D20 – Metóda na vytváranie a popisovania klastrov (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 KPB4-3 bol vypracovaný výstup Metóda extrakcie kľúčových pojmov. Tento dokument obsahuje poznámky k metóde extrakcie kľúčových pojmov pre právne texty založenej na kombinácií štatistických a sémantických metód použitím referenceií popísaného v článku:

Metóda využíva metódu porovnania vektorových

Grafovo orientovaný kategorizátor súdnych rozhodnutí

Kategorizátor rozhodnutí je projekt založený na jazyku Python, ktorý je navrhnutý na spracovanie súdnych rozhodnutí, tvorbu multipartitných grafových reprezentácií a identifikáciu latentných právnych oblastí prostredníctvom detekcie komunít. Nasledujúce časti opisujú hlavné funkcionality, metodológiu konštrukcie grafov a experimentálne výsledky.

2. Vlastnosti

Kľúčová funkcionalita projektu je zhrnutá do niekoľkých základných procesov:

- **Príprava dát:** Extrakcia právnych odkazov pomocou regulárnych výrazov a metrík podobnosti, ako je Levenshteinova vzdialenosť, príprava množiny potencionálnych kľúčových pojmov, zákódovanie

Extrakcia kľúčových pojmov z jednotlivých častí rozhodnutia: Vypočíta sa skóre pre jednotlivé potencionálne pojmy pre daný dokument a jeho referencie. Skóre je určené na základe kombinácie TF-IDF a kosínovskej podobnosti potencionálnych pojmov a samotného dokumentu.

- **Výpočet váh:** Určenie váhy sémantickej pododbnosti medzi referencovaným textom a samotným právnym textov použitím kombinácie TF-IDF 0

- **Určenie finálnej sady pojmov:** Vypočíta sa finálne skóre daného pojmu pre daný dokument zlúčením skóre zo samotného textu a skóre z jeho referencií ako aj následné prahovanie ktoré vráti konečnú sadu pojmov

3. Požiadavky

Pred nasadením systém vyžaduje štruktúrovaný súbor súdnych rozhodnutí a špecifické výpočtové moduly na analýzu grafov. Projekt je založený na jazyku Python a knižnici NetworkX na grafové operácie.

Požadované komponenty:

- Python 3.x a TFIDFVectorizer (pre štatistickú zložku skóre), resp. bge-m3 pre vytváranie vektorovej reprezentácie
- Extraktor právnych odkazov alebo hotová databáza takýchto odkazov
- Databáza právnych pojmov slúžiaca ako kandidátne alebo metóda na výber takýchto pojmov
- Vektorová reprezentácia právnych textov, jeho referencií a potencionálnych kľúčových pojmov

4. Konštrukcia grafu a váhy

Graf *G*, pozostávajúci zo súdnych rozhodnutí, právnych odkazov, kľúčových slov a hrán, zachytáva viacrozmerný charakter právnych textov. Váhy sú vypočítané tak, aby významné, ale zriedkavé citácie mali vyššiu váhu než často sa vyskytujúce.

Výpočet váh pre hrany medzi súdnym rozhodnutím a právnym odkazom:

Váha medzi súdnym rozhodnutím a právnym odkazom sa vypočíta ako podiel lokálneho počtu citácií daného odkazu v konkrétnom rozhodnutí a globálneho počtu citácií toho istého odkazu v celom dátovom súbore.

- **lokálna váha** predstavuje počet výskytov právneho odkazu v konkrétnom rozhodnutí,

- **globálna váha** predstavuje celkový počet výskytov rovnakého právneho odkazu vo všetkých rozhodnutiach,

všetky váhy sú normalizované na úrovni rozhodnutia tak, aby ich súčet bol rovný jednej.

5. Detekcia komunít

Systém identifikuje klastre, v ktorých sú väzby vnútri komunít hustejšie než väzby medzi komunitami.

Fáza 1: Uzly sa presúvajú medzi komunitami s cieľom maximalizovať prírastok modularity.

Fáza 2: Komunity sa zhlukujú do superuzlov a proces sa opakuje rekurzívne.

Modularita: Meria kvalitu rozdelenia a jej hodnota sa zvyčajne pohybuje v intervale od mínus jedna po jedna.

6. Experimentálne výsledky

Analýza tisíc súdnych rozhodnutí z oblasti trestného práva viedla k identifikácii štrnástich samostatných komunít s hodnotou modularity 0,53.

Vybrané príklady klastrov:

Komunita C1 (55 prípadov)

Kľúčové právne odkazy: § 415 (podmienečné prepustenie), § 37 (odplata)

Kľúčové slová: študent, maloletý, škola, rodina

Komunita C2 (34 prípadov)

Kľúčové právne odkazy: § 289 (návykové látky), § 61 (zákaz činnosti)

Kľúčové slová: zdravie, test, vozidlá, spôsobilosť

Komunita C3 (21 prípadov)

Kľúčové právne odkazy: § 352 (falšovanie), § 394 (obnova konania)

Kľúčové slová: falšovanie, odôvodnenie, sťažnosti

Komunita C4 (71 prípadov)

Kľúčové právne odkazy: § 117 (trest odňatia slobody), § 469 (zahladenie odsúdenia)

Kľúčové slová: sankcie, bezpečnosť, zásah

7. Zdrojové súbory

Zdrojový kód projektu a dátový súbor sú udržiavané v repozitári a sú dostupné na adrese: https://github.com/simonHorvat/GCDC