

Keyphrase extraction from Slovak court decisions

Dávid Varga, Šimon Horvát, Zoltán Szoplák, Ľubomír Antoni, Stanislav Krajčí, Peter Gurský and Laura Bachňáková Rózenfeldová

Pavol Jozef Šafárik University in Košice, Faculty of Science, Institute of Computer Science, Jesenná 5, 040 01 Košice, Slovakia

Abstract

Keyphrase extraction is a vital subtask of text summarization and comparison, through which we can obtain the most relevant set of words and phrases that describe the content of a given document. In this paper we test multiple approaches of unsupervised keyword extraction on a set of court decisions. These approaches are TF-IDF, YAKE! and a graph-based weighted PageRank algorithm. We combine these algorithms with a dictionary-based word embedding method in order to capture the semantic relationships between the potential keyphrases. Extracted keyphrases can be used for semantic indexing of court decisions, which can help with finding decisions with similar content.

Keywords

keyphrase, keyword, extraction, legal text, word network, embedding, court decision

1. Introduction

In their decision-making, judges need to ensure the consistency of decisions with the standard practice of courts. Getting an overview of similar relevant court decisions is a time-consuming process. Currently, available tools have limited options for filtering a set of all decisions, often resulting in an extensive collection of documents. In the Slovak court system, only the Supreme Court has an analytical department that has human resources to create overviews of relevant court decisions for judges. With a vast number of court cases, common judges often do not have time and resources to get to all relevant documents, which can cause essential decisions to be overlooked by judges. The analytical department of the Supreme Court manually creates metadata to all Supreme Court decisions, including keyphrases, to speed up the overview-making process, especially by narrowing the search results down to a reasonable size. Automatic keyphrase extraction can help with manual annotation by providing hints, thus making the annotation process semi-automatic and faster. This increases the number of court decision annotations that can be used for searching and filtering.

In the field of natural language processing, automatic keyphrase extraction can be used as a form of text summarization. Manually extracting keyphrases consists of reading the whole document, understanding its content

and selecting the phrases used, or generating phrases that aptly describe the document. Manual extraction of keyphrases from long texts or from a large number of texts is time-consuming and demanding on human resources. These are the reasons why it is appropriate to automate this process. The process of automated extraction of keyphrases consists of selecting candidate phrases from a document or external source, which are evaluated according to how well they describe the document. An evaluation algorithm is used to evaluate the candidate phrases, which calculates the score according to statistics, semantics, or both at the same time. The candidate phrases with the highest score are then selected as keyphrases.

Keyphrase extraction algorithms are divided into two main groups, supervised and unsupervised algorithms. We can train supervised algorithms on a labeled dataset, while the resulting models often achieve high accuracy [1]. If a dataset that is labeled with keyphrases is not available, it is advisable to use unsupervised algorithms. These types of algorithms usually use statistical metrics that take into account the number of occurrences of phrases, the co-occurrence of phrases, the position of phrases within the document and others. These algorithms are often combined with graph algorithms, word embeddings, or other language models.


In this article, we will focus on the extraction of keyphrases from Slovak court decisions. This dataset does not contain manually extracted keyphrases, therefore we decided to use a combination of unsupervised statistical and semantic approaches.

The objectives of this article are:

- design and implementation of an algorithm for extracting keyphrases from Slovak court decisions;
- evaluation of the results of extracted keyphrases on a set of court decisions.

ITAT'22: Information technologies – Applications and Theory, September 23–27, 2022, Zuberec, Slovakia

✉ david.varga@student.upjs.sk (D. Varga);
simon.horvat@student.upjs.sk (Š. Horvát);
zoltan.szoplak@student.upjs.sk (Z. Szoplák);
lubomir.antoni@upjs.sk (L. Antoni); stanislav.krajci@upjs.sk
(S. Krajčí); peter.gursky@upjs.sk (P. Gurský);
laura.rozenfeldova@upjs.sk (L. B. Rózenfeldová)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

This article is organized into four sections. In Section 2, we describe the related approaches to automated keyphrase extraction and other works related to legal document processing. In Section 3, we propose the multiple algorithms to extract the keyphrases from Slovak court decisions. Finally, we analyze the results of the algorithms in Section 4.

2. Related works

A lot of research has been done on applying NLP techniques to law texts and court decisions. NLP techniques are used in different tasks, for example: predicting the outcomes of court decisions [2, 3, 4], searching for insufficiently reasoned court decisions [5], creating electronic versions of court decisions [6] or creating a collection of datasets for evaluating performance across different legal text understanding tasks [7].

