to summarize, visualize, explore, and theorize about a corpus [13].

Latent Dirichlet allocation, LDA, first proposed by Blei, Ng, and Jordan in 2003 [14], is currently one of the most popular topic modelling techniques [15; 16]. LDA is a probabilistic model of texts that is based on two assumptions: (1) there are a fixed number of groups of terms that tend to occur together in documents (topics); (2) each document in the corpus exhibits the topics to varying degree [13].

The exact application of probabilistic topic modeling has varied depending on the goals of researchers and domains that have been studied. Since 2003 there have been very many studies that researched the usage of the LDA and its derivatives for creating models for scientific papers [17–19] and for researching historical newspapers and magazines [20–23]. Judging from the use cases presented by the developers of methods of probabilistic topic modeling, the method was created primarily for non-fiction corpora, however, also prose fiction and poetry has been analyzed [25].

It has been theorized that historians and literary scholars use this technique differently: historians hope to work with clear, unambiguous topics while literary scholars find ambiguous topics even more informative [26]. Efforts that are aimed at producing coherent and trustworthy topics that are usable for exploring large repositories of academic papers or news sources is one of directions that is pursued in the field of topic modeling [27]. Indeed, in the context of academic repositories, the presence of intrusion words in topics and unsubstantiated mixing of topics is undesirable. In the literary studies, on the contrary, topics often are not represented as coherent, highly readable lists of semantically linked keywords. However, this is not always perceived as a deficiency, e.g., it is posited that instead of forming thematic topics around a single referent, lists of words can repre-

sent a discourse, a sociolect, or a kind of poetic rhetoric, [26], i.e., a topic model can be used not only for discovering *what* people are writing about, rather also *how* they are writing [28].

The approach that puts less weight on the coherence and comprehensibility of topics have been extended also to the studies of historical non-fiction. Since, in many cases, the output of the modelling is ambiguous and difficult to interpret, it is often emphasized that topic modeling is not necessarily useful as evidence but makes an excellent tool for discovery [29]. Topic models can point toward topics that might have otherwise remained unnoticed (since it is impossible to read all texts or read all texts with similar intensity) [30]. Hence, especially in the research of historical and literary sources, topic models can facilitate qualitative analysis and the results of topic modelling are especially useful when used concurrently with hermeneutic close reading (qualitative analysis of texts) [25; 30; 31]. Abovementioned two approaches to topics have been described also as topic realism and topic instrumentalism: topic realism is characterized as a view that the modelling process can capture representations of theoretical constructs (frames, discourses, narratives) that actually exist in the texts; topic instrumentalism is a view that topic models merely provide information about word patterns that can be useful to guide interpretation of the texts [16].

Although certain types of documents may respond poorly to formalization attempts while other can be organized and summarized more easily, it has to be recognized the quality of a topic model is determined also by the soundness of the sample, by the procedures of pre-processing, and the training parameters set for the model [32]. The parameters used for designing the Oskars Kalpaks use case is to a large extent based on the default parameters and recommendations posited by the developers of the Gensim Python library that is used to conduct this study [33]. The calculations for the optimal number of topics are based on the (C_V) coherence measure that has shown good performance with strong correlations to human ratings [34].

More than one application of LDA methodologies could be potentially useful for developing services for digital libraries of historical periodicals: 1) one topic model can be pre-trained, finetuned, and used for the entire digital collection to support the browsing functionality and recommendations for the users; 2) separate pre-trained topic models can be created for various segments of periodicals (individual titles, or types of periodicals, or time periods), implemented as supplementary functionalities with visualizations; 3) auxiliary interface could be created to allow building corpora and training models based on the search results. The use case explored in detail in this paper is to a large extent consistent with the application (3), however, as a most basic example it provides valuable insights for all applications.

One of the aspects of topic modelling that is not addressed in this paper but nevertheless highly relevant for collections that span many decades, is the dimension of time (especially for the application (1)). In its base form, the LDA method does not account for temporal aspect of the data, i.e., time is not a variable in the model, yet many use cases in humanities research include archives that span long time periods [15]. To mitigate this limitation, LDA can be trained separately for consecutive slices or data or other additional techniques of sampling and weighting of the results can be applied. Other models that take the temporal dimension into account, such as dynamic topic models, are also available [35].

