D19 – Algoritmus pre prepájanie právnych textov (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 AKCELERAČNÁ 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 KPB4 bol vypracovaný výstup Algoritmus pre prepájanie právnych textov. Systém na prepájanie právnych dokumentov je softvérový modul implementovaný v jazyku Python, navrhnutý tak, aby – po zadaní zdrojového dokumentu – vyhľadal najpodobnejšie slovenské súdne rozhodnutia. Metóda kombinuje štatistické metriky, sémantické embeddingy a analýzu štruktúrovaných právnych údajov s cieľom identifikovať dokumenty s podobným právnym obsahom, odôvodnením alebo skutkovým základom.

Systém využíva hybridný prístup založený na kombinácii:

- sémantickej podobnosti textu,
- podobnosti jednotlivých obsahových častí dokumentu,
- zhody právnych referencií,
- zhody citovaných právnych predpisov,
- podobnosti extrahovaných kľúčových pojmov.

Výsledkom je zoznam dokumentov zoradených podľa konečného skóre podobnosti so zdrojovým dokumentom.

Funkcionalita systému

Hlavná funkcionalita systému pozostáva z nasledujúcich krokov:

1. Príprava vstupných dát

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

- text samotného rozhodnutia,
- odkazy na súvisiace súdne rozhodnutia (dockets),
- citované právne predpisy,
- extrahované kľúčové pojmy.

Text dokumentu je normalizovaný jednoduchým čistením, ktoré odstraňuje nadbytočné medzery a zjednocuje formátovanie textu.

1.1. Segmentácia dokumentov

Každé rozhodnutie je rozdelené na menšie textové segmenty („chunky"). Segmentácia je realizovaná na úrovni odsekov, pričom systém zachováva prirodzenú štruktúru právneho textu. Jednotlivé odseky sú agregované do chunkov približne rovnakej veľkosti.Tento prístup umožňuje analyzovať podobnosť nielen na úrovni celého dokumentu, ale aj medzi konkrétnymi časťami právnej argumentácie.

1.2. Filtrovanie procesného a formulárového textu

Pred samotným výpočtom podobnosti systém identifikuje a odstraňuje procesné alebo formulárové časti textu, ktoré sa v súdnych rozhodnutiach pravidelne opakujú.

Filtrované sú najmä texty obsahujúce:

- označenia súdov,
- procesné poučenia,
- štandardizované formulácie rozhodnutí,
- opakujúce sa právnické frázy.

Odstraňujú sa aj veľmi krátke textové segmenty, ktoré nenesú dostatočný významový obsah.

Tento krok znižuje vplyv všeobecného procesného jazyka na výsledné skóre podobnosti a zvyšuje dôraz na vecný obsah rozhodnutí.

2. Sémantická reprezentácia dokumentov

Na výpočet vektorových reprezentácií dokumentov a textových segmentov je použitý jazykový model SlovakBERT doladený na úlohu sémantickej podobnosti viet.

Pre každý dokument sa vytvára:

- globálny embedding reprezentujúci celý dokument,
- embeddingy jednotlivých chunkov.

Embeddingy sú normalizované a následne porovnávané pomocou kosínovej podobnosti.

3. Výpočet podobnosti dokumentov

3.1. Globálna sémantická podobnosť

Systém vypočíta kosínovú podobnosť medzi embeddingmi celých dokumentov. Táto metrika reprezentuje všeobecnú tematickú a významovú podobnosť rozhodnutí.

3.2. Podobnosť textových segmentov

Pre všetky chunky oboch dokumentov sa vypočíta matica kosínových podobností. Systém následne identifikuje najpodobnejšie dvojice chunkov medzi dokumentmi. Podobnosť je vyhodnocovaná symetricky v oboch smeroch:

- pre každý chunk prvého dokumentu sa hľadá najlepší zodpovedajúci chunk druhého dokumentu,
- pre každý chunk druhého dokumentu sa hľadá najlepší zodpovedajúci chunk prvého dokumentu.

Do výsledného skóre vstupujú iba najrelevantnejšie podobnosti nad definovaným prahom. Finálne chunk skóre vzniká ako priemer najlepších zhôd.

Tento prístup umožňuje zachytiť lokálne podobné právne argumentácie alebo skutkové okolnosti aj v prípade, že sa celkové dokumenty výrazne líšia rozsahom alebo štruktúrou.

4. Porovnanie právnych referencií

Systém porovnáva odkazy na súvisiace súdne rozhodnutia pomocou Jaccardovej similarity množín identifikátorov.

Vyššie skóre získavajú dokumenty, ktoré odkazujú na podobnú judikatúru alebo rovnaké právne prípady.

5. Porovnanie citovaných právnych predpisov

Citované právne predpisy sú reprezentované ako množiny identifikátorov alebo

URL adries právnych noriem. Ich podobnosť je taktiež vypočítaná pomocou Jaccardovej similarity, čím systém zohľadňuje zhodu právneho základu rozhodnutí.

6. Porovnanie kľúčových pojmov

Extrahované kľúčové pojmy sú porovnávané pomocou váženého prekryvu, ktorý zohľadňuje:

- prítomnosť rovnakých pojmov,
- ich relatívnu významnosť v dokumentoch.

Tento prístup umožňuje lepšie zachytiť tematickú podobnosť dokumentov aj pri odlišnej formulácii textu.

7. Kombinované skóre podobnosti

Finálne skóre podobnosti dokumentov je vypočítané ako lineárna kombinácia viacerých komponentov:

- podobnosti právnych referencií,
- podobnosti citovaných právnych predpisov,
- podobnosti kľúčových pojmov,
- globálnej sémantickej podobnosti dokumentov,
- podobnosti textových segmentov.

Najväčšiu váhu má podobnosť chunkov, pretože najlepšie reprezentuje vecnú a argumentačnú podobnosť rozhodnutí.

8. Zoradený zoznam relevantných dokumentov

Pomocou opísanej metriky podobnosti porovnáme referenčný dokument so všetkými ostatnými dokumentmi v našom datasete, vypočítame podobnosť medzi každým z nich a vrátime zoradený zoznam identifikátorov na základe skóre podobnosti.

Použité technológie

Systém je implementovaný v jazyku Python a využíva nasledujúce technológie:

- Sentence Transformers pre sémantické embeddingy,
- SlovakBERT pre reprezentáciu slovenského právneho textu,
- cosine similarity pre porovnanie embeddingov,
- scikit-learn pre výpočet podobnosti a prácu s vektormi,

- NumPy pre numerické operácie.

Podpora GPU (CUDA) umožňuje efektívne spracovanie rozsiahlych korpusov súdnych rozhodnutí.

Zdrojové súbory

Zdrojový kód projektu a dátový súbor sú udržiavané v repozitári a sú dostupné na adrese: https://gitlab.science.upjs.sk/altaml/document-relevance-extractor