Topic Modelling of the Czech Supreme Court Decisions

Tereza NOVOTNÁ [a,1], Jakub HARAŠTA [a] and Jakub KÓL [b]
[a] *Institute of Law and Technology, Masaryk University, Brno, Czech republic*
[b] *Atlas Consulting spol. s r.o., Ostrava, Czech Republic*

Abstract. The Czech Supreme Court produces significant amount of decisions totalling more than 130 000 decisions since 1993. The amount makes it difficult for law practitioners to research this case law. This work focuses on topic models for enhanced information retrieval through identification of case law approaching the same or similar issues. We provide initial quantitative evaluation of Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) models according to *CV* coherence score for different number of topics modelled n_k= {10, 20, ..., 90, 100}. Additionally, we provide qualitative evaluation for LDA and NMF models n_k= {20, 30} that will serve as a starting point for subsequent expert-user evaluation.

Keywords. topic modelling, Latent Dirichlet Allocation, Non-negative Matrix Factorization, court decisions, coherence score

1. Introduction and Motivation

The Czech Supreme Court produces significant amount of decisions totalling more than 130 000 decisions since 1993. The amount makes it difficult for law practitioners to research this case law. In this paper, we apply topic modelling methods in order to provide less time-consuming and more efficient court decisions retrieval.

Ultimately, our aim is to provide more accurate legal information retrieval methods that take into consideration specifics of the Czech law and the Czech language and are extensively evaluated by lawyers knowledgeable in the Czech law and practicing within the jurisdiction.

2. Related Work

A general purpose of topic modelling methods is to discover underlying topic structures in the given set of documents. These topics are probability distributions over a set of words. This method is beneficial for many information retrieval tasks. It is fundamentally unsupervised, however it has many supervised or semi-supervised extensions or ap-

[1] E-mail: tereza.novotna@mail.muni.cz.

Proceedings of the 2020 Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL)

plications [13]. For our experiment, we have selected two well-known topic modelling algorithms, Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF).

LDA was first introduced as an unsupervised model in [1]. It was successfully used for classifying journalistic texts in [6,13] and it was applied on Twitter tweets in [7] demonstrating that this method may not be the best option for short texts. Additionally, it was used for summarizing of scientific papers [5]. In legal IR, this method is used for topic detection or clustering of similar documents [8,11]. NMF was first introduced in [15] and gained subsequent popularity through [10] as an innovative data clustering algorithm. NMF algorithm was used for polyphonic music transcription [17] or for document clustering in [19,9].

Part of the research in topic modelling focuses on comparing different models. For this purpose, various extrinsic and intrinsic evaluation methods were designed [16]. Coherence score, as one of the intrinsic methods, is a metric expressing the logical order or coherence of topics and thus enables machine (quantitative) validation and comparison. Coherence calculation is applied to the most important words of the topic and the result is then calculated as the sum or arithmetic mean of all these values. The *CV* (or sometimes referred to as "c_v") coherence score which is used as an automatic validation measure in this paper was introduced in [16]. It is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity [12].

It was reported that NMF outperforms LDA in topic coherence score when using corpora of relatively shorts texts [4,2]. On the other hand, LDA provided more coherent topics in the case of longer texts and outperformed NMF [18,12]. Considering that purpose of our research is to provide more accurate legal information retrieval, expert-user evaluation of the relevance of topics assigned to documents is more important than coherence measure. However, in this phase of our research, we use coherence measure to compare LDA and NMF in different settings (different number of topics n_k) to identify most coherent setting to be later subjected to expert-user evaluation.

3. Method

We used CzCDC 1.0 corpus [14], specifically its Supreme Court subset containing 111 977 decisions (published between 01/1994 and 09/2018). We removed so called unifying opinions due to their highly specific nature (both substantive and procedural), which left us with 111 187 decisions. Subsequently, we removed headers (containing names of judges, identification of parties and their representatives etc.), numbers, punctuation symbols and stop words based on the general list of Czech stop words. Subsequently, we used part-of-speech tagging to select nouns and adjectives (because these are usual bearers of meaning in legal language), and used lemmatization and short words removal (all words shorter than three characters were removed). Finally, all the remaining characters were transformed into lower-case form. We used spaCy python library with its extension for the Czech language via ud-pipe.[2]

[2] https://pypi.org/project/spacy-udpipe/.