DATA SET

The corpus of the Newspaper *Latvian Soldier* (*Latvijas Kareivis*) has been used for the case study. The *Latvian Soldier* was an official daily newspaper of the Latvian Army issued from 1920 to 1940; up to year 1925 it was printed in the German black letter script and transitioned to the modern script in the later years; thus, it represents a typical challenge of the orthography change characteristic to the historical newspapers that span several decades. To design the case study, only the modern part of the text was selected. In addition to that, the articles containing the string “kalpaks” were extracted from the corpus to create a model for a typical course of enquiry that is focused on researching a particular concept or topic in contrast to researching the distribution of all topics in the corpus. The pre-selection of the corpus allows to access a particular topic in greater detail: only one from 50 topics of the whole corpus of *Latvian Soldier* contained the keyword “kalpaks” while the most coherent model of topics in the preselected Oskars Kalpaks’ subcorpus contained 6 topics. Highly specialized, domain specific topic models with many high-quality, fine-grained topics provide a perspective that is remarkably different from the topic model of the whole corpus. Another motivation for working with pre-selected subcorpora rather than the whole corpus in the research situations, is avoiding working with very large data sets that can create significant challenges to handling and analyzing data.

The corpus of *Latvian Soldier* contains 55.9 M tokens; the Kalpaks’ subcorpus – 1.3 M tokens. The corpus has been segmented on the level of

articles, however, the sections that contain short news and announcements have been consolidated resulting in segments that include many themes. This aspect could be reckoned as an impediment to creation of coherent topic model, however, as illustrated in the case study, these segments are clustered together as a topic due to salient features that allow to identify (filter) material that does not contain the discussion about topics related to Oskars Kalpaks.

The corpus has been lemmatized by using the Latvian natural language processing tool pipeline NLP-PIPE [36]. The size of the vocabulary has been further reduced by omitting tokens that are less than two symbols long. These steps of pre-processing contribute to mitigating the noise created by the OCR errors. Arguably, tokens that contain only two symbols could also be removed based on the presumption that these strings are not semantically significant; however, the Kalpaks' case study has demonstrated that, at least in some cases, numbers can play a role in the grouping of documents.

Colonel Oskars Kalpaks was the Commander of the First Latvian Independent Battalion and is considered to be the first Commander in Chief of Latvian Armed Forces. Under Kalpak's leadership, the Independent Latvian Battalion became combat-ready and won the first battles that were instrumental on the road to winning the Latvian War of Independence (1918–1920). Kalpaks was posthumously awarded Latvia's highest military award, the Order of Lāčplēsis, and became a prominent figure in Latvian war history and a subject of legends.

Colonel Kalpaks exemplifies a topic that is relevant for historians and history learners and retains relevance during the time period analyzed; in addition to that, the keyword "kalpaks" is related to a range of different topics. The string "kalpaks" exemplifies several issues related to disambiguation, e.g., the corpus contains mentions of Kalpaka Boulevard, Kalpaka Street, Kalpaka Bridge that in most contexts are not related to Oskars Kalpaks; at the same time, the surname "Kalpaks" is comparatively rare and Kalpaks' many namesakes do not add complexity to this experimental model.

RESULTS

The case study was conducted by using the open source Python library Gensim. The bag-of-words dictionary was created by using individual words, bigrams and trigrams resulting in 5030 features. Following filters were applied subsequently: the vocabulary includes features that have occurred at least 20 times; features that occur in more than 50 % of the documents were omitted. Every LDA model tested in the case study was trained by applying 400 iterations and 20 epochs.

The optimal number of topics was determined by calculating the (C_V) coherence scores; the models were also subjectively reviewed by the authors of this paper. The highest (C_V) score at 0.61 was achieved for the model that consists of six topics.

