D5 - Method for the extraction of references

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

INTRODUCTION

The project **"Automation of Legal Text Analysis Based on Machine Learning"** (hereinafter referred to as **"ALTAML"**) represents a key initiative aimed at integrating innovative approaches in the field of data processing and subsequent analysis, specifically within the legal domain, which also includes the law of information and communication technologies. The objective of this project is to develop and validate 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 other legal acts, and other attributes.

The ALTAML project thus aims to contribute to a more efficient access to legal information and to the acceleration of 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.

Within Work Package KPB2, **Deliverable D5 – Method for the Extraction of References** was developed. This document presents notes on the automatic extraction of references to Slovak and European legal acts and previous court decisions. The methods combine rule-based approaches (regular expressions, Levenshtein distance, fuzzy search) with Named Entity Recognition (NER) using the SlovakBERT model. The system enables the creation of hypertextual and semantic links between legal documents, thereby supporting the long-term goal of the project – building a connected and searchable ecosystem of Slovak and European legal texts.

1. ALTAML referex

The ALTAML Referex project provides automated extraction of references to:

- slovak legislation,
- past Slovak court decisions,
- decisions of the European Court of Human Rights (ECHR),
- decisions of the Court of Justice of the European Union (CJEU),
- decisions of the Office for Personal Data Protection of the Slovak Republic.

It can be deployed as a **REST API** service for extracting references from legal texts.

MongoDB setup

Create a MongoDB database named lawDB and add the following collections:

- law validities,
- law_aliases,
- law_aliases_lemmatized,
- codebook_courts,
- zakony_aliases,
- dict_lemmatized_laws.

Then import data into each collection from the files located in the mongo_collections folder. If needed, adjust database connection URLs in the config.yaml file.

Dependencies for ECHR Extraction

Install the required dependencies using the following commands:

python -m pip install --upgrade pip setuptools wheel
pip install torch torchvision torchaudio --index-url
https://download.pytorch.org/whl/cpu

REST-API

Install **uvicorn**:
pip install uvicorn

In your IDE or terminal, configure the Python run configuration as follows:

- **Module name:** uvicorn
- **Parameters:** rest:app --reload --port 8001
- **Working directory:** YOUR_PATH_TO_PROJECT/altaml-referex/rest

Once the service is running, you can use the following endpoints:

- **Extract references to Slovak legislation**
  Endpoint: /v1/referex/laws
  Request model: ExtractLawsRequest (defined in law_extractor_rest.py)
- **Extract references to court decisions**
  Endpoint: /v1/referex/dockets
  Request model: ExtractDocketsFromDecisionsTextRequest (defined in docket_extractor_rest.py)

2. Tvaroslovnik-API

This project uses the **Tvaroslovnik API** for lemmatization purposes. To enable high-performance mode, please contact us to obtain an **API key**.

3. Source files

The project's source code is maintained in the internal UPJŠ GitLab repository. It can be accessed at the following address: https://gitlab.science.upjs.sk/altaml/referex