We relied on LDA and NMF implementation algorithm in gensim package.[3] One of the parametres to be set for both methods is the extremes filtering removing words that appear either very often or very little. We removed all tokens appearing in less than 5 documents and all tokens appearing in more than half of the documents. The upper limit stems from the fact that the Czech Supreme Court serves as an apex court for civil and criminal cases. Hence, if the token appears in more than half of the documents, we assumed that it appears in both civil and criminal branch as a very common term without any specific legal importance. Furthermore, NMF model is built on the tf-idf corpus.

We used the resulting dataset to train both LDA and NMF models over ten instances with a different number of topics $n_k = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$.

4. Results and Discussion

The *CV* coherence score comparing LDA and NMF models for the same n_k is presented in Figure 1. LDA model with $n_k = 30$ and NMF model with $n_k = 20$ achieved the highest *CV* coherence score at 0.6387 or 0.7418 respectively. At the same time, all of the instances of NMF models achieved higher *CV* coherence score than instances of LDA models with the same n_k. All models and their topics as pyLDAvis graphs are available at github page.[4] Generally, scores for both methods in different settings are relatively high compared to those reported in [16].

This means that topics are composed of highly coherent keywords or semantically related terms. Following the interpretation in [16], topics and their distribution in the corpus of court decisions should be relatively similar to human evaluation. To confirm this assumption, further qualitative analysis is necessary. As is evident from [13,16,3], the correlation between human evaluation and coherence score of different models varies greatly and depends on various parameters. It is not sufficient to declare success in this experiment based on *CV* coherence score of different models alone. Initial subsequent analysis is precursor of successful larger expert-user evaluation.

Given the scope of this paper, it is impossible to provide for qualitative evaluation of all twenty models. For small-scale initial qualitative analysis, we selected the LDA model with the highest *CV* coherence score $n_k = 30$ and the NMF model with the highest *CV* coherence score $n_k = 20$ plus corresponding NMF model with $n_k = 30$.

The list of topics for both selected models reveals interesting tendencies. First of all, while NMF model offers higher *CV* coherence score in general, topics are more general. LDA topics are more specific. Therefore, it can discover more of the less common topics, that the NMF method does not reveal at all. For example, the topic no. 29 contains words related to the inheritance (*"child"*, *"adult"*, *"inheritance"*, *"testator"* etc.). The NMF method does not contain these words at all in the most coherent model. Logically, this may be caused by the fact that the model has a third smaller number of topics all together but furthermore not even the 30 topics NMF model contains this topic. If we look at the 30 topics NMF model, we can see that a higher number of topics means more general or even interchangeable topics (for example topics no. 15 and no. 16). At the same time,

[3] https://radimrehurek.com/gensim/.

[4] https://github.com/tm-czech-supreme-court/lda-nmf-models

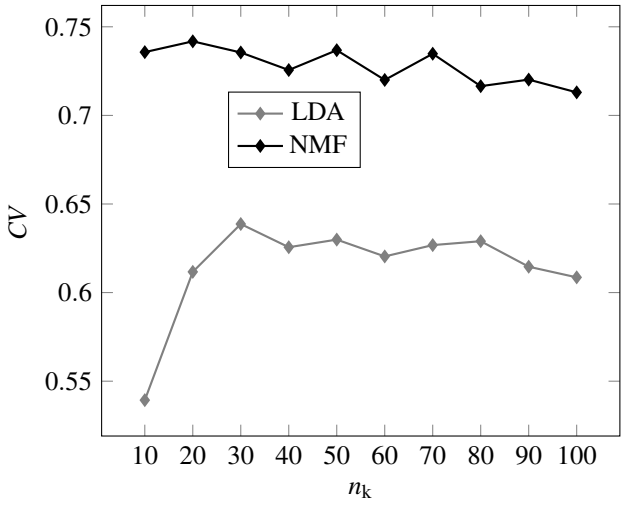

Figure 1. Comparison of *CV* coherence scores for LDA and NMF models

NMF and LDA model differs in tendencies towards more substantive (LDA) or more procedural (NMF) topics. Both of these approaches can be used in practice, as lawyers research case law focusing on both procedural and substantive aspects. As such, these deserve specific consideration within the subsequent expert-user evaluation.