An international voluntary association called the Free Access to Law Movement (FALM) [8] was founded in 1992 and has more than 60 member organisations from around the globe. FALM members provide free access to legal information, group legal documents into one place and analyse law texts. FALM member CanLII [9] uses software to process canadian court decisions. CanLII creates links to articles that are used in court decisions and to other court decisions used as citations. This software also creates a short description of the court decision and selects keyphrases. These can be used to save time and effort for legal experts such as judges and lawyers.

Algorithms used for legal text summarization are summed up in a survey paper [10]. However, in this section, we focus specifically on keyphrase extraction by use of statistical approaches and unsupervised algorithms. We also summarize the principles of selecting appropriate keyphrases based on observations.

The simplest approach to select keyphrases is to count the n-grams in the text and select the most common n-grams [11]. This approach is also called Bag of Words or BoW and does not take into account synonyms, grammar or the meaning of individual n-grams. The downside of using a BoW approach is that it does not select those keyphrases that are concise to the text and at the same time occur rarely in the text.

A significant improvement over the BoW method is TF-IDF [12]. TF-IDF takes into account the whole corpus and penalizes phrases that occur in many documents. It is often used as a baseline method or in one of the steps of an algorithm, for example KP-Miner [13] or Liu's clustering algorithm [14]. We will describe TF-IDF in more detail in the next chapter.

Three desirable properties of keyphrases are described in [14]:

- **Understandable.** Keyphrases should be easy to understand.
- **Relevant.** Keyphrases should relate to the main topic of the document.
- **Good coverage.** Keyphrases should cover all parts of the document appropriately.

According to these properties, the Liu's clustering algorithm [14] was created, which used statistical, semantic and clustering methods simultaneously. The first step of the algorithm was to search for candidate words. From these, keyphrases of several words will be composed in the next steps of the algorithm. Subsequently, the candidate phrases were calculated semantic closeness scores, according to their common occurrences within a fixed-length window and also according to an external source - Wikipedia. For each word, they created an embedding, where on each index of the vector, a value representing the relationship between the word and a specific article from Wikipedia was calculated using TF-IDF. Candidate words were clustered according to semantic closeness, which grouped semantically similar words into individual clusters. Subsequently, exemplary words representing the entire cluster were selected from individual clusters, which had to be extended to phrases composed of several words. The keyphrases for the document were selected so that the algorithm processed all the words of the document, and if the word type was a noun that was also an exemplary word, then the word was selected in the list of keyphrases along with adjectives in its neighbourhood in the original text.

One of the latest language-independent unsupervised keyphrase extraction algorithms is YAKE! [15]. It uses statistical information, such as word counts and word occurrences, to identify keyphrases in unstructured texts. Its great advantage is that it only works with the current document during extraction, so it is not necessary to have the whole corpus of similar texts or other text sources available. The algorithm consists of five steps: (1) preprocessing the document into a machine-readable format, which results in tagged individual words; (2) for each word, a representation is created consisting of a set of properties evaluated by statistical measurements; (3) the individual properties of the words are heuristically combined into one score, which represents the importance of the word; (4) generating n-grams from candidate words and assigning a degree of relevance; (5) deduplication of keyphrases that are too similar and ranking by relevance.

Another approach to extracting keyphrases is to use graphs and graph algorithms. The text may be represented by a graph such that the vertices of the graph are candidate phrases and the edges represent the relationship between these phrases. Subsequently, a value for each vertex of the graph is assigned using the selected

evaluation function, and the edges and their weights are used to calculate this value. Thus, the individual methods differ in the use of different types of graphs and evaluation functions. One of the first algorithms to extract keywords from a text that uses a graph is TextRank [16], which has inspired a number of other graph-based algorithms. Its evaluation function calculates the values for the vertices of the oriented graph recursively and the information at the input of this function is global, ie in each step, it comes from the whole graph. The evaluation function used is the PageRank [17] algorithm, which is iterative and its input is the oriented graph. PageRank was originally designed for scoring web pages by importance on the web, but in TextRank it is used to give score to candidate keyphrases.

RAKE [18] is another graph-based unsupervised algorithm and it uses word frequency and word co-occurrence to create a graph and assign scores to phrases. It needs a list of stop-words and delimiters at the input, but it is able to identify interior stop-words in phrases.

3. Methods

3.1. Background knowledge

In this section we describe how we mine knowledge from sources other than the document from which we want to extract key phrases. This background knowledge is used as weighting mechanism in methods described in the next section.

3.1.1. Term frequency – inverse document frequency.

TF-IDF is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. This measure is multiplication of two metrics:

1. term frequency expresses how many times a word appears in a document,
2. the inverse document frequency expresses how unique a given word is to a document. It is the frequency of the word across a set of all documents:

$$\text{idf}(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

where $|\{d \in D : t \in d\}|$ is the number of documents where the term t appears.

