DOI 10.5817/MUJLT2020-1-5

DOCUMENT SIMILARITY OF CZECH SUPREME COURT DECISIONS*

by

TEREZA NOVOTNÁ**

Retrieval of court decisions dealing with a similar legal matter is a prevalent task performed by lawyers as it is a part of a relevant decision-making practice review. In spite of the natural language processing methods that are currently available, this legal research is still mostly done through Boolean searches or by contextual retrieval. In this study, it is experimentally verified whether the doc2vec method together with cosine similarity, can automatically retrieve the Czech Supreme Court decisions dealing with a similar legal issue as a given decision. Furthermore, the limits and challenges of these methods and its application on the Czech Supreme Court decisions are discussed.

KEY WORDS

Automatic Court Decisions Processing, Cosine Similarity, Czech Supreme Court, Document Semantic Similarity, Doc2Vec

1. INTRODUCTION AND MOTIVATION

Although the Czech legal system belongs to continental legal systems based on the statutes and regulations, the role of the top-tier court's decisions is significant. This role is continuously theoretically examined by legal scholars and academics in order to find its position in the common law and continental law spectrum.[1] There is a consensus among the Czech legal professional public that judicial decisions are not generally binding as they

* Author gratefully acknowledges the institutional support of *Masaryk University*.

** tereza.novotna@mail.muni.cz, Ph.D. candidate at the Institute of Law and Technology, Masaryk University in Brno, The Czech Republic.

1 See: MacCormick, N., Summers, R. S. (1997) *Interpreting Precedents. A Comparative Study*. Dartmouth: Aldeshot; Smejkalová, T. (2019) Judikatura, nebo precedens? *Právník. Teoretický časopis pro otázky státu a práva*, 158 (9), pp. 852–864.

are in the common law systems.[2] As nowadays we see the binding effect of the decisions more as a spectrum then a binary option[3], the question is where in this spectrum Czech top-tier court decisions lie. Therefore, there are several characteristics to be taken into account when it comes to the binding effect of the highest court decisions in the Czech Republic. One of the attributes of the role of court decisions of *Supreme*, *Supreme Administrative*, and *Constitutional Court* is the consistency of decision--making practice. The decision-making practice should be predictable and repetitive in order to fulfil the conditions of the principle of legal certainty.

In spite of the fact that the reality is usually more complicated, the analysis of previous decision-making practice in the similar matter is still a significant part of work of every judge, lawyer, legal scholar or student. Furthermore, the analysis of court decisions takes a great part in the academic journals, scientific publications, students' books or different kinds of commentaries. In the Czech legal society, whole journals are dedicated to overview current court decisions[4], other journals have special sections for the overview and annotations of current decisions[5], famous legal focused accounts on social media provide actualities from the current decision-making practice of individual courts,[6] or generally used commercial systems provide the newest decisions in special sections.

The *Czech Supreme Court* publishes approximately between 5 and 7 thousands of decisions per year, and this number is continuously increasing (see Figure 1).[7]

---

[2] Different legal scholars come up with different approaches to tackle the binding effect of decisions. See for example: Harvánek, J. et al. (2008) *Teorie práva*. Plzeň: Vydavatelství a nakladatelství Aleš Čeněk, p. 261; Gerloch, A. (2017) *Teorie práva*. 7th ed. Plzeň: Vydavatelství a nakladatelství Aleš Čeněk, p. 90; Kubů, L., Hungr P. and Osina P. (2007) *Teorie práva*. Praha: Linde, p. 56; Bobek, M. et al. (2013) *Judikatura a právní argumentace*. 2nd ed. Praha: Auditorium, pp. 112, 113, 117, 118.

[3] The idea of the spectrum was formulated in: Peczenik, A. (1997) The Binding Force of Precedent. In: MacCormick, N., Summers, R. S. (eds.). *Interpreting Precedents. A Comparative Study*. Dartmouth: Aldeshot, pp. 461–479.

[4] For example Czech journal *Soudní rozhledy* from *C. H. Beck*.

[5] For example Czech journal *Revue pro právo a technologie* contains special section "Aktuální judikatura" dealing with recently published court decisions.

[6] For example Czech *Facebook* account *Iuridum daily* or Czech TV series *Týden v justici*.

[7] Data statistics of the *Supreme Court* decisions contained in the *Czech Court Decisions Corpus*. See: Novotná, T. and Harašta, J. (2019) *The Czech Court Decisions Corpus (CzCDC): Availability as the First Step*. ArXiv:1910.09513. [online] Available from: http://arxiv.org/abs/1910.09513 [Accessed 20 January 2020].

