D5 - Metóda pre extrakciu odkazov

Projekt financovaný Európskou úniou Next GenerationEU prostredníctvom Plánu obnovy a odolnosti SR v rámci projektu č. 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 AKTIVITA
VÝSKUMNÁ AGENTÚRA

ÚVOD

Projekt „**Automatizácia analýzy právnych textov na základe strojového učenia**“ (ďalej len „**ALTAML**“) predstavuje kľúčovú iniciatívu zameranú na integráciu inovatívnych prístupov v oblasti spracovania údajov a následnej analýzy, so zameraní na oblasť práva, ktorá zahŕňa aj právo informačných a komunikačných technológií. Cieľom tohto projektu je vyvinúť a overiť účinné metódy automatizovanej analýzy právnych textov pomocou techník strojového učenia. To zahŕňa najmä vývoj nástrojov, ktoré uľahčujú spracovanie a analýzu veľkých objemov údajov vo forme rôznych právnych dokumentov, ako aj extrakciu relevantných informácií (atribútov) z takýchto dokumentov vrátane identifikácie kľúčových pojmov, odkazov na právne predpisy a ďalších atribútov.

Cieľom projektu ALTAML je teda prispieť k efektívnejšiemu prístupu k právnym informáciám a k zrýchleniu právnych procesov, čím sa zabezpečí vyššia miera právnej istoty a zlepší sa dostupnosť právnych textov, výsledkov ich analýzy a relevantných právnych informácií pre odbornú aj širokú verejnosť.

V rámci pracovného balíka KPB2, bola vyvinutá **D5 – Metóda pre extrakciu odkazov.** Tento dokument prezentuje poznatky o automatickej extrakcii odkazov na slovenské a európske právne predpisy, ako aj na predchádzajúce súdne rozhodnutia. Použitá metóda kombinuje pravidlami riadené prístupy (regulárne výrazy, Levenshteinovu vzdialenosť, fuzzy vyhľadávanie) s rozpoznávaním pomenovaných entít (NER) založeným na modeli SlovakBERT. Systém umožňuje vytváranie hypertextových a sémantických prepojení medzi právnymi dokumentmi, čím podporuje dlhodobý cieľ projektu – budovanie prepojeného a vyhľadateľného ekosystému slovenských a európskych právnych textov.

1. ALTAML Referex

Projekt ALTAML Referex poskytuje automatizovanú extrakciu odkazov na:

- slovenské právne predpisy,
- predchádzajúce rozhodnutia slovenských súdov,
- rozhodnutia Európskeho súdu pre ľudské práva (ESĽP),
- rozhodnutia Súdneho dvora Európskej únie (SDEÚ),
- rozhodnutia Úradu na ochranu osobných údajov Slovenskej republiky.

Tento projekt je možné nasadiť ako REST API službu určenú na extrakciu referencií z právnych textov.

Nastavenie databázy MongoDB

Vytvorte databázu MongoDB s názvom lawDB a pridajte do nej nasledujúce kolekcie:

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

Následne naimportujte dáta do každej kolekcie zo súborov nachádzajúcich sa v priečinku mongo_collections. V prípade potreby upravte URL pripojenia k databáze v súbore config.yaml.

Závislosti pre extrakciu rozhodnutí ESĽP

Nainštalujte požadované závislosti pomocou nasledujúcich príkazov:

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

REST-API

Nainštalujte **uvicorn** príkazom:
pip install uvicorn

Vo vašom vývojovom prostredí (IDE) alebo termináli nastavte konfiguráciu spustenia Pythonu nasledovne:

- **Module name(názov modulu):** uvicorn
- **Parameters(parametre):** rest:app --reload --port 8001
- **Working directory(pracovný adresár):** CESTA_K_VASMU_PROJEKTU/altaml-referex/rest

Po spustení služby môžete použiť nasledujúce endpointy:

- **Extrakcia odkazov na slovenské právne predpisy**
  Endpoint: /v1/referex/laws
  Model požiadavky: ExtractLawsRequest (definovaný v law_extractor_rest.py)
- **Extrakcia odkazov na súdne rozhodnutia**
  Endpoint: /v1/referex/dockets
  Model požiadavky: ExtractDocketsFromDecisionsTextRequest (definovaný v docket_extractor_rest.py)

2. Tvaroslovnik-API

Tento projekt využíva **Tvaroslovník API** na účely lematizácie. Kontaktujte nás prosím pre získanie **API kľúča**, ktorý je potrebný pre vysokovýkonný režim.

3. Zdrojové súbory

Zdrojový kód projektu je spravovaný v internom repozitári UPJŠ GitLab. Je dostupný na nasledujúcej adrese: https://gitlab.science.upjs.sk/altaml/referex