So idf examines the frequency values in all documents to reduce the impact of frequent words.

3.1.2. Phrase network.

To use keyphrase extraction methods it is in our best interest to develop a vocabulary of potential keyphrases. Keyphrase extraction methods such as TF-IDF prioritize keyphrases that are unique to a specific document that might not be suitable for the purposes of topic clustering. Using this vocabulary as a basis for creating phrase embeddings can help us semantically compare the keyphrases in order to facilitate better keyphrase selection.

Let V be the set of all unigrams, bigrams and trigrams used in court decisions documents¹. Relations among phrases in V are denoted as set of E . We mine these relations mainly from Slovak Law Thesaurus (SLT) [19] as follows:

1. Let phrase_i and phrase_j be words or phrases defined in SLT. In case that phrase_j occurs in definition of phrase_i , we expand our set E by the pair $(\text{phrase}_i, \text{phrase}_j)$.
2. Let phrase_i be word or phrase defined in SLT. Let $\{\text{phrase}_1, \text{phrase}_2, \dots, \text{phrase}_j\} \subseteq V$ be the words used in definition of phrase_i , but without definition in SLT. We add pairs $(\text{phrase}_i, \text{phrase}_1)$, $(\text{phrase}_i, \text{phrase}_2)$, \dots , $(\text{phrase}_i, \text{phrase}_j)$ to the set E . Not all words appearing in the definition are related to a defined phrase, therefore we weigh these relations with the TF-IDF used in our global weight function. The set of documents used for IDF calculation D is the set of all definitions from SLT.
3. Let $\text{phrase}_i \in V$ be a phrase that is not found in SLT. We find the definitions of individual words that make up the phrase in the Dictionary of Slovak language and continue as in the previous step. Let $\{\text{phrase}_1, \text{phrase}_2, \dots, \text{phrase}_j\} \subseteq V$ be the words used in definition of phrase_i . We add pairs $(\text{phrase}_i, \text{phrase}_1)$, $(\text{phrase}_i, \text{phrase}_2)$, \dots , $(\text{phrase}_i, \text{phrase}_j)$ to the set E . The set of documents used for IDF calculation D is the set of all definitions from Dictionary of Slovak language.

Using this set of definitions, we model a network of legal phrases defined as follows:

Let $G = (V, E, \phi)$ be a directed evaluated graph, where $\phi : E \rightarrow R$ is a function:

$$\phi(e) = \begin{cases} 1, & \text{if } e \text{ gained in 1} \\ \text{tf-idf}(\text{phrase}_i, \text{phrase}_j), & \text{if } e \text{ gained in } 2 \vee 3 \end{cases}$$

such that $e = (\text{phrase}_i, \text{phrase}_j)$, $\text{phrase}_i \in V$ is defined phrase, $\text{phrase}_j \in V$ is phrase occurring in definition of phrase_i and $e \in E$.

¹Vocabulary V does not contain stop words.

In the next step, we use the graph embedding techniques described in [20] which produce a semantic representation for each phrase from V . In our approach, we use the Node2Vec algorithm, described in [21] which is one of the graph embedding techniques based on a random walk. These vectors with semantic interpretation are used as background knowledge for the algorithms described below. A detailed description of the method for obtaining embeddings is described in [22].

Suppose we need embedding for a phrase² consisting of more than one word. We compute it as an element-wise average of all word embedding occurring in the phrases.

3.2. Weighted PageRank

In order to incorporate our vocabulary and embeddings, we can use a keyphrase selection method described in [23] in conjunction with our phrase embeddings.

First, we create an undirected weighted graph representing a given court decision, with each node corresponding to a phrase of the decision present in our vocabulary V . A pair of nodes v_1 and v_2 , each representing a potential keyphrase, will be connected by an edge if they are located within a fixed-size sliding window. The weight of these edges represents the similarity between the potential keyphrases that make up its nodes. This similarity is defined by two metrics. One of them is the dice coefficient which measures the interlinkedness of the two phrases. It is calculated as the number of times the phrases appear in the decision as a tuple, divided by the sum of frequencies of phrases individually:

$$\text{dice}(v_i, v_j) = \frac{2 \times \text{freq}(v_i, v_j)}{\text{freq}(v_i) + \text{freq}(v_j)} \quad (1)$$

where v_i and v_j are vertices connected by an edge, $\text{freq}(v_i)$ is the number of times the vertex v_i appears in the document, and $\text{freq}(v_i, v_j)$ is the frequency where the vertices v_i and v_j form a tuple, in whichever order.