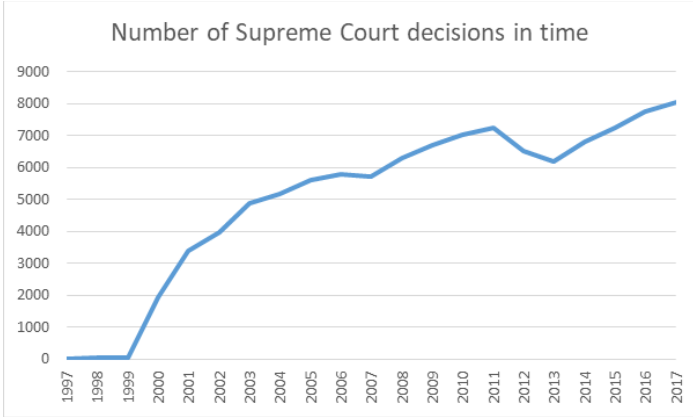

Figure 1: The time evolution of the number of Supreme Court decisions

The *Supreme Court* has a whole department just for analysis of its decisions. The employees of the *Supreme Court* are manually reading through the decisions and classifying them in order to create different kinds of collections with decisions related to different codes and articles, different keywords and topics etc. These collections then serve for better orientation in the court decisions for the judges and assistants. The manual processing of texts of court decisions is time-consuming and subjectively affected.

In this study, I choose a natural language processing (hereby *"NLP"*) method *doc2vec*[8] to automatically process the *Supreme Court* decisions and *cosine similarity* measure to compute the similarity value of the decisions[9]. *NLP* is a computer science and linguistic field dedicated to automatically process natural language in order to perform different tasks. These tasks lead in particular to better information retrieval. Different methods and different approaches are used to obtain more and more accurate and efficient results from the information retrieval systems. *NLP* methods are recently mainly based on machine learning methods and language statistics. The most common tasks in legal language processing are segmentation of legal texts[10], its summarization[11], extraction of different parts of legal texts (citations, entities etc.)[12], extraction of topics or keywords[13] and semantic similarity counting[14], just to name a few.

[8] This method was introduced in: Le, Q. and Mikolov, T. (2014) Distributed representations of sentences and documents. In: *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, Beijing, China, pp. 1188–1196.

[9] Gomaa, W. H. and Fahmy, A. A. (2013) A survey of text similarity approaches. *International Journal of Computer Applications*, 68 (13), pp. 13–18.

I apply the methods in order to experimentally answer the following hypothesis: the court decisions with a high *cosine similarity* of their *doc2vec* representations are dealing with the qualitatively similar legal issue. The hypothesis is created on the hypothetical situation where a lawyer disposes of one decision in a specific matter. She wants to obtain the *Supreme Court* decisions in a similar matter to help her to build an argumentation in her pending case. That is a standard task that every lawyer needs to perform – to review recent decision-making practice in order to choose the right strategy in a legal case. The hypothesis is based on the previous work in the similarity of the legal documents, and specifics of the Czech language described in Section 2.[15] The methodology and the data is described in Section 3. The result of this experiment and its general evaluation is in Section 4. The limits of the method, possible development and improvement as well as future work are described in Section 5. Section 6 is concluding the study with a short summarization.

2. RELATED WORK

The *doc2vec* method is based on the *word2vec* method that was originally proposed by *Mikolov et al*.[16] The *doc2vec* was proposed by *Mikolov and Le* as an extension of *word2vec* using the neural vector embedding for

[10] For example: Savelka, J. and Ashley, K. D. (2018) Segmenting U.S. Court Decisions into Functional and Issue Specific Parts. In: Palmirani, M. (ed.). *Legal Knowledge and Information Systems JURIX 2018*. IOS Press Ebooks, pp. 111–120. Available from: http://ebooks.iospress.nl/volume/legal-knowledge-and-information-systems-jurix-2018-the-thirty-first-annual-conference [Accessed 20 January 2020]; Harašta, J. et al. (2019) Automatic Segmentation of Czech Court Decision into Multi-Paragraph Parts. *Jusletter IT*, 23 May 2019, pp. 1–11.

[11] For example: Barzilay, R. and Elhadad, M. (1997) Using Lexical chains for Text Summarization. In: *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 10–17. Hearst, M. A. (1997) TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics*, 23 (1), pp. 33–64.

[12] For example: Harašta, J. et al. (2018) Annotated Corpus of Czech Case Law for Reference Recognition Tasks. In: *Text, Speech, and Dialogue: 21st International Conference proceeding*, pp. 239–250; Kríž, V. et al. (2014) Statistical Recognition of References in Czech Court Decisions. In: *Proceedings of MICAI*, pp. 51–61.

[13] For example: Ercan, G. and Cicekli, I. (2007) Using Lexical Chains for Keyword Extraction. *Information Processing & Management*, 43 (6), pp. 1705–1714.

[14] For example: Hearst, M. A. (1997) TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics*, 23 (1), pp. 33–64; Wagh, R. and Anand, D. (2017) Application of citation network analysis for improved similarity index estimation of legal case documents: A study. In: *2017 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC)*, pp. 1–5. Available from: doi:10.1109/ICCTAC.2017.8249996 [Accessed 20 January 2020].

[15] See notes 31–33.