The design of the use case was based on a hypothetical research questions: what topics are related to the discussion about Oskars Kalpaks in the *Latvian Soldier* from 1925 to 1940? In what contexts his name is mentioned? How many

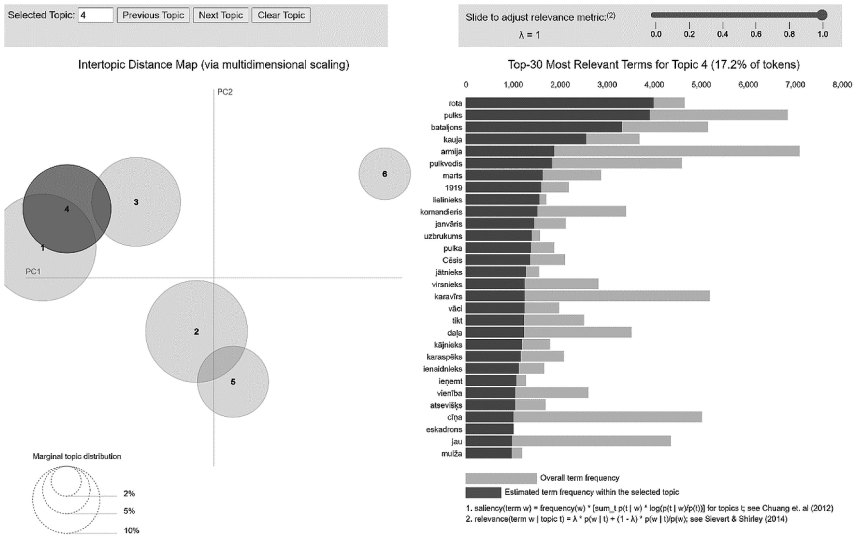

Figure 1. The distance map of six topics in the Oskars Kalpaks subcorpus (on the left); 30 most relevant terms in the topic '4' (on the right)[1]

[1] The visualization is generated by pyLDAvis Python library [37].

different topics are related to Kalpaks? How is the popularity of these topics changing over time?

Working with a corpus that is generated based on a keyword is different from the use case scenario where the model is created with a goal to discover latent discourses in a data driven way, without guiding the results by pre-selected keywords or other presuppositions. It has to be taken into account that the model that is built based on the preselected, domain-specific corpus is skewed in a particular way and can only be used to explore the contexts of the preselected keyword. Nevertheless, as discussed in the "Data set" section, the method provides advantages in some research situations.

Since the subject of study is preselected, namely, a Latvian war hero, it was expected that topics 1, 3, 4 share such keywords as regiment, battalion, soldier (see **Table 1**). On the other hand, topics 2, 5 and 6 do not display terms that we would expect to see in the discussion about Kalpaks. Although the lists are informative, they do not provide information sufficient for interpretation without referring to articles associated with each topic. This observation concurs with the insight discussed above: researchers may find that digital analysis is most helpful when used jointly with qualitative inspection of texts.

The subjective inspection of texts reveals that topics 1, 3 and 4 represent three main contexts in which Oskars Kalpaks is discussed (the LDA model treats texts as mixtures of topics, hence every article can contain several topics with different weights):

- references to remembrance days of Oskars Kalpaks that are accompanied by emotional display of patriotism (topic 1);
- accounts of remembrance days of Oskars Kalpaks that describe the ceremonies that are carried out during the celebrations and commemorations (topic 3);
- accounts that narrate the events of 1919, the forming of the Latvian army and fighting against Bolshevik troops (topic 4).

Topic 5 is present in announcements, advertisements of concerts and other culture events. Events related to Oskar Kalpaks may be referenced, however, in most cases Kalpaka Boulevard or Kalpaka Street is mentioned (the numbers are dates and times when events take place; the string "o'clock" is present in three different versions).

Topic 6 is mostly present in announcements related to schools that carry the name of Oskars Kalpaks, usually in the context of fundraising events and donations to the schools; articles from 1939 contain announcements about donations to strengthen Latvian defence forces (the numbers are sums that are being donated; the string "ls" is the abbreviation of the Latvian currency — lats).

Topic 2 appears to contain the most random mixture of keywords; it is present in a range of articles that sometimes refer to Oskar Kalpaks but, in most cases, mention the steamer "Kalpaks"; in one instance another person with the surname "Kalpaks" is mentioned.