The second metric is inspired by Newton’s law of universal gravitation. The frequencies of the phrases are used as the mass of the objects, and the distance is calculated as the cosine distance between the embeddings of the two phrases.

$$\text{attr}(v_i, v_j) = \frac{\text{freq}(v_i) \times \text{freq}(v_j)}{d(v_i, v_j)^2} \quad (2)$$

where $d(v_i, v_j)$ is the cosine between the embeddings of phrases v_i and v_j .

The weight of an edge is then calculated combining the attraction force and the dice coefficient:

²We already have embeddings for phrases defined in SLT. Here we talk about phrases from V (or unseen) that do not occur in any relation to E .

$$w_{ij} = \text{attr}(v_i, v_j) \times \text{dice}(v_i, v_j) \quad (3)$$

To extract keywords from the keywords of a graph, we will make use of the weighted PageRank algorithm. The PageRank algorithm is an iterative algorithm that calculates a score for each node of the graph, with a higher score indicating higher suitability as a keyphrase. The weighted PageRank algorithm ranks a node according to the rank of the sum of all its adjacent nodes, as well as the weights that connect them.

Then, the PageRank score is calculated, for each node of the graph recursively. The score at a given time step is calculated as:

$$P_t(v_i) = (1 - d) + d \times \sum_{v_j \in C(v_i)} \frac{w_{ij}}{\sum_{v_k \in C(v_i)} w_{jk}} P_{t-1}(v_j) \quad (4)$$

where $P_t(v_i)$ is the PageRank score for the node v_i at time t , $C(v_i)$ is the set of edges adjacent to node v_i , d is the dumping factor.

The results obtained from the PageRank algorithm can then be used to determine the most likely keyphrase candidates, with a higher score representing a more suitable keyphrase.

The issue with using the weighted PageRank algorithm on its own is that it works only with a given document, which makes it useful in extracting keyphrases that describe the text itself, but not what differentiates it from other texts. Since the texts are judicial decisions, many court-centric phrases would hinder our ability to differentiate court decisions by topic. Therefore the score for each phrase we obtained from the weighted PageRank was multiplied by its IDF score, calculated from all available court decisions as described by the TF-IDF metric. Multiplying the PageRank score by the IDF should favor keywords that are not as frequent and would therefore probably not be court-centric and thus more relevant to the specific topic of that decision.

3.3. Autoencoders

Keyword extraction methods like TF-IDF penalize phrases that are frequent in many documents, but infrequent phrases are not necessarily semantically informative. The task of removing court-centric phrases would be better achieved by using some form of semantic comparison. Phrases that are semantically dissimilar to the meaning of the majority of phrases are more likely to be keyphrases that can be used to meaningfully cluster documents. To perform semantic comparisons, we can combine our phrase embeddings with the autoencoder method.

Autoencoders, described in detail in [24] are unsupervised neural networks that aim to create a representation

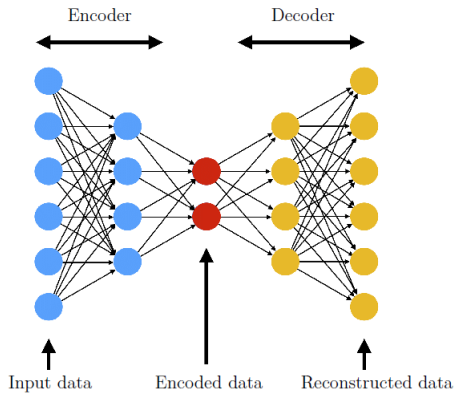


Figure 1: The scheme of autoencoder

of data that selects only the most relevant parameters, which can be used to reconstruct the original data. Autoencoders consist of two main parts: the encoder, which converts the input into an encoding (usually of lesser dimensionality than the input), and a decoder that tries to reconstruct the input from the encoding (Fig. 1). Using simple feedforward neural networks, the encoding h be calculated as:

$$h = \omega(Wx + b) \quad (5)$$

where x is the input, ω is the element-wise activation function, W is a weight matrix and b is the bias. This encoding can then be used to obtain x' , the reconstruction of the input. The reconstruction is calculated as:

$$x' = \omega'(W'h + b') \quad (6)$$

where ω' , W' and b' might be different from ω , W and b .