Similarly, there are differences between models related to civil law and criminal law Supreme Court decisions. The NMF 20 and 30 topics models tend to cluster criminal law terms into a few main topics without further distinction, on the other hand, the 30 topics LDA model tends to contain more criminal law topics with finer distinction. The 20 topics NMF model contains 3 criminal law topics, the 30 topics NMF model contains 4 criminal law topics and the LDA 30 topics model contains 5 purely criminal topics.

5. Conclusion and Future Work

Quantitative results show that *CV* coherence score for all models is relatively high and that both methods provide relatively meaningful and coherent legal topics. NMF models generally score higher in *CV* coherence score, while LDA models appear to provide more details and specific topics. Qualitative analysis of results also shows that NMF models are better at identification of procedural topics, while LDA models are better at identification of substantive topics. This suggests that ability to use these models to allow better information retrieval is goal-specific.

This short paper is part of larger project, our future work requires extensive expert-user evaluation to identify whether results of our initial quantitative and qualitative analysis of LDA and NMF models are supported by expert-user experience. The expert-user evaluation will include topics identified by both LDA and NMF models and expert-users will be tasked to identify which topics describe the specific court decisions more accurately.

Acknowledgment

This publication was supported by Masaryk University (MUNI/A/1454/2019, Automatic Processing of Court Decisions: User Experiment). We would like to thank Vincent Kríž for his consultations and advices.

References

[1] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2003, 3(4–5), p. 993–1022.

[2] O'Callaghan D, Green D, Carthy J, Cunningham P. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 2015, 42(13), p. 5645–5657.

[3] Chang J, Boyd-Graber J, Gerrish S, Wang C, Blei DM. Reading Tea Leaves: How Humans Interpret Topic Models. *Proceedings of Neural Information Processing Systems (NIPS) 2009*, p. 288-296.

[4] Chen Y, Zhang H, Liu R, Ye Z, Lin J. Experimental explorations on short text topic mining between LDA and NMF based Schemes. *Knowledge-Based Systems*, 2019, 163(1), 1–13.

[5] He L, Li W, Zhuge H. Exploring Differential Topic Models for Comparative Summarization of Scientific Papers. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, p. 1028–1038.

[6] Jacobi C, Van Atteveldt W, Welbers K. Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 2016, 4(1), p. 89–106.

[7] Jonsson E, Stolee J. An Evaluation of Topic Modelling Techniques for Twitter. Research paper. 2016, `https://www.cs.toronto.edu/~jstolee/projects/topic.pdf`.

[8] Kumar VR, Raghuveer K. Legal Document Summarization using Latent Dirichlet Allocation. *International Journal of Computer Science and Telecommunications*, 2012, 3(7), p. 114-117.

[9] Laxmi L, Kumar PK, Shankar K, Lakshmanaprabu SK, Vidhyavathi RM, Maseleno A. Charismatic Document Clustering Through Novel K-Means Non-negative Matrix Factorization (KNMF) Algorithm Using Key Phrase Extraction. *International Journal of Parallel Programming*, 2002, 48(3), p. 496–514.

[10] Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature*, 1999, 401(6755), p. 788–791.

[11] Lu Q, Conrad JG, Al-Kofahi K, Keenan W. Legal Document Clustering with Built-in Topic Segmentation. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management 2011*, p. 383–392.

[12] Mifrah S. Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID'19 Corpus. *International Journal of Advanced Trends in Computer Science and Engineering*, 2020, 9(4), p. 5756–5761.

[13] Nikolenko SI, Koltcov S, Koltsova. Topic modelling for qualitative studies. *Journal of Information Science*, 2017, 43(1), p. 88–102.

[14] Novotná T, Harašta J. The Czech Court Decisions Corpus (CzCDC): Availability as the First Step. 2019, arXiv:1910.09513.

[15] Paatero P, Tapper U. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*,1994, 5(2), p. 111-126.

[16] Röder M, Both A, Hinneburg A. Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM 2015*, p. 399–408.

[17] Smaragdis P, Broxn JC. Non-negative matrix factorization for polyphonic music transcription. *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, p. 177–180.

[18] Suri P, Roy NR. Comparison between LDA NMF for event-detection from large text stream data. *Proceedings of the 3rd International Conference on Computational Intelligence Communication Technology (CICT 2017)*, p. 1–5.

[19] Xu W, Liu X, Gong Y. Document clustering based on non-negative matrix factorization. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval 2003*, p. 267–273.