The subjective analysis of texts is supported by the representation of data in the pyLDAvis visualization: topics 1, 3 and 4 are grouped together

Table 1

30 most relevant terms in each topic (translated from Latvian to English)

Topic 1	nation, also, this, our own, battle, to be able, army, more, then, when, you, only, country, because, already, soldier, power, entire, Latvian, I, good, myself, she, if, life, one
Topic 2	street, city, committee, yesterday, place, on, ministry, society, 10, predict, ls, minister, already, division, take place, part, head, o'clock, someone, by now, room, to be able, decide, if, house, 000, find, evening, police
Topic 3	minister, general, celebration, city, militia, president, regiment, army, soldier, chief, remembrance, grave, commander, memorial, colonel, also, take place, president of the state, fall, garrison, place, church, society, jubilee, liepāja, ceremonial, battalion, organization, church service
Topic 4	squad, regiment, battalion, battle, army, colonel, march, 1919, bolshevik, commander, january, attack, cēsis, horseman, officer, soldier, germans, get, division, infantry, troops, enemy, seize, band, separate, fight, squadron, already, manor
Topic 5	o'clock, 30, o'cl, 20, concert, 19, announcement, 15, 18, 12, 10, 17, 00, today, sound, 22, 16, record, morning, evening, street, 13, music, song, at, city, 21, take part, opera, choir
Topic 6	ls [lats], count [county], 10, school, primary school, 10 ls, 50, cl [class], 20, 25, 100, 50 ls, county, ba [unknown], committee, book, employee, teacher, 000, boy scouts, pupil, 20 ls, 15, 100 ls, soc [society], 24 ls, 30, cit [city], city, association

in the multidimensional scaling map as (see **Figure 1**).

The distribution of topics in each year from 1920 to 1940 is displayed in the Figure 2. The counts used for this visualization are simplified and differ from the method used in the pyLDAvis visualization: the pyLDAvis is a weighted calculation of all topics in all documents where the size of the bubbles is influenced also by the length of the documents; the **Figure 2** bar chart only displays the raw counts of documents sorted by the topic that is most prominent. Nevertheless, the diagram could be helpful for evaluating the prominence of topics over time and can be interpreted in the context of the historical events (coup d' tat and the regime change in 1934, beginning of the Second world war in 1939.

CONCLUSIONS

The LDA methodology for topic modelling is a promising candidate to be adopted as a service in digital collections of NLL due to the fact that it is well understood and extensively tested, and is considered especially suitable for modelling academic texts, magazines, and newspapers; it is available as open code; several modified and improved versions have been developed to overcome the limitations of standard LDA.

After the analysis of the publications, three plausible scenarious for the usage of LDA in the digital collections is determined: 1) creation of a topic model for the whole collection of periodicals to enhance browsing and discoverability; 2) creation of several topic models for different sections of collections to enhance browsing and discoverability; 3) creation of an auxiliary service that would allow prompt creation of corpora and training of models. Owing to this study, the third application is already available to the users of NLL as a service, albeit the GUI has not been developed.

During this study, a sequence of procedures necessary for preparing and processing the corpus and representing the results is being established (compiling corpus, cleaning, morphological tagging, pre-training, establishing topic counts with highest coherence scores, creating the final model, creating visualizations and full text documents with topic scores for qualitative exploration of the context).

The model created for the corpus of Oskars Kalpaks affirms that the selected methodologies of LDA and text pre-processing can result in coherent, usable topics, however, to interpret the results fully, texts of articles need to be consulted; namely, topic instrumentalism approach can be implemented but further testing and improvement

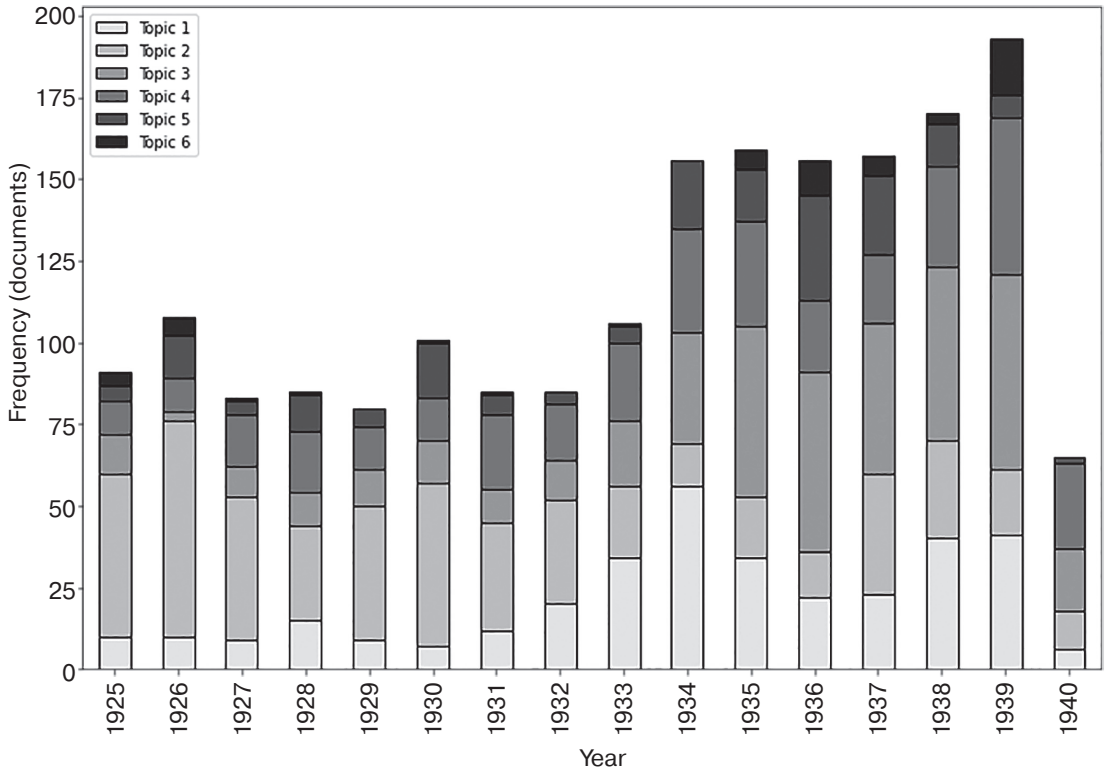

Figure 2. The distribution of topics 1–6 in the consecutive years 1925–1940. The frequency counts refer to the number of documents where a particular topic has the highest value

of the method is needed to accommodate needs of topic realism. Current application can be viewed as a significant tool to enhance the qualitative analysis of a given research subject.

The segmentation practice that joins together short news items is a potential impediment to creating topics. In this use case, however, topic model allowed to identify and group together these sections, which was helpful. Further testing is needed to determine what is the impact for larger corpora.

As a next step, the LDA method will be tested for larger corpora, without preselecting articles; the method of dynamic topic models will be applied for corpora that span more than two decades.

ACKNOWLEDGMENTS

The case study of application of Latent Dirichlet allocation for historical newspapers is a part of a research project "Text Analysis Methods and Tools for Similarity Metrics in Large National Text Corpora: the Case of the Latvian National Digital Library (LNDL) and the National Repository of Academic Texts of Ukraine (NRATU)" that aims at determining algorithms and methods that would be suitable for researching similarity in the text corpora of LNDL and NRATU [38]. The project is conducted within the Latvian-Ukrainian Joint Programme of Scientific and Technological Cooperation [39].

REFERENCES

1. The main page of the official web portal National Digital Library of Latvia. Retrieved from: https://www.lndb.lv/.
2. Krūmiņa, L. (2012). Digitalizācija Latvijā pasaules pieredzes kontekstā. *Bibliotēku pasaule.* Vol. 57. P. 39-45. Retrieved from: https://dom.lndb.lv/data/obj/file/162387.pdf.
3. Zariņš, U. (2014). Eiropas kultūras mantojums digitālajā vidē. *Latvijas intereses Eiropas Savienībā.* No. 2. P. 41–55. Retrieved from: https://dom.lndb.lv/data/obj/61436.html
4. The comprehensive list of the digitized periodicals can be viewed in the website periodika.lv. Retrieved from: https://periodika.lndb.lv/#allPeriodical.
5. Ehrmann, M., Romanello M., & Clematide, S. et. al. (2020). Language Resources for Historical Newspapers: The Impresso Collection. *LREC 2020 Proceedings.* P. 958–968. Retrieved from: http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.121.pdf.
6. Digital Approaches in Cultural Heritage: towards a pan-Baltic cooperation network: final report. Riga: National Library of Latvia, 2019. Retrieved from: https://dom.lndb.lv/data/obj/781145.html.
7. McGillivray, B.; Schuster, K., Dunn, S. (Eds.) (2021). Computational methods for semantic analysis of historical texts. *Routledge International Handbook of Research Methods in Digital Humanities.* London; New York: Routledge, Taylor & Francis Group. P. 261–274. https://doi.org/10.4324/9780429777028-20
8. Bollmann, M. (2019). A Large-Scale Comparison of Historical Text Normalization Systems. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers) — Association for Computational Linguistics, P. 3885–3898.
9. Abney, S., & Bird, S. (2010). The Human Language Project: building a universal corpus of the world's languages. *Proceedings of the 48th Meeting of the Association for Computational Linguistics Association for Computational Linguistics*. P. 88–97.
10. Skadiņa, I., Veisbergs, A., Vasiļjevs, A. et al. (2012). The Latvian Language in the Digital Age / Latviešu valoda digitālajā laikmetā. *META-NET White Paper.* Berlin: Springer.
11. Alabi, J. O., Amponsah-Kaakyire, K., & Adelani, D., et.al. (2020). Massive vs. Curated Embeddings for Low-Resourced Languages: the Case of Yorub` a and Twi. *Proceedings of the 12th Language Resources and Evaluation Conference.* European Language Resources Association, P. 2754–2762.
12. Alves, D., Thakkar, G., & Tadić, M. (2020). Evaluating Language Tools for Fifteen EU-official Under-resourced Languages. *Proceedings of the 12th Language Resources and Evaluation Conference.* European Language Resources Association, P. 1866–1873.
13. Blei, D. M. (2012). Topic modeling and digital humanities. Journal of Digital Humanities, 2 (1). Retrieved from: http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/
14. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3* (January). P. 993–1022. Retrieved from: https://dl.acm.org/doi/10.5555/944919.944937.
15. Marjanen, J., Zosa, E., Hengchen, S., et. al. (2020). Topic Modelling Discourse Dynamics in Historical Newspapers. *Post-Proceedings of the 5th Conference Digital Humanities in the Nordic Countries (DHN 2020).* P. 63–77. Retrieved from: http://ceur-ws.org/Vol-2865/paper6.pdf
16. Pääkkönen, J., & Ylikoski, P. (2020). Humanistic interpretation and machine learning. *Synthese* Retrieved from: https://link.springer.com/article/10.1007/s11229-020-02806-w. https://doi.org/10.1007/s11229-020-02806-w
17. Blei D. M., & Lafferty, J. (2007). A correlated topic model of Science. // *Annals of Applied Statistics,* Vol.1(1). P. 17–35. Retrieved from: https://projecteuclid.org/journals/annals-of-applied-statistics/volume-1/issue-1/A-correlated-topic-model-of-Science/10.1214/07-AOAS114.full. https://doi.org/10.1214/07-aoas114
18. Newman, D., Chemudugunta, C., Smyth P., & Steyvers, M. (2006). Analyzing entities and topics in news articles using statistical topic models. *Intelligence and Security Informatics, Lecture Notes in Computer Science.* Retrieved from: https://www.researchgate.net/publication/221246920_Analyzing_Entities_and_Topics_in_News_Articles_Using_Statistical_Topic_Models. https://doi.org/10.1007/11760146_9
19. Hall, D., & Jurafsky, C. D. (2008). Manning. Studying the history of ideas using topic models. In EMNLP. Retrieved from: https://web.stanford.edu/~jurafsky/hallemnlp08.pdf. https://doi.org/10.3115/1613715.1613763
20. Block, S. (2006). Doing More with Digitization: An introduction to topic modeling of early American sources. *Common-place: The Interactive Journal of Early American Life.* 6.2. Retrieved from: http://commonplace.online/article/doing-more-with-digitization/.

