D09 – Metóda prideľovania kľúčových pojmov (SK)

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 AUTORITA
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 bol vypracovaný výstup Metóda prideľovania kľúčových pojmov. Navrhnutý systém predstavuje hybridný extraktor právnych kľúčových pojmov zo slovenských súdnych rozhodnutí, ktorý kombinuje štatistické metódy spracovania textu, sémantické vektorové reprezentácie a jazykový model veľkého rozsahu. Extrakcia je obmedzená na vopred vytvorenú množinu právnych pojmov (viď. Výstup *D08-KPB2.4-Databáza slovenských a anglických právnych pojmov*), čím je zabezpečená terminologická presnosť a konzistentnosť výstupov.

Každému kandidátnemu pojmu je priradené skóre vypočítané na základe kombinácie TF-IDF váženia a kosínovej podobnosti medzi vektorovou reprezentáciou pojmu a samotného rozhodnutia. Skóre je ďalej modifikované s ohľadom na výskyt pojmov v referenčných textoch, ako sú odkazované súdne rozhodnutia a citované právne predpisy. Tento prístup umožňuje zohľadniť širší právny kontext dokumentu.

Finálna selekcia kľúčových pojmov je realizovaná prostredníctvom lokálne nasadeného jazykového modelu, ktorý zabezpečuje odstránenie synonymických a duplicitných výrazov a uprednostňuje právne ustálené pojmy. Výsledkom je zoradená množina kľúčových pojmov, ktorá verne reprezentuje obsah a právnu podstatu analyzovaného rozhodnutia. Stručnejší popis nájdete nižšie.

Hybridný extraktor právnych kľúčových pojmov zo súdnych rozhodnutí

Extraktor kľúčových pojmov je softvérový modul implementovaný v jazyku Python, navrhnutý na automatickú identifikáciu právne relevantných pojmov zo slovenských súdnych rozhodnutí. Systém kombinuje štatistické metódy spracovania prirodzeného jazyka, sémantické vektorové reprezentácie a veľký jazykový model (LLM) na dosiahnutie opisných a výstižných výstupov. Extrakcia je obmedzená na vopred definovanú množinu právnych pojmov, čím sa minimalizuje riziko generovania nepresných alebo neustálených výrazov.

1. Funkcionalita systému

Hlavná funkcionalita extraktora pozostáva z nasledujúcich krokov:

Príprava vstupných dát

Systém pracuje so štruktúrovaným korpusom súdnych rozhodnutí, pričom ku každému rozhodnutiu môžu byť priradené:

- texty referencovaných rozhodnutí (dockets),
- texty citovaných právnych predpisov.

Texty sú normalizované jednoduchým čistením (odstránenie nadbytočných medzier) a agregované do jednotnej textovej reprezentácie dokumentu.

Definícia kandidátnych pojmov

Kandidátne kľúčové pojmy sú načítané z externého JSON súboru (viď. Výstup *D08-KPB2.4-Databáza slovenských a anglických právnych pojmov*). Každý pojem môže obsahovať aj jednu alebo viac definícií, ktoré sa používajú na obohatenie jeho sémantickej reprezentácie. Tento prístup umožňuje lepšie zachytiť význam pojmu v kontexte právnej terminológie.

Sémantická reprezentácia pojmov a dokumentov

Na výpočet vektorových reprezentácií dokumentov aj kandidátnych pojmov je použitý jazykový model *SlovakBERT* doladený na úlohu sémantickej podobnosti viet. Pre každý kandidátny pojem sa vytvorí jeden embedding kombinujúci samotný názov pojmu a jeho definície.

2. Extrakcia a skórovanie kľúčových pojmov

Štatistická zložka (TF-IDF)

Pre každý dokument sa vypočíta TF-IDF skóre, avšak výhradne nad množinou kandidátnych pojmov. Tým sa zabezpečí, že systém hodnotí len právne relevantné termíny a ignoruje ostatný lexikálny obsah dokumentu. Slovenské stopslová sú explicitne definované a odstránené z analýzy.

Sémantická podobnosť

Pre každý kandidátny pojem sa vypočíta kosínová podobnosť medzi jeho embeddingom a embeddingom celého dokumentu. Táto podobnosť reprezentuje mieru významovej príbuznosti pojmu k obsahu rozhodnutia.

Zohľadnenie referencií

Skóre pojmov je ďalej modifikované na základe ich výskytu v referenčných textoch:

- súvisiacich súdnych rozhodnutiach,
- citovaných právnych predpisoch.

Zvýšenie skóre je normalizované počtom referenčných textov, čím sa zabezpečí porovnateľnosť medzi dokumentmi s rôznym počtom citácií.

Kombinované skóre

Finálne skóre pojmu je vypočítané ako lineárna kombinácia TF-IDF skóre a sémantickej podobnosti, pričom váhový parameter umožňuje regulovať vplyv oboch zložiek.

3. Výber a redukcia kandidátov

Z celkovej množiny kandidátnych pojmov je vybraný obmedzený počet najlepšie skórovaných pojmov, ktoré postupujú do finálnej fázy spracovania. Tento krok znižuje výpočtovú náročnosť a zároveň eliminuje nízko relevantných pojmov.
Voliteľne je možné použiť princíp maximálnej marginálnej relevancie (MMR), ktorý zabezpečuje, že vybrané pojmy nie sú iba vysoko relevantné, ale aj významovo rozmanité.

4. Reranking pomocou jazykového modelu

Finálna selekcia kľúčových pojmov je realizovaná pomocou lokálne nasadeného veľkého jazykového modelu (prostredníctvom nástroja Ollama). Model je riadený prísnym promptom formulovaným v slovenskom jazyku, ktorý zabezpečuje:

- výber výhradne zo zadaných kandidátov,

- odstránenie synonymických a duplicitných výrazov,
- preferenciu právne ustálených a terminologicky presných pojmov,
- deterministický výstup vo forme JSON poľa bez dodatočného textu.

Tento krok umožňuje simulovať expertné rozhodovanie právnika pri výbere kľúčových pojmov.

5. Výstup systému

Výstupom systému je JSON štruktúra, ktorá pre každý identifikátor rozhodnutia obsahuje zoradený zoznam kľúčových pojmov spolu s ich finálnym skóre. Poradie pojmov reflektuje ich celkovú relevanciu po započítaní štatistických, sémantických a referenčných faktorov.

6. Požiadavky a technológie

Systém je postavený na jazyku Python a využíva nasledujúce technológie:

- TF-IDF (scikit-learn) pre štatistickú analýzu textu,
- Sentence Transformers pre sémantické embeddingy,
- kosínovú podobnosť na porovnanie vektorov,
- lokálny LLM (LLaMA 3.3) pre finálny reranking.

Odporúča sa RAM 32GB a viac.
Podpora GPU (CUDA) umožňuje efektívne spracovanie rozsiahlych právnych korpusov.

7. Zdrojové súbory

Zdrojový kód projektu a dátový súbor sú udržiavané v repozitári a sú dostupné na adrese: https://gitlab.science.upjs.sk/sk-lii/keyphrase-extractor-new