**Rozšírené zadanie diplomovej práce**

**Názov:** *Sémantické a štrukturálne rozdelenie súdnych rozhodnutí*

**Autor:** Bc. Attila Papán

**Vedúci:** RNDr. Peter Gurský, PhD.

**Konzultant:** RNDr. Zoltán Szoplák

**Ciele:**

1. Oboznámiť sa so štruktúrou Slovenských súdnych rozhodnutí a metódami sémantickej segmentácie
2. Navrhnúť a implementovať metódu na automatizáciu členenia textu rozhodnutia na jednotlivé časti podľa obsahu.
3. Vyhodnotiť úspešnosť implementovaného prístupu s anotovanou sadou od právnych expertov.

**Popis:**

Náš ústav už niekoľko rokov spolupracuje s rôznymi právnikmi a s Právnickou fakultou UPJŠ. Ich spoločným cieľom je zjednodušiť, zlepšiť či zrýchliť prácu právnikov. Na začiatku tohto projektu sa úspechy hľadali ťažko, keďže právnici nevedeli vysloviť, čo presne by potrebovali od nás. Preto v počiatočných fázach informatici vymysleli rôzne možné riešenia a komunikovali s právnikmi, či im to čo spravili, pomáha. Po čase prichádzali úspechy, keďže právnici, už podľa uvedených príkladov použitia dát, vedeli usmerniť informatikov, akou ďaľšou cestou by mohli ísť. Spolupráca je stále lepšia a lepšia, lebo právnici vidia, ako by mohli využiť produkty poskytnuté našou fakultou a tieto riešenia ponúkajú čoraz viac úspechov.

Máme k dispozícií databázu vyše 4 miliónov súdnych rozhodnutí, s ktorými pracujeme s rôznymi spôsobmi. Právnici počas ich práce hľadajú rozsudky, ktoré súvisia s ich prípadom alebo boli podobné, a aj podľa týchto historických rozhodnutí súdov sa pripravia na ich prácu. My im môžme pomáhať hlavne vo vyhľadávaní v tejto obrovskej databáze. Vieme vyhľadávať podľa kľúčových slov, ktoré právnik zadá alebo podľa typu prípadu (trestné právo, občianske právo, rodinné právo atď.)

V našej práci sa pokúsime zrýchliť a zjednodušiť toto vyhľadávanie pomocou rozdelenia textu súdnych rozhodnutí sémanticky aj štrukturálne. Tieto texty zvyčajne majú viacero častí: hlavička rozhodnutia (základné informácie o prípade), úvod (stručný popis predmetu sporu), opis (podrobný opis prípadu, vyšetrovania, faktov), právna analýza, záver (rozhodnutie súdu), odôvodnenie rozhodnutia. Súdne rozhodnutia majú takisto viacero typov, ako rozsudok, uznesenie, nariadenie alebo výnos. Našim hlavným cieľom je rozdeliť tieto rozhodnutia na jednotlivé časti, keďže oni v základe nie sú rozdelené, sú viac-menej jedným súvislým textom. Tým pádom by sa následne ľahšie vyhľadávalo v databáze, jednotlivé slová by sa nehľadali v celom texte, ale len napr. v odôvodnení rozhodnutia. Napr. opis prípadu je z právneho hľadiska málovravný, keďže opisuje okolnosti prípadu, ale môže byť využívaní pri filtrovaní a vyhľadávaní. Na túto prácu potom by mohli byť založené aj iné práce alebo iní by mohli využívať túto, už rozdelenú databázu na ďalšiu prácu na tomto projekte so súdnymi rozhodnutiