D09 – Method of assigning keyphrases (EN)

Project funded by the European Union Next GenerationEU through the Slovak Republic's Recovery and Resilience Plan under project no. 09I05-03-V02-00049.

Financované Európskou úniou NextGenerationEU
PLÁN OBNOVY
ÚRAD PODPREDSEDU VLÁDY SLOVENSKEJ REPUBLIKY PRE PLÁN OBNOVY A ZNALOSTNÚ EKONOMIKU
VAIA VÝSKUMNÁ A INOVAČNÁ AUTORITA
VÝSKUMNÁ AGENTÚRA

INTRODUCTION

The project "**Automation of Legal Text Analysis Based on Machine Learning**" (hereinafter referred to as "ALTAML") is a key initiative aimed at integrating innovative approaches in the field of data processing and subsequent analysis, with a focus on the field of law, which also includes information and communication technology law. The aim of this project is to develop and verify effective methods for the automated analysis of legal texts using machine learning techniques. This includes, in particular, the development of tools that facilitate the processing and analysis of large volumes of data in the form of various legal documents, as well as the extraction of relevant information (attributes) from such documents, including the identification of key terms, references to legislation and other attributes.

The aim of the ALTAML project is therefore to contribute to a more efficient approach to legal information and to speed up legal processes, thereby ensuring a higher degree of legal certainty and improving the accessibility of legal texts, the results of their analysis and relevant legal information for both professionals and the general public.

As part of the KPB2 work package, the output Method of assigning keyphrases was developed. The proposed system is a hybrid extractor of legal keyphrases from Slovak court decisions, combining statistical text processing methods, semantic vector representations and a large-scale language model. Extraction is limited to a pre-defined set of legal terms (see Output *D08-KPB2.4-Database of Slovak and English Legal Terms*), which ensures terminological accuracy and consistency of outputs.

Each candidate term is assigned a score calculated based on a combination of TF-IDF weighting and cosine similarity between the vector representation of the term and the decision itself. The score is further modified to take into account the occurrence of phrases in reference texts, such as referenced court decisions and cited legislation. This approach allows for the broader legal context of the document to be taken into account.

The final selection of key terms is carried out using a locally deployed language model, which ensures the removal of synonymous and duplicate terms and gives preference to legally established terms. The result is a sorted set of key terms that accurately represents the content and legal substance of the analysed decision. A more concise description can be found below.

Hybrid extractor of legal key terms from court decisions

The key term extractor is a software module implemented in Python, designed to automatically identify legally relevant terms from Slovak court decisions. The system combines statistical methods of natural language processing, semantic vector representations and a large language model (LLM) to achieve descriptive and concise outputs. Extraction is limited to a predefined set of legal terms, minimising the risk of generating inaccurate or inconsistent expressions.

System functionality

The main functionality of the extractor consists of the following steps:

1. Input data preparation

The system works with a structured corpus of court decisions, whereby the following can be assigned to each decision:

- texts of referenced decisions (dockets),
- texts of cited legal regulations.

The texts are normalised by simple cleaning (removal of redundant spaces) and aggregated into a uniform text representation of the document.

1.1. Definition of candidate terms

Candidate key terms are loaded from an external JSON file (see Output *D08-KPB2.4 Database of Slovak and English legal terms*). Each term may also contain one or more definitions that are used to enrich its semantic representation. This approach allows for a better understanding of the meaning of the term in the context of legal terminology.

1.2. Semantic representation of terms and documents

The *SlovakBERT* language model, fine-tuned for the task of semantic similarity of sentences, is used to calculate vector representations of documents and candidate terms. For each candidate term, a single embedding is created, combining the term name itself and its definitions.

2. Extraction and scoring of key terms

2.1. Statistical component (TF-IDF)

A TF-IDF score is calculated for each document, but only over the set of candidate terms. This ensures that the system evaluates only legally relevant terms and ignores the rest of the document's lexical content. Slovak stop words are explicitly defined and removed from the analysis.

2.2. Semantic similarity

For each candidate term, the cosine similarity between its embedding and the embedding of the entire document is calculated. This similarity represents the degree of semantic relatedness of the term to the content of the decision.

2.3. Consideration of references

The scores of the terms are further modified based on their occurrence in reference texts:

- related court decisions,
- cited legal regulations.

The increase in the score is normalised by the number of reference texts, ensuring comparability between documents with different numbers of citations.

2.4. Combined score

The final term score is calculated as a linear combination of the TF-IDF score and semantic similarity, with a weighting parameter allowing the influence of both components to be adjusted.

3. Selection and reduction of candidates

From the total set of candidate terms, a limited number of the best-scoring terms are selected to proceed to the final processing stage. This step reduces computational complexity and eliminates low-relevance terms.

Optionally, the maximum marginal relevance (MMR) principle can be used to ensure that the selected terms are not only highly relevant but also semantically diverse.

4. Reranking using a language model

The final selection of key terms is performed using a locally deployed large language model (via the Ollama tool). The model is controlled by a strict prompt formulated in Slovak, which ensures:

- selection exclusively from the specified candidates,
- the removal of synonymous and duplicate terms,
- preference for legally established and terminologically accurate terms,
- deterministic output in the form of a JSON array without additional text.

This step allows the simulation of expert decision-making by a lawyer when selecting key terms.

System output

The system output is a JSON structure that contains a sorted list of key terms for each decision identifier, together with their final scores. The order of the terms reflects their overall relevance after taking into account statistical, semantic and reference factors.

Requirements and technologies

The system is built on Python and uses the following technologies:

- TF-IDF (scikit-learn) for statistical text analysis,
- Sentence Transformers for semantic embeddings,
- cosine similarity for vector comparison,
- local LLM (LLaMA 3.3) for final reranking.

32GB RAM or more is recommended.
GPU (CUDA) support enables efficient processing of large legal corpora.

Source files

The project source code and data file are maintained in a repository and are available at: https://gitlab.science.upjs.sk/altaml/keyphrase-extractor