[16] Mikolov, T. et al. (2013) Efficient estimation of word representations in vector space. In: *Proceedings of Workshop at the International Conference on Learning Representations*, Scottsdale, USA.

the whole documents (sentences, paragraphs etc.).[17] Original *word2vec* method is based on the principle, where the text is split into unique words and these words are embedded with a vector representation. Vectors emerge from the large text corpora training models in a way where the model predicts the current word from the word neighborhood (the words that the current word is often surrounded). The output of this method is a vector space model where the semantically similar words are embedded with similar vector representations.[18] The *doc2vec* is a subsequent work based on the same principles. This method is generally applicable to the text segments of any length – from sentences to whole documents. The methodology is very similar to the *word2vec*, except entire segments (referred as "paragraphs" in the original paper[19]) are embedded with a vector representation as well as unique words in the text. Furthermore, the prediction of a current word is based on both segment and word vectors. In this way, it is possible to capture the semantic context of the text segments.[20]

There is a large number of applications of this method in different fields, including legal text analysis. The empirical work proving the highest efficiency of the *doc2vec* is the one from *Lau and Baldwin*, where the authors compare the method to other embedding-based methods.[21]

The *doc2vec* is used to classify different types of documents. *Trieu, Tran and Tran* used this method to classify *Twitter* news according to their topics (or labels).[22] These news-based documents were transformed into vectors using the *doc2vec* method. The label for a current document is chosen according to the vector similarity of other document vectors in the pre-

---

[17] Le, Q. and Mikolov, T. (2014) Distributed representations of sentences and documents. In: *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, Beijing, China, pp. 1188–1196.

[18] This method was proposed with the use of two possible architectures: continuous bag-of--words and skip-gram, where in the skip-gram the word *neighborhood* is predicted based on the current word. Furthermore, the continuous bag-of-words respect word orders. For more detailed information see: Mikolov, T. et al. (2013) Efficient estimation of word representations in vector space. In: *Proceedings of Workshop at the International Conference on Learning Representations*, Scottsdale, USA.

[19] Detailed information about the *doc2vec* in: Le, Q. and Mikolov, T. (2014) Distributed representations of sentences and documents. In: *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, Beijing, China, pp. 1188–1196.

[20] Ibid.

[21] Lau, J. H. and Baldwin, T. (2016) An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. In: *Proceedings of the 1st Workshop on Representation Learning for NLP*, Berlin: Association for Computational Linguistics, pp. 78–86. Available from: https://www.aclweb.org/anthology/W16-1609.pdf [Accessed 20 January 2020].

-trained model. Sentiment analysis, as *Bilgin and Şentürk* suggest in their study, is another use of the *doc2vec*.[23]

Sentiment analysis is an analysis of emotions contained in the text. Authors successfully used *doc2vec* on the *Twitter*-based text corpus in order to define the specific product feedbacks as positive, negative or neutral. The opinion mining[24] is a method close to the semantic analysis retrieving public opinion on the specific matter. As such, this method is widely used for social media analysis. The study from *Maslova and Potapov* proves the usability for flexional languages as well (authors use Russian texts).

Recommendation of similarly focused new texts is the last example to show the possible use of *doc2vec*. In the comparative empirical study performed on the news texts, *Nandi et al.* showed that this method outperforms other widely used *NLP* methods, such as *Latent semantic analysis* and *Latent Dirichlet allocation*.[25] This study shows that the recommendation model based on *doc2vec* retrieved more contextually similar news to the original text than the two other compared methods.

All of the above-suggested uses of *doc2vec* are performed to the relatively short texts. In the legal domain, the documents are mostly longer than tweets or news articles. Despite that, successful studies applying *doc2vec* to longer legal documents were published. Firstly, the *doc2vec* can be used to determine and merge controversial issues in the case law.[26] In the *Tian*

---

[22] Trieu, L. Q., Tran, H. Q. and Tran, M.-T. (2017) News Classification from Social Media Using Twitter-based Doc2Vec Model and Automatic Query Expansion. In: *Proceedings of the Eighth International Symposium on Information and Communication Technology*. Nha Trang City, Viet Nam: Association for Computing Machinery, pp. 460–467. Available from: doi:10.1145/3155133.3155206 [Accessed 20 January 2020]; Kim, D. et al. (2019) Multi-co--training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. In: *Information Sciences*, 477, pp. 15–29. Available from: doi:10.1016/j.ins.2018.10.006 [Accessed 20 January 2020].

[23] Bilgin, M. and Şentürk, I. F. (2017) Sentiment analysis on Twitter data with semi-supervised Doc2Vec. In: *2017 International Conference on Computer Science and Engineering (UBMK)*, pp. 661–666. Available from: doi:10.1109/UBMK.2017.8093492 [Accessed 20 January 2020].

[24] Maslova, N. a Potapov, V. (2017) Neural Network Doc2vec in Automated Sentiment Analysis for Short Informal Texts. In: Karpov, A. et al. (eds.). *Speech and Computer*. Cham: Springer International Publishing, pp. 546–554. Lecture Notes in Computer Science. Available from: doi:10.1007/978-3-319-66429-3_54 [Accessed 20 January 2020].

[25] Nandi, N. R. et al. (2018) Bangla News Recommendation Using doc2vec. In: *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pp. 1–5. Available from: doi:10.1109/ICBSLP.2018.8554679 [Accessed 20 January 2020].