20. Block, S. (2006). Doing More with Digitization: An introduction to topic modeling of early American sources. *Common-place: The Interactive Journal of Early American Life.* 6.2. Retrieved from: http://commonplace.online/article/doing-more-with-digitization/.
21. Nelson, R. K. (2011). Mining the Dispatch. Retrieved from: https://dsl.richmond.edu/dispatch/introduction.
22. Templeton, T. C., Brown, T., Battacharyya, S., & Boyd-Graber, J. (2012). Mining the Dispatch under Supervision: Using Casualty Counts to Guide Topics from the Richmond Daily Dispatch Corpus, Chicago Colloquium on Digital Humanities and Computer Science.
23. Hengchen, S. (2017). When Does it Mean? Detecting Semantic Change in Historical Texts. *Ph.D. thesis*. Universite libre de Bruxelles.
24. Viola, L., & Verheul, J. (2019). Mining ethnicity: Discourse-driven topic modelling of immigrant discourses in the USA, 1898–1920. Digital Scholarship in the Humanities. https://doi.org/10.1093/llc/fqz068
25. Rhody Lisa M. (2012). Topic Modeling and Figurative Language. Journal of Digital Humanities. Vol. 2, No. 1. Retrieved from: http://journalofdigitalhumanities.org/2-1/topic-model-data-for-topic-modeling-and-figurative-language-by-lisa-m-rhody/.
26. Underwood, T. (2012). Topic modeling just made simple enough. Blog post. Retrieved from: https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/.
27. Chang, J., Boyd-Graber, J., & Gerrish, S., et. al. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems 22 (NIPS 2009).* Retrieved from: https://proceedings.neurips.cc/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf.
28. Goldstone, A., & Underwood, T. (2012). What Can Topic Models of PMLA Teach Us About the History of Literary Scholarship? *Journal of Digital Humanities — 2012*. Retrieved from: http://journalofdigitalhumanities.org/2-1/what-can-topic-models-of-pmla-teach-us-by-ted-underwood-and-andrew-goldstone/.
29. Brett, M. R. (2012). Topic Modeling: A Basic Introduction. *Journal of Digital Humanities.* Vol. 2. No. 1. Retrieved from: http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/.
30. Kurvinen, H.; Fridlund, M., Oiva, M., & Paju, P. (Ed.). (2020). Towards Digital Histories of Women's Suffrage Movements. *Digital Histories: Emergent Approaches within the New Digital History.* Helsinki University Press. P. 159.
31. Viola, L., & Verheul, J. (2019). Mining ethnicity: Discourse-driven topic modelling of immigrant discourses in the USA, P. 1898–1920. *Digital Scholarship in the Humanities.* Retrieved from: https://www.researchgate.net/publication/339140752_Mining_ethnicity_Discourse-driven_topic_modelling_of_immigrant_discourses_in_the_USA_1898-1920. https://doi.org/10.1093/llc/fqz068
32. Wallach, H., Mimno, D., & McCallum, A. (2009). Rethinking LDA: Why priors matter. *Advances in Neural Information Processing Systems.* Vol. 23. Jaunuary, P. 1973–1981. Retrieved from: http://dirichlet.net/pdf/wallach09rethinking.pdf.
33. Řehůřek, R., & Sojka, P. Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. 2010.* Retrieved from: http://is.muni.cz/publication/884893/en.
34. Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining — WSDM '15.* P. 399-408. Retrieved from: https://svn.aksw.org/papers/2015/WSDM_Topic_Evaluation/public.pdf. https://doi.org/10.1145/2684822.2685324
35. Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd international conference on Machine Learning.* P. 113–120. https://doi.org/10.1145/1143844.1143859
36. Znotiņš, A., & Cīrule, E. (2018). NLP-PIPE: Latvian NLP Tool Pipeline. *Human Language Technologies — The Baltic Perspective, IOS Press.* Vol. 307. P. 183–189.
37. Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. Proceedings of the workshop on interactive language learning, visualization, and interfaces. *Association for Computational Linguistics.* P. 63–70. Retrieved from: https://www.researchgate.net/publication/265784473_LDAvis_A_method_for_visualizing_and_interpreting_topics. https://doi.org/10.3115/v1/w14-3110
38. Project: Text Analysis Methods and Tools For Similarity Metrics in Large National Text Corpora. Retrieved from: https://lnb.lv/en/projects/text-analysis-methods-and-tools-similarity-metrics-large-national-text-corpora.
39. Latvian-Ukrainian Bilateral Cooperation Programme projects. Retrieved from: https://www.lu.lv/en/science/programmes-and-projects/international-programmes/latvian-ukrainian-bilateral-cooperation-programme-projects/.

