D10 – Poloautomatizovaná metóda na segmentáciu 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 KPB2 bol vypracovaný výstup Poloautomatizovaná metóda na segmentáciu právnych textov. Navrhnutý systém predstavuje hybridný segmentačný a klasifikačný model právnych textov, ktorý kombinuje štatistické metódy spracovania prirodzeného jazyka, sémantické vektorové reprezentácie, heuristické pravidlá a sekvenčný postprocessing. Cieľom systému je automaticky rozdeliť slovenské súdne rozhodnutia na logické obsahové časti a každej časti priradiť právne relevantný typ segmentu.

Systém pracuje viacúrovňovo. Najskôr vykonáva hrubú segmentáciu dokumentu na úrovni tematických blokov pomocou analýzy sémantických zmien medzi vetami. Následne aplikuje hybridný klasifikačný mechanizmus založený na kombinácii TF-IDF podobnosti, embeddingovej sémantickej podobnosti a pravidlových heuristík. V ďalšej fáze je realizovaný jemný vetný relabeling, ktorý umožňuje spresniť hranice jednotlivých segmentov a odstrániť lokálne nekonzistencie klasifikácie. Výsledkom je štruktúrované rozdelenie rozhodnutia na právne významové celky, ktoré reprezentujú jednotlivé fázy súdneho rozhodovania. Stručnejší popis systému je uvedený nižšie.

Hybridný systém segmentácie a klasifikácie slovenských súdnych rozhodnutí

Systém je implementovaný v jazyku Python a je navrhnutý na automatické spracovanie rozsiahlych korpusov slovenských súdnych rozhodnutí. Kombinuje metódy štatistického spracovania textu, transformerové jazykové modely a heuristické pravidlá vytvorené na základe štruktúry slovenských právnych dokumentov.

Funkcionalita systému

Hlavná funkcionalita systému pozostáva z nasledujúcich krokov:

1. Príprava vstupných dát

Systém pracuje s textami slovenských súdnych rozhodnutí uloženými v štruktúrovanom JSON formáte. Texty sú najskôr normalizované a čistené od technických artefaktov, nadbytočných medzier a špeciálnych markerov. Následne sú dokumenty rozdelené na jednotlivé vety pomocou pravidlovo orientovaného segmentátora prispôsobeného právnym textom.

1.1. Sémantická reprezentácia viet

Každá veta dokumentu je transformovaná na vektorovú reprezentáciu (embedding) pomocou transformerového jazykového modelu SlovakBERT doladeného na úlohu sémantickej podobnosti viet. Embeddingy slúžia ako základ pre identifikáciu tematických zmien v dokumente a následnú segmentáciu textu.

1.2. Detekcia hraníc segmentov

Na identifikáciu hraníc medzi obsahovými časťami rozhodnutia je použitá analýza sémantických vzdialeností medzi okolitými vetami. Systém vypočítava skóre tematickej zmeny medzi vetnými oknami a na základe lokálnych maxím určuje kandidátne hranice segmentov. Segmentácia je ďalej rozšírená o makrohranice identifikované pomocou štruktúrnych markerov typických pre slovenské súdne rozhodnutia.

2. Klasifikácia segmentov

Po vytvorení počiatočných segmentov systém vykonáva ich klasifikáciu do vopred definovaných právnych kategórií:

- Predmet_sporu
- Vyrokova_cast
- Posudenie_sudu
- Skutkove_okolnosti
- Pravny_ramec
- Dokazy
- Predchadzajuce_konanie_vo_veci
- Zalobca_Prokurator
- Zalovany_Obzalovany
- Dalsie

2.1. TF-IDF reprezentácia segmentov

Každý segment je reprezentovaný pomocou znakových TF-IDF n-gramov. Na rozdiel od klasických prístupov je použitá znaková reprezentácia typu char_wb, ktorá umožňuje robustnejšie zachytenie morfologických variantov slovenského právneho jazyka a znižuje citlivosť na drobné textové odlišnosti.

2.2. Sémantické skórovanie

Pre každý segment je vypočítaná podobnosť voči referenčným opisom jednotlivých tried. Referenčné popisy reprezentujú typický obsah konkrétnej právnej kategórie a slúžia ako prototypové reprezentácie tried.

2.3. Pravidlové heuristiky

Výsledné skóre segmentov je upravované pomocou rozsiahlej množiny heuristických pravidiel založených na:

- výskyte právnych formulácií,
- procesných výrazov,
- odkazov na právne predpisy,
- identifikácii strán konania,
- detekcii výrokových formulácií,
- analýze súdneho odôvodnenia,
- štruktúrnej pozícii segmentu v dokumente.

Pravidlá umožňujú zvýšiť presnosť klasifikácie v prípadoch, kde samotná sémantická podobnosť neposkytuje dostatočne spoľahlivé rozlíšenie právnych kategórií.

3. Sekvenčný postprocessing

Po počiatočnej klasifikácii segmentov systém vykonáva sekvenčný postprocessing, ktorého cieľom je odstrániť nekonzistencie medzi susediacimi segmentmi.
Postprocessing pozostáva najmä z:

- korekcie nesprávne identifikovaných výrokových častí,
- úpravy segmentov na základe ich pozície v dokumente,
- zjednocovania krátkych anomálnych segmentov medzi obsahovo podobnými blokmi,
- opravy prechodových segmentov medzi jednotlivými právnymi časťami.

Tento krok umožňuje zohľadniť sekvenčný charakter právnych dokumentov a stabilizovať výslednú štruktúru segmentácie.

4. Vetný relabeling

Na zvýšenie presnosti hraníc segmentov je implementovaná druhá úroveň klasifikácie na úrovni jednotlivých viet.

Každá veta dedí počiatočný label nadradeného segmentu, pričom následne prechádza samostatným procesom relabelingu založeným na:

- embeddingovej podobnosti,
- lokálnom kontexte susedných viet,
- sekvenčných pravidlách typu ABA a ABC,
- heuristických pravidlách pre právne formulácie.

Systém iteratívne opravuje lokálne nekonzistentné sekvencie viet a následne opätovne spája susediace vety s rovnakými labelmi do finálnych segmentov.

5. Výstup systému

Výstupom systému je štruktúrovaná JSON reprezentácia dokumentu obsahujúca:

- identifikátor segmentu,
- typ segmentu,
- rozsah viet segmentu,
- počet viet segmentu,
- text segmentu,
- klasifikačné skóre,
- alternatívne kandidátne labely.

Systém zároveň generuje analytické CSV výstupy obsahujúce štatistiky relabelingu, prechodové matice medzi triedami a agregované metriky segmentácie.

6. Požiadavky a technológie

Systém je implementovaný v jazyku Python a využíva nasledujúce technológie:

- Sentence Transformers pre embeddingy viet,
- SlovakBERT pre sémantickú reprezentáciu právnych textov,
- scikit-learn pre TF-IDF reprezentácie a kosínovú podobnosť,
- PyTorch pre neuronové modely a GPU akceleráciu,
- NumPy a Pandas pre dátové spracovanie a analýzu,
- heuristické regulárne výrazy pre právne pravidlá.

Odporúča sa RAM 32 GB a viac. GPU akcelerácia prostredníctvom CUDA výrazne zrýchľuje výpočet embeddingov a 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/altaml/legal-text-segmentation