[26] Tian, X. et al. (2018) K-Means Clustering for Controversial Issues Merging in Chinese Legal Texts. In: Palmirani, M. (ed.). *Legal Knowledge and Information Systems JURIX 2018*. IOS Press Ebooks, pp. 215–219. Available from: http://ebooks.iospress.nl/volume/legal-knowledge-and-information-systems-jurix-2018-the-thirty-first-annual-conference [Accessed 20 January 2020].

et al. published paper, the causes of action are analyzed in order to define the controversial issue in them.

Alternatively, counting the similarity of the legal document combined with *cosine similarity* measure or retrieval of similar court decisions are common tasks to apply *doc2vec* to the legal documents. *Renjit and Idicula* applied *doc2vec* to both statutes and precedents in order to obtain the most similar to the in-hand legal document.[27] *Barco Ranera, Solano and Oco* achieved high accuracy of semantically similar court decisions retrieved using *doc2vec* model comparing to the expert evaluation.[28] A comparison of several legal court decision retrieval methods from *Mandal et al.* showed that *doc2vec* outperformed other well knows methods for legal text analysis.[29] The *cosine similarity* was compared to the network analysis in the work from *Wagh and Anand* in combination with different vector--based *NLP* method. In this study, citation network analysis had more accurate results than the *cosine similarity* of document vectors.[30]

However, there are not many studies in legal analysis field using vector embedding for the Czech language. *Novotný and Ircing* compared *doc2vec* method application on English dataset and two Czech datasets to state the high efficiency in the classification tasks even for Czech texts.[31] *Kocmi* used document embedding for machine translation in his dissertation.[32]

Additionally, there is general unavailability of the Czech legal texts corpora. In this experiment, the recently published *Czech Court Decisions*

---

[27] Renjit, S. and Idicula, S. M. (2019) CUSAT NLP@AILA-FIRE2019: Similarity in Legal Texts using Document Level Embeddings. In: Bhattacharya, P. et al. *Overview of the FIRE 2019 AILA track: Artificial Intelligence for Legal Assistance*. Proc. Of FIRE, pp. 12–15.

[28] Ranera, L. T. B., Solano, G. A., and Oco, N. (2019) Retrieval of Semantically Similar Philippine Supreme Court Case Decisions using Doc2Vec. In: *2019 International Symposium on Multimedia and Communication Technology (ISMAC)*. IEEE, pp. 1–6; Mandal, A. et al. (2017) Measuring similarity among legal court case documents. In: *Proceedings of the 10th Annual ACM India Compute Conference*, pp. 1–9.

[29] Mandal, A. et al. (2017) Measuring similarity among legal court case documents. In: *Proceedings of the 10th Annual ACM India Compute Conference*, pp. 1–9.

[30] Wagh, R. and Anand, D. (2017) Application of citation network analysis for improved similarity index estimation of legal case documents: A study. In: *2017 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC)*, pp. 1–5. Available from: doi:10.1109/ICCTAC.2017.8249996 [Accessed 20 January 2020].

[31] Novotný, J. and Ircing, P. (2018) The Benefit of Document Embedding in Unsupervised Document Classification. In: Karpov, A. et al. (eds.). *Speech and Computer*. Cham: Springer International Publishing, pp. 470–478. Lecture Notes in Computer Science. Available from: doi:10.1007/978-3-319-99579-3_49 [Accessed 20 January 2020].

[32] Kocmi, T. (2020) Exploring Benefits of Transfer Learning in Neural Machine Translation. [pre-print] Available from: https://arxiv.org/abs/2001.01622 [Accessed 20 January 2020].

Corpus was used as a source dataset of the decisions of the *Czech Supreme Court*.[33] The dataset is further described in Section 3.

3. METHODOLOGY

The *NLP* method to count the semantic similarity was chosen having regard to the nature of the documents. The *doc2vec* is a generally applicable method based on *word2vec* method transforming whole documents into vectors while building vector space model. There are several reasons why this method was chosen. First of all, *doc2vec* outperforms other vector space model methods.[34] Secondly, this method can capture the semantics of the texts because it respects the order of words.[35] Furthermore, as it was described, it can be used to the documents of different lengths.[36] Another reason is that *doc2vec* generally does not require a lemmatization step (as it will be discussed below),[37] which is very time saving considering the corpus size. Finally, *doc2vec* is easily applicable through further described python--based libraries.

The *doc2vec* method was applied to the whole dataset of the *Supreme Court* decisions available in the *Czech Court Decisions Corpus 1.0*. It was proven that this algorithm performs better for large corpora of texts[38], such as the whole corpus of *Supreme Court* decisions published which was used.

3.1. DATA

Publicly available *Czech Court Decisions Corpus 1.0* was used to build a vector space model. This dataset contains plain texts of decisions of the *Supreme Court* published between 1st January 1993 and 30th September 2018. The dataset consists of 111,977 decisions in total. According

[33] Novotná, T. and Harašta, J. (2019) *The Czech Court Decisions Corpus (CzCDC): Availability as the First Step*. ArXiv:1910.09513. [online] Available from: http://arxiv.org/abs/1910.09513 [Accessed 20 January 2020].

[34] See note 25 or 29.

[35] Le, Q. and Mikolov, T. (2014) Distributed representations of sentences and documents. In: *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, Beijing, China, pp. 1188–1196.

[36] *Czech Supreme Court* decisions vary in length from half-page documents to several pages length decisions.

[37] Hrala, M. and Král, P. (2013) Evaluation of the Document Classification Approaches. In: Burdul, R. et al. (eds.). *Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013*. Heidelberg: Springer International Publishing, pp. 877–885. Advances in Intelligent Systems and Computing. Available from: doi:10.1007/978-3-319-00969-8_86 [Accessed 20 January 2020].

[38] Mikolov, T. et al. (2013) Efficient estimation of word representations in vector space. In: *Proceedings of Workshop at the International Conference on Learning Representations*, Scottsdale, USA.

to the accompanying description, it contains approximately 91 % of all decisions of the *Supreme Court* issued within the relevant time range. These decisions are already anonymized and contain all the parts of the decisions, including the metadata heading. The heading is contained since the decisions were downloaded from the court website. The texts of the decisions are in Czech language and unstructured.

Several steps of data preprocessing must have been performed in order to build a vector space model and to process the decisions. First of all, the texts where tokenized. Tokenization means splitting the texts into predefined tokens. In this case, words were used as tokens. The second step was punctuation removal. Further off, lemmatization of tokens is a usual following step.[39] The lemmatization is a transformation of the words (as tokens) into their "lemmas". The lemma is the dictionary form of a word in the languages where a word can take different forms. This step is desirable when processing inflected languages such as the Czech language. In this particular case, using lemmatization was considered but not used in the final training model. According to the relevant literature, the lemmatization of texts does not generally improve the performance of the *doc2Vec* method.[40] As the required time when lemmatizing the model increases significantly and better performance is not expected, the lemmatization was not used in this particular case.

The last preprocessing step was the stop words removal. Stop words are words in the natural text that do not bear any meaning from the perspective of semantics, and those words are usually prepositions, conjunctions etc. Removing these words can help to increase the accuracy of text transformation, although the question of whether to remove stop words is uncertain. This is because since the vector space model is based on the statistical appearance of unique words, removing common words without a specific meaning helps to highlight other, meaningful words. For

---

[39] See for example: Schweighofer, E., Winiwarter, W. and Merkl, D. (1995) Information filtering: the computation of similarities in large corpora of legal texts. In: *Proceedings of the 5th international conference on Artificial intelligence and law*, p. 119–126; Kannan, Subbu and Gurusamy, V. (2014) Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, 5 (1), pp. 7–16.

[40] Hrala, M. and Král, P. (2013) Evaluation of the Document Classification Approaches. In: Burdul, R. et al. (eds.). *Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013*. Heidelberg: Springer International Publishing, pp. 877–885. Advances in Intelligent Systems and Computing. Available from: doi:10.1007/978-3-319-00969-8_86 [Accessed 20 January 2020].

this task, I used the general list of Czech stop words available for the *Python* libraries.[41]

Unique tasks described above are the standard preprocessing steps for many *NLP* techniques and methods.[42] It is obvious that after this process, the texts are represented rather as sets of single words (tokens), which is desirable input for many different methods of the *NLP*.

3.2. TRAINING VECTOR SPACE MODEL

The sets of words are then transformed into vectors using the *doc2vec* method. This method is based on the *word2vec* method based on machine learning. Both documents and words are represented as vectors of the dimension of N, this dimension (or length) are chosen arbitrarily according to the length of documents that are transformed. The principle of this method is a prediction of the current word according to its context (word surrounding).[43] This way, the words that usually appear in the same context (close to each other) have similar vector embeddings (such as "Supreme Court"). Creation of these vectors is taking place in the training phase of the process. The algorithm goes through all the words in many repetitions and gradually refines vector representations of unique words.

The training is based on several parameters such as the length of the vectors, the number of epochs, which is the number of iterations during the training, or the statistical limits of words that are ignored during the training. These limits are based on the prediction that words appearing in the text less than the limit do not bear any vital information about the text. Although this is a relative statistical prediction, in the case of court decisions, it is applicable, since we are looking for the common words among the individual documents. When the parameters are set, the model is trained. The training time of a model of 111,977 *Supreme Court* decisions is approximately 30 hours. Parameters can be reset, and the model can be trained again to achieve better performance of the model when the results

---

[41] This *Python* library is available from: https://pypi.org/project/stop-words/

[42] See for example: Schweighofer, E., Winiwarter, W. and Merkl, D. (1995) Information filtering: the computation of similarities in large corpora of legal texts. In: *Proceedings of the 5th international conference on Artificial intelligence and law*, p. 119–126; Kannan, Subbu and Gurusamy, V. (2014) Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, 5 (1), pp. 7–16.

[43] In the case contiuous bag-of-words algorithm is used. If skip-gram is used, then the surrounding is predicted on the base of a current word.

are unsatisfying. This is very time-consuming since every new training takes approximately the same number of hours. Once the model is successfully trained, it can be stored and used repeatedly.

3.3. COSINE SIMILARITY OF DOCUMENTS

Cosine similarity of vectors is a common method for computing semantic similarity of different text parts. It counts the *cosine value* of an angle between two vectors as two documents.[44]

In this case, I trained a vector space model consisting of vector representations of the decisions from the dataset of the *Supreme Court* decisions. To prove or disprove the initial hypothesis, I use the evaluation based on a comparison of the one decision that a lawyer possesses in the beginning, as it was set in Section 1. This decision is pre-processed in the same way as the decisions contained in the model, and then it is transformed into a vector. This vector is compared to the trained model, and the *cosine similarity* is computed in order to obtain the most similar vectors out of the training dataset. Using this methodology, I was able to obtain the list of most semantically similar decisions relatively fast.

There is a second way of how to count the similarity among the documents. Within this method, after the model training, every vector is compared to every other. This method is very time-consuming and very demanding as regards computational capacity. On the other hand, it can provide more information on the semantic similarity among the whole dataset of documents. This approach is desirable when building a network for the network or cluster analysis.

3.4. TASK DEFINITION

For the evaluation of the method and proving the hypothesis, let us consider the situation from the Section 1 as a legal information retrieval query. Let us assume that a lawyer has a *Supreme Court* decision, *26 Cdo 1471/2013*, dealing mainly with the lease agreement and its validity. She wants to obtain more *Supreme Court* decisions dealing with the same matter – the lease agreement. Therefore, according to the methodology and

[44] Gomaa, W. H. and Fahmy, A. A. (2013) A survey of text similarity approaches. *International Journal of Computer Applications*, 68 (13), pp. 13–18; Wagh, R. and Anand, D. (2017) Application of citation network analysis for improved similarity index estimation of legal case documents: A study. In: *2017 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC)*, pp. 1–5. Available from: doi:10.1109/ICCTAC.2017.8249996 [Accessed 20 January 2020].

the process described in Section 3, we compare the decision *26 Cdo 1471/2013* to the vector space model. As a result, a set of 10 most semantically similar decisions is retrieved. The number of the most similar decision is chosen mostly arbitrarily regarding the scope of this article. This number can vary according to the needs of the hypothetical lawyer from the example situation.

4. RESULTS

The list of 10 the most semantically similar decisions and their *cosine similarity* values are in Table 1.

The evaluation methodology is a qualitative analysis of retrieved decisions and its comparison from the legal point of view. Retrieved decisions were manually analyzed in order to extract the most important legal issue. These issues are in the third column of Table 1. It was observed whether the clue legal issue is thematically related to the original decision concerning the validity of a lease agreement. The *cosine similarity* value between the vector representations of a certain decision and the original decision is in the second column of Table 1.

Docket number	**Similarity value**	**Legal Issue**
20 Cdo 1003/2000	0.5084	The implicitly concluded lease agreement
26 Cdo 1143/2004	0.4667	The transition of lease of an apartment, evidence
26 Cdo 761/2003	0.4655	The termination of the lease
26 Cdo 1136/2003	0.4617	The termination of the lease
26 Cdo 567/2009	0.4611	The validity of lease agreement
33 Cdo 2593/2015	0.4605	The termination of the lease
26 Cdo 4331/2011	0.4576	Procedural decision – dismissal
26 Cdo 4801/2016	0.4572	The validity of lease agreement
26 Cdo 4898/2008	0.4536	The conclusion of lease agreement

Table 1: The most similar decisions retrieved from the model

The *doc2vec* model retrieved the 10 most semantically similar decisions to the original decision in a hierarchical order. All of the decisions are decisions arising from a civil procedure. We can observe that 8 out of the 10 decisions were decided by the same senate. Generally, senates of the *Supreme Court* are divided according to the specific matters they deal with. Furthermore, 9 of these are dealing with relatively close topics – it is either the validity of the lease agreement itself or its conclusion, termination

or transition. The question of the validity of an agreement is strongly related mainly to the question of concluding or termination of the agreement.

However, there is also a procedural dismissal – *26 Cdo 4331/2011*. This decision is very short, only a few sentences and the heading. When taking a deeper look, one can discover that the problematic issue here is the heading. The heading is very similar to the heading of the original decision; these decisions are decided by the same judges, which means that the heading contains a number of identical words. As the decision is very short, the heading forms a great part of the whole text of the decision. This uncovers one of the limits of the *doc2Vec* method, that is further discussed in following section. This method does not perform very well when combining relatively long and short texts as in this example. This is due to the fact that the dimension of word and paragraph vectors is the same for every word and document not regarding the length difference of unique documents.

Generally, it can be concluded that the hypothesis was proved – it can be stated that the semantically similar decisions are dealing with similar legal issues. However, this method has certainly many limits, and the evaluation is further discussed in Section 5.

5. DISCUSSION AND FUTURE WORK

There are a few issues to be discussed concerning mainly the method and its limits, but also data and its preprocessing and the evaluation of the experiment.

The data preprocessing part is mostly straightforward and standardized, although there are some questions to be addressed. The first is the list of stop words. The general list used in this study contains natural language Czech stop words. As the legal language is very specific from the natural one, this list could be extended with legal words bearing no important meaning. Words as court, procedure, civil etc. are words frequently appearing in the resulted decisions, although they are very general with no specific meaning for the context. For this task to be performed transparently, I suggest using a statistical density of unique words in the dataset of *Supreme Court* decisions and removing those generally most common.