We have trained our autoencoder to reconstruct the embeddings of phrases of the vocabulary V , described in 3.1.2.³ Due to the vocabulary being made up primarily of phrases relevant to court decisions, we can infer that the reconstruction performance will be better with phrases explicitly related to court decisions. However, these phrases are detrimental to topic-based differentiation. Therefore by penalizing a high reconstruction success of a keyphrase, we can filter out those that are not relevant to the topic of that court decision. In our case, we multiplied the TF-IDF score of keyphrases with the cosine distance between the input embedding and the reconstructed embedding from the autoencoder:

$$\text{score}(v_i) = \text{tf-idf}(v_i) * \cos(\text{emb}(v_i), \text{rec}(\text{emb}(v_i))) \quad (7)$$

³Link to lemmatized court decisions. <https://bit.ly/3zUwbYA>

4. Evaluation

We have implemented two algorithms to serve as our baseline. The first is the regular TF-IDF metric used for keyword extraction, using all available court decisions to calculate the IDF value. This method is corpus-dependent, so other documents are taken into account. The second is the YAKE! algorithm [15], which takes into account only the current document. The algorithm described in 3.2 combines weighted PageRank with our phrase embeddings and multiplies the result by the IDF score of the TF-IDF metric. We will refer to this algorithm as WPR. The algorithm that multiplies regular TF-IDF score with cosine distance between and the algorithm described in 3.3 we labelled as AE.

Since we did not have access to extracted keyphrases of any court decisions, we have chosen five random court decisions for manual and expert evaluation. We have asked a legal expert to evaluate results in three ways:

- creation of abstracts that offer a brief summary of the content of the decisions (see figures 1 and 3),
- manual extraction of keyphrases from the decisions using dictionary of keyphrases used by the analytical department of the Supreme Court (see figures 1 and 3),
- the expert's opinion on the potential of the computed keyphrases to be included in the dictionary or to be used in any other way (see section 4.1).

We summarized the outputs of the algorithms into tables 2 and 4, where the rows are documents and the columns are algorithms. Each table cell consists of the top five keyphrases found by the given algorithm for the given document.

We have compared the computed key phrases with abstracts and manually extracted keyphrases. The phrases that are present in the abstract are highlighted in yellow. If the keyphrase matches the manually extracted keyphrase, it is highlighted by a black frame.

As we can see, the YAKE! algorithm provides many keyphrases that cannot be found in abstracts or manual keyphrases. This is due to the chosen keyphrases being too long and heavily related to the topic of judicial decisions that offer little in phrases of differentiating decisions from one another since the method is corpus-independent.

The weighted WPR algorithm multiplied by the IDF score performs quite a bit better, achieving good performance on documents 3 and 5, but is outclassed by the algorithms using TF-IDF as the basis of selection. This is likely because the WPR algorithm prefers phrases that are frequent and that are semantically similar to the other keyphrases, which is a good approach for general

keyphrase extraction; however those might not be well suited to clustering within a corpus.

TF-IDF on its own achieves good performance, as the metric is built for extracting phrases that are good unique descriptors of documents. It brings many matches on all of the documents, with the top five keyphrases being good topic descriptors for all documents.

The most abstract and manual keyphrase matches were achieved by the AE algorithm, combining TF-IDF with the reconstruction error of the autoencoder.

An interesting finding of all evaluated methods is that the resulting phrases are found mainly in abstracts and less among manually obtained phrases. We would also like to point out that several manually extracted phrases are not even in the abstracts themselves.

We asked a legal expert to weigh in on the results from her perspective. We present her statement in full in the next section.

4.1. Legal expert statement

The keyphrases selected by the analysis define the nature of the respective judicial decisions to varying degrees. In some cases, the selected keyphrases sufficiently characterize the decisions, e. g. as regards the second decision where it is clear that the decision regards the cancellation of the child support obligation. In other cases, the keyphrases extracted from the decisions' text describe the factual circumstances of the case rather than the relevant legal institutes applied in them or the legal process as such. To illustrate, the keyphrases describing the first decision focus on the factual background of the case, namely the asserting of warranty ("refund") for the services provided ("to train"), but do not specifically define the applicable legal institute (liability for defects), or the type of contract concluded between the parties to a dispute (framework agreement on cooperation), which would be most likely the keyphrases used by the legal expert to search for decisions in analogous cases. Similarly, it is unclear from the keyphrases characterizing other decisions examined what type of a decision is adopted (decision on the merits of the case or a procedural decision). To demonstrate, it is not apparent that the third decision regards the appellant's court reversal and referral of the decision of the court of the first instance, that in the fourth decision, the court discontinued the execution of a judgment or that the fifth decision approves the agreement on guilt and punishment (although in this case the phrase "approve the agreement" has been selected). This is, however, understandable, as these are all legal categories that may not be immediately identifiable from the decisions' text alone without previous legal input.