А. БАКЛАНЕ, маг. філософії
В. САУЛЕСПУРЕНС, маг. комп'ютерних наук

ЗАСТОСУВАННЯ ЛАТЕНТНОГО РОЗПОДІЛУ ДІРІХЛЕ ДЛЯ АНАЛІЗУ ЛАТВІЙСЬКИХ ІСТОРИЧНИХ ГАЗЕТ: ПРИКЛАД ОСКАРА КАЛПАКА

Резюме. *Упродовж останніх 20-ти років тематичне моделювання і, зокрема, застосування моделі LDA (прихованого розподілу Діріхле) стало одним із найчастіше використовуваних методів дослідницького аналізу та пошуку інформації з текстових джерел. Хоча тематичне моделювання використовувалося для досліджень у великій кількості проєктів, ця технологія ще не стала частиною загальних стандартних функцій цифрових історичних колекцій, що куруються бібліотеками, архівами та іншими установами пам'яті. Окрім того, чимало широко поширених і добре вивчених методів обробки природної мови, включаючи тематичне моделювання, недостатньо застосовувалися для роботи з джерелами нечисленних або малоресурсних мов, включаючи латиську. У статті представлені результати першого тематичного дослідження, у якому методологія LDA використовувалася для аналізу набору даних історичних газет латиською мовою. Для проведення аналізу*

використовується корпус газети «Латвійський солдат», на прикладі виконання тем, пов'язаних із першим командувачем Латвійської армії Оскаром Калпаксом. У дослідженнях цифрових гуманітарних наук результати тематичного моделювання використовувалися й інтерпретувалися декількома різними способами залежно від типу та жанру тексту, наприклад, для отримання семантичних зв'язних, які заслуговують на довіру для списків ключових слів або для отримання лексичних ознак, які не допомагають тематичному аналізу, але замість цього дають інші відомості про використання мови. Автори статті пропонують додатки, які могли б бути найбільш підходящими для аналізу історичних газет у великих цифрових колекціях установ пам'яті, а також розповідають про проблеми, пов'язані з роботою з текстовими джерелами, що містять помилки оптичного розпізнавання, проблематичну сегментацію статей та інших несучасних даних.

Ключові слова: *моделювання тем, латентне розподілення Діріхле, когерентність тем, історичні газети, обробка природної мови для латиської мови, цифрові гуманітарні науки, Оскарс Калпакс.*

INFORMATION ABOUT THE AUTHORS

Anda Baklāne — Master of Philosophy, Researcher and the Head of Digital Research Services at the Department of the Development of Digital Services, at the National Library of Latvia, 3, Mūkusalas Str., Riga, Latvia, LV-1423; +(371)67806100; anda.baklane@lnb.lv; ORCID: 0000-0002-0301-2504

Valdis Saulespurēns — Master of Computer Science, Researcher and data analyst at the Technology Department of the National Library of Latvia, 3, Mūkusalas Str., Riga, Latvia, LV-1423; +(371)67806100; valdis.saulespurens@lnb.lv; ORCID: 0000-0002-9665-0125

ІНФОРМАЦІЯ ПРО АВТОРІВ

Анда Баклане — магістр філософії, дослідниця та керівниця проєктів цифрових досліджень у Департаменті розвитку цифрових послуг Латвійської національної бібліотеки, Латвія, Рига, вул. Мукусалас, 3, LV-1423; +(371)67806100; anda.baklane@lnb.lv; ORCID: 0000-0002-0301-2504

Валдіс Саулеспуренс — магістр комп'ютерних наук, дослідник та розробник цифрових сервісів Технологічного відділу Латвійської національної бібліотеки, Латвія, Рига, вул. Мукусалас, 3, LV-1423; +(371)67806100; valdis.saulespurens@lnb.lv; ORCID: 0000-0002-9665-0125