To avoid the mistake of receiving the procedural dismissal not relevant to the legal information retrieval query, the document segmentation task and removal of some parts of the texts would be suitable. It is upon further discussion what parts of decisions are relevant for this specific method, however facts or argumentation parts usually bear the most important information about the case, on the contrary, the heading or even the metadata heading should be removed.

Regarding the method itself and the vector space model training, different parameters can be set to refine the results. At this point, the length of the vectors and the limits for word occurrences should be emphasized. The limits are closely related to the stop words and potentially could be stressed within the word density analysis .

The discussion on the limits of the *doc2vec* method highlights the fact that this method performed worse on the set of long texts. This issue could be partly solved with document segmentation, although the argumentation part is usually the longest one in the court decision. Therefore, I do not expect a significant improvement in this regard.

The evaluation is the last issue discussed in this Section. Although the results are transparent, the limits of qualitative analysis do not allow for comparison with different studies with a similar topic. For the experiment to be comparable, information retrieval measures such as precision and recall should be computed. For the qualitative evaluation itself, the evaluation group rather than only one evaluator should be considered, and the relevant assessment should be involved.

6. CONCLUSION

In this experimental study, I used the *doc2vec* method to count semantic similarity of the *Czech Supreme Court* decisions to prove the hypothesis that decisions with high semantic similarity deal with a similar legal issue.

I used a whole dataset of the *Supreme Court* decisions from the *Czech Court Decisions Corpus 1.0* to build a training vector space model, that was used afterwards to compute the *cosine similarity* between the decisions. I used the pre-selected *Supreme Court* decision as a test one to retrieve decisions most similar to it out of the dataset, and I qualitatively analyzed the legal issue concerned in the decisions to evaluate the method. Finally, I stated that 9 out of 10 retrieved decisions dealt with a similar legal issue, and I considered the hypothesis proven. The retrieved documents dealt

mainly with conclusion or termination of a rent agreement taken into account that the initial decision dealt with the validity of the rent agreement. From a legal point of view, these questions are related.

The *doc2vec*, as well as data preprocessing and evaluation method, have their limits that influence the performance of the method. These limits were discussed further, and possible improvements were suggested.

LIST OF REFERENCES

[1] Barzilay, R. and Elhadad, M. (1997) Using Lexical chains for Text Summarization. In: *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 10–17.

[2] Bilgin, M. and Şentürk, I. F. (2017) Sentiment analysis on Twitter data with semi--supervised Doc2Vec. In: *2017 International Conference on Computer Science and Engineering (UBMK)*, pp. 661–666. Available from: doi:10.1109/UBMK.2017.8093492 [Accessed 20 January 2020].

[3] Bobek, M. et al. (2013) *Judikatura a právní argumentace*. 2nd ed. Praha: Auditorium.

[4] Ercan, G. and Cicekli, I. (2007) Using Lexical Chains for Keyword Extraction. *Information Processing & Management*, 43 (6), pp. 1705–1714.

[5] Gerloch, A. (2017) *Teorie práva*. 7th ed. Plzeň: Vydavatelství a nakladatelství Aleš Čeněk.

[6] Gomaa, W. H. and Fahmy, A. A. (2013) A survey of text similarity approaches. *International Journal of Computer Applications*, 68 (13), pp. 13–18.

[7] Harašta, J. et al. (2018) Annotated Corpus of Czech Case Law for Reference Recognition Tasks. In: *Text, Speech, and Dialogue: 21st International Conference proceeding*, pp. 239–250.

[8] Harašta, J. et al. (2019) Automatic Segmentation of Czech Court Decision into Multi--Paragraph Parts. *Jusletter IT*, 23 May 2019, pp. 1–11.

[9] Harvánek, J. et al. (2008) *Teorie práva*. Plzeň: Vydavatelství a nakladatelství Aleš Čeněk.

[10] Hearst, M. A. (1997) TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics*, 23 (1), pp. 33–64.

[11] Hrala, M. and Král, P. (2013) Evaluation of the Document Classification Approaches. In: Burdul, R. et al. (eds.). *Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013*. Heidelberg: Springer International Publishing, pp. 877–885. Advances in Intelligent Systems and Computing. Available from: doi:10.10 07/978-3-319-00969-8_86 [Accessed 20 January 2020].

[12] Kannan, Subbu and Gurusamy, V. (2014) Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, 5 (1), pp. 7–16.

[13] Kim, D. et al. (2019) Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. In: *Information Sciences*. 477, pp. 15–29. Available from: doi:10.1016/j.ins.2018.10.006 [Accessed 20 January 2020].

[14] Kocmi, T. (2020) *Exploring Benefits of Transfer Learning in Neural Machine Translation*. [preprint] Available from: https://arxiv.org/abs/2001.01622 [Accessed 20 January 2020].

[15] Kríž, V. et al. (2014) Statistical Recognition of References in Czech Court Decisions. In: *Proceedings of MICAI*, pp. 51–61.

[16] Kubů, L., Hungr, P. and Osina, P. (2007) *Teorie práva*. Praha: Linde.