4.2. Summary and future work

This paper proposes and evaluates unsupervised keyword extraction methods because we lack labeled data as a proof of concept. We can conclude from the statement of a legal expert that the most relevant keyphrases are legal institutes and legal processes.

In our new project, we plan to cooperate with the Supreme Court of the Slovak Republic, in which we should be able to work with manually extracted phrases from their court decisions. This cooperation will allow us to design and test supervised keyword extraction methods and compare them with the methods presented in this paper. In our future work, we want to include laws and regulations cited by court decisions as a source of names of legal institutes.

5. Conclusion

In the article, we studied the problem of revealing keyphrases in the court decisions of the Slovak Republic. We proposed two unsupervised algorithms and evaluated them on five arbitrary court decisions. We have compared computed keyphrases with expert-written abstracts and manually extracted keyphrases. The results show that the methods extract keyphrases that are mainly included in abstracts rather than manually extracted keyphrases. The best results proposed the AE algorithm, combining TF-IDF with the reconstruction error of the autoencoder.

We believe that the results of the algorithms can be used as recommendations for manual annotation of court decisions with keyphrases if the intersection of found keyphrases with a dictionary of legal phrases is applied. It can also be used to enrich search results and expand filtering options.

Acknowledgement

This work was supported by the Slovak Research and Development Agency under contract No. APVV-21-0336 Analysis of court decisions by methods of artificial intelligence. This work was supported by the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic under contract VEGA 1/0177/21 Descriptive and computational complexity of automata and algorithms. This work was supported by the internal project at the Faculty of Science at Pavol Jozef Šafárik University in Košice vvgp-pf-2021-1789 Legal text analysis using computer linguistics.

References

- [1] E. Papagiannopoulou, G. Tsoumakas, A review of keyphrase extraction, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10 (2020) e1339.
- [2] M. Medvedeva, M. Vols, M. Wieling, Using machine learning to predict decisions of the european court of human rights, *Artificial Intelligence and Law* 28 (2020) 237–266.
- [3] N. Aletras, D. Tsarapatsanis, D. Preoŕiuc-Pietro, V. Lampos, Predicting judicial decisions of the european court of human rights: A natural language processing perspective, *PeerJ Computer Science* 2 (2016) e93.
- [4] D. Alghazzawi, O. Bamasag, A. Albeshri, I. Sana, H. Ullah, M. Z. Asghar, Efficient prediction of court judgments using an lstm+ cnn neural network model with an optimal feature set, *Mathematics* 10 (2022) 683.
- [5] D. Varga, Z. Szoplák, S. Krajci, P. Sokol, P. Gurský, Analysis and prediction of legal judgements in the slovak criminal (2021).
- [6] P. H. Luz de Araujo, T. E. de Campos, F. Ataiades Braz, N. Correia da Silva, VICTOR: a dataset for Brazilian legal documents classification, in: *Proceedings of the 12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, 2020, pp. 1449–1458. URL: <https://www.aclweb.org/anthology/2020.lrec-1.181>.
- [7] I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androustopoulos, D. M. Katz, N. Aletras, LexGLUE: A benchmark dataset for legal language understanding in english, *arXiv preprint arXiv:2110.00976* (2021).
- [8] The Free Access to Law Movement (FALM), 2022. URL: <http://falm.info/>.
- [9] The Canadian Legal Information Institute (CanLII), 2022. URL: <https://www.canlii.org/>.
- [10] A. Kanapala, S. Pal, R. Pamula, Text summarization from legal documents: a survey, *Artificial Intelligence Review* 51 (2019) 371–402.
- [11] Z. S. Harris, Distributional structure, *Word* 10 (1954) 146–162.
- [12] K. S. Jones, A statistical interpretation of term specificity and its application in retrieval, *Journal of documentation* (1972).
- [13] S. R. El-Beltagy, A. Rafea, Kp-miner: A keyphrase extraction system for english and arabic documents, *Information systems* 34 (2009) 132–144.
- [14] Z. Liu, P. Li, Y. Zheng, M. Sun, Clustering to find exemplar terms for keyphrase extraction, in: *Proceedings of the 2009 conference on empirical methods in natural language processing*, 2009, pp. 257–266.
- [15] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, A. Jatowt, Yake! keyword extraction from single documents using multiple local features, *Information Sciences* 509 (2020) 257–289.
- [16] R. Mihalcea, P. Tarau, Textrank: Bringing order into text, in: *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004, pp. 404–411.
- [17] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, *Computer networks and ISDN systems* 30 (1998) 107–117.
- [18] S. Rose, D. Engel, N. Cramer, W. Cowley, Automatic keyword extraction from individual documents, *Text mining: applications and theory 1* (2010) 10–1002.
- [19] Slovak law thesaurus, Legislative and information portal, Ministry of Justice of the Slovak Republic (2022). URL: <https://www.slov-lex.sk/zoznam-tezaurov>.
- [20] P. Goyal, E. Ferrara, Graph embedding techniques, applications, and performance: A survey, *Knowl. Based Syst.* 151 (2018) 78–94.
- [21] A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016).
- [22] S. Horvát, S. Krajci, L. Antoni, Semantic representation of slovak words, *CEUR Workshop Proceedings Vol-2718* (2020).
- [23] R. Wang, Corpus-independent generic keyphrase extraction using word embedding vectors, *Software engineering research conference Vol. 39* (2014).
- [24] D. Bank, N. Koenigstein, R. Giryes, Autoencoders, *CoRR abs/2003.05991* (2020). URL: <https://arxiv.org/abs/2003.05991>. arXiv: 2003.05991.

Table 1

Abstracts from court decisions and manually extracted keyphrases by legal expert translated to English.

No.	Abstract	Manually extracted keyphrases
1	The complainant (lector) demanded via judicial proceedings that the defendant pays the full price of the in-voice for the services provided (realization of professional training). The defendant, who was the complainant's customer, paid the invoice only in part (liability for delay) due to considering the services provided by the complainant to be of poor quality (liability for defects). The defendant has also demanded a refund.	contract, liability for defects, liability, default, client, innominate contract, warranty, service, action
2	The complainant demanded the court to cancel the duty to support and maintain against the two defendants, who graduated from high school, are legal adults who are able to earn a living wage. The defendants agreed with the cancellation of the duty to support and maintain.	alimony, duty to support and maintain
3	The complainant applied a bill of exchange against the defendant, which was rejected by the district court. The reasoning of rejection was the fact that the district court called for the complainant to fill in additional data in to the proposal form, which the complainant did not do. The court of appeals ruled in favour of the complainant, affirming that he did not need to fill in his proposal with additional data. The first instance court arrived at the decision by applying incorrect legislation and incorrect interpretation of the legislation and EU rights.	bill of exchange, claim, commercial paper, appeal, referral, reversing decision
4	The court rejected the proposal of granting authorization to a court distrainor and stopped all distraint proceedings. The court didn't assign the distraint expenses to the court distrainor.	discontinue distraint, distraint proceedings, distraint, court distrainor
5	The accused was negligently driving a motor vehicle, not paying attention to the traffic situation on the road and did not give way to a crossing pedestrian. A collision occurred, where the pedestrian suffered injuries consisting of multiple bone fractures and internal bleeding. The accused inflicted grievous bodily harm to the pedestrian due to negligence, due to which the accused was charged with inflicting injury. The accused was received a fine had their driving license revoked from all types of motor vehicles and she entered a plea agreement.	bodily harm, agreement on guilt and punishment, negligence, punishment, criminal offence, punishment by disqualification

Table 2

Top 5 keyphrases translated to English language.

No.	TF-IDF	YAKE!	WPR	AE
1	to train customer lector project studies	according to the PRINCE methodology between the participants of the proceedings PRINCE methodology training according to the commercial law section participants of the proceedings was	to train lector trainer accreditation studies	to train customer lector email refund
2	studies duty to support and maintain support and maintain to work court of Námestovo	district court of Námestovo by the judgment of the district court on the basis of an employment contract to support according to the paragraph he finished high school studies	loader high school worker to take care of part-time job	duty to support and maintain court of Námestovo cancel the duty to support and maintain contract of employment obligation towards
3	bill of exchange form first instance court first instance fill out	low value of the dispute to apply the claim of the court to apply the claim the first instance court in connection to the court of appeals	assumption receiving bill of exchange form stage	bill of exchange the first instance court to apply the claim form of application owner of the bill of exchange
4	court distrainor Dolný Kubín Dolný to grant authorization to grant	court of Dolný Kubín first instance court district court of Dolný Kubín apartment Dolný Kubín Dolný Kubín case reference	Dolný Kubín court distrainor Dolný to apply to instruct case reference	court distrainor Dolný Kubín to grant a warrant court court expenses of distraint
5	penalty guilt bone to charge fracture	by paragraph paragraph paragraph paragraph letter health by paragraph months by paragraphs Euro by paragraph	pedestrian pedestrian crossing shovel bone lane	road traffic fracture bone penalty approve the agreement

Table 3

Abstracts from court decisions and manually extracted keyphrases by legal expert in Slovak.