[17] Lau, J. H. and Baldwin, T. (2016) An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. In: *Proceedings of the 1st Workshop on Representation Learning for NLP*, Berlin: Association for Computational Linguistics, pp. 78–86. Available from: https://www.aclweb.org/anthology/W16-1609.pdf [Accessed 20 January 2020].

[18] Le, Q. and Mikolov, T. (2014) Distributed representations of sentences and documents. In: *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, Beijing, China, pp. 1188–1196.

[19] MacCormick, N. and Summers, R. S. (1997) *Interpreting Precedents. A Comparative Study*. Dartmouth: Aldeshot.

[20] Mandal, A. et al. (2017) Measuring similarity among legal court case documents. In: *Proceedings of the 10th Annual ACM India Compute Conference*, pp. 1–9.

[21] Maslova, N. and Potapov, V. (2017) Neural Network Doc2vec in Automated Sentiment Analysis for Short Informal Texts. In: Karpov, A. et al. (eds.). *Speech and Computer*. Cham: Springer International Publishing, pp. 546–554. Lecture Notes in Computer Science. Available from: doi:10.1007/978-3-319-66429-3_54 [Accessed 20 January 2020].

[22] Mikolov, T. et al. (2013) Efficient estimation of word representations in vector space. In: *Proceedings of Workshop at the International Conference on Learning Representations*. Scottsdale, USA.

[23] Nandi, N. R. et al. (2018) Bangla News Recommendation Using doc2vec. In: *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pp. 1–5. Available from: doi:10.1109/ICBSLP.2018.8554679 [Accessed 20 January 2020].

[24] Novotná, T. and Harašta, J. (2019) *The Czech Court Decisions Corpus (CzCDC): Availability as the First Step*. ArXiv:1910.09513. [online] Available from: http://arxiv.org/abs/1910.09513 [Accessed 20 January 2020].

[25] Novotný, J. and Ircing, P. (2018) The Benefit of Document Embedding in Unsupervised Document Classification. In: Karpov, A. et al. (eds.). *Speech and Computer. Cham: Springer International Publishing*, pp. 470–478. Lecture Notes in Computer Science. Available from: doi:10.1007/978-3-319-99579-3_49 [Accessed 20 January 2020].

[26] Peczenik, A. (1997) The Binding Force of Precedent. In: MacCormick, N., Summers, R. S. (eds.). *Interpreting Precedents. A Comparative Study*. Dartmouth: Aldeshot, pp. 461–479.

[27] Ranera, L. T. B., Solano, G. A. and Oco, N. (2019) Retrieval of Semantically Similar Philippine Supreme Court Case Decisions using Doc2Vec. In: *2019 International Symposium on Multimedia and Communication Technology (ISMAC)*. IEEE, pp. 1–6.

[28] Renjit, S. and Idicula, S. M. (2019) CUSAT NLP@AILA-FIRE2019: Similarity in Legal Texts using Document Level Embeddings. In: Bhattacharya, P. et al. *Overview of the FIRE 2019 AILA track: Artificial Intelligence for Legal Assistance*. Proc. of FIRE, pp. 12–15.

[29] Savelka, J. and Ashley, K. D. (2018) Segmenting U.S. Court Decisions into Functional and Issue Specific Parts. In: Palmirani, M. (ed.). *Legal Knowledge and Information Systems JURIX 2018*. IOS Press Ebooks, pp. 111–120. Available from: http://ebooks.iospress.nl/volume/legal-knowledge-and-information-systems-jurix-2018-the-thirty-first-annual-conference [Accessed 20 January 2020].

[30] Schweighofer, E., Winiwarter, W. and Merkl, D. (1995) Information filtering: the computation of similarities in large corpora of legal texts. In: *Proceedings of the 5th international conference on Artificial intelligence and law*, p. 119–126.

[31] Smejkalová, T. (2019) Judikatura, nebo precedens? *Právník. Teoretický časopis pro otázky státu a práva*, 158 (9), pp. 852–864.

[32] Tian, X. et al. (2018) K-Means Clustering for Controversial Issues Merging in Chinese Legal Texts. In: Palmirani, M. (ed.). *Legal Knowledge and Information Systems JURIX 2018*. IOS Press Ebooks, pp. 215–219. Available from: http://ebooks.iospress.nl/volume/legal-knowledge-and-information-systems-jurix-2018-the-thirty-first-annual-conference [Accessed 20 January 2020].

[33] Trieu, L. Q., Tran, H. Q. and Tran, M.-T. (2017) News Classification from Social Media Using Twitter-based Doc2Vec Model and Automatic Query Expansion. In: *Proceedings of the Eighth International Symposium on Information and Communication Technology*. Nha Trang City, Viet Nam: Association for Computing Machinery, pp. 460–467. Available from: doi:10.1145/3155133.3155206 [Accessed 20 January 2020].

[34] Wagh, R. and Anand, D. (2017) Application of citation network analysis for improved similarity index estimation of legal case documents: A study. In: *2017 IEEE International*

Conference on Current Trends in Advanced Computing (ICCTAC), pp. 1–5. Available from: doi:10.1109/ICCTAC.2017.8249996 [Accessed 20 January 2020].