No.	Abstract	Manually extracted keyphrases
1	Navrhovateľ (lektor) sa súdnym konaním domáhal, aby odporca uhradil faktúru za poskytnuté služby (realizácia odborných školení) v plnej výške. Odporca, ktorý bol zákazníkom navrhovateľa, uhradil faktúru iba čiastočne (zodpovednosť za omeškanie) kvôli tomu, že navrhovateľ podľa neho poskytol vadné služby (zodpovednosť za vady). Navrhovateľ taktiež podal reklamáciu.	zmluva, zodpovednosť za vady, zodpovednosť, omeškanie, objednávateľ, nepomenovaná zmluva, reklamácia, služba, žaloba
2	Navrhovateľka žiadala, aby súd zrušil jej vyživovaciu povinnosť voči dvom odporcom, ktorí ukončili stredoškolské štúdium, sú plnoletí a zarábajú si sami na živobytie. Odporcovia súhlasili so zrušením vyživovacej povinnosti.	vyživné, vyživovacia povinnosť
3	Navrhovateľ si v návrhu uplatnil voči odporcovi pohľadávku, ktorú mu okresný súd zamietol. Dôvodom zamietnutia bol ten, že okresný súd vyzval navrhovateľa o doplnenie údajov prostredníctvom tlačiva na doplnenie návrhu, ktoré navrhovateľ nedoplnil. Odvolací súd dal navrhovateľovi za pravdu, teda že navrhovateľ nemusel dopĺňať svoj návrh o ďalšie údaje. Prvostupňový súd dospel k rozhodnutiu na základe aplikácie nesprávnych právnych predpisov a nesprávnej interpretácie príslušných právnych predpisov a práva EÚ.	zmenka, pohľadávka, cenné papiere, odvolanie, vrátenie vecí, zrušujúce rozhodnutie
4	Súd zamietol žiadosť o udelenie poverenia pre súdnu exekútorku a zastavil exekučné konanie. Súd exekútorku trovy exekúcie neprisúdil.	zastavenie exekúcie, exekučné konanie, exekúcia, exekútor
5	Obvinená viedla motorové vozidlo a nevenovala plnú pozornosť vedeniu vozidla. Nesledovala situáciu v cestnej premávke a nedala prednosť chodcovi prechádzajúceho cez priechod pre chodcov. Došlo k zrážke, pričom chodec utrpel poranenia pozostávajúce zo zlomením viacerých kostí a vnútorných krvácaní. Z nedbanlivosti spôsobila ťažkú ujmu na zdraví chodcovi, čím spáchala prečin ublíženia na zdraví. Obvinená dostala peňažný trest a trest zákazu činnosti viesť všetky druhy motorových vozidiel, pričom uzavrela dohodu o vine a treste.	ujma na zdraví, dohoda o vine a treste, nedbanlivosť, trest, trestný čin, trest zákazu činnosti

Table 4

Top 5 keyphrases in Slovak language.

No.	TF-IDF	YAKE!	WPR	AE
1	školiť zákazník lektor projekt štúdium	podľa metodiky PRINCE medzi účastníkmi konania školenia metodiky PRINCE podľa ods obchodného účastníkmi konania bola	školiť lektor školiť akreditácia štúdiá	školiť zákazník lektor email reklamácia
2	štúdium vyživovacia povinnosť vyživovací pracovať súd Námestovo	okresného súdu námestovo rozsudkom okresného súdu základe pracovnej zmluvy živiť podľa ods ukončil stredoškolské štúdium	nakladač stredoškolský robotník opatrovať brigáda	vyživovacia povinnosť súd námestovo zrušiť vyživovaciu povinnosť pracovná zmluva povinnosť voči
3	zmenka tlačivo prvostupňový súd prvostupňový vyplniť	nízkou hodnotou sporu uplatnenie pohľadávky súdu uplatnenie pohľadávky prvostupňový súd súvislosti odvolací súd	dohad príjímací zmenka tlačivo etapa	zmenka prvostupňový súd uplatniť pohľadávku tlačivo návrh majiteľ zmenky
4	súdna exekútorka Dolný Kubín dolný udelenie poverenia udelenie	súd Dolný Kubín súd prvého stupňa okresný súd dolný bytom Dolný Kubín dolný kubín spisová	Dolný Kubín súdna exekútorka Dolný uplatniť poučif spisová značka	súdna exekútorka dolný kubín udelenie poverenia súd súdny trovy exekúcie
5	trest vina kosť obviniť zlomenina	podľa ods ods ods pism zdraví podľa ods mesiacov podľa ods eur podľa ods	chodec priechod lopata kosť pruh	cestná premávka zlomenina kosť trest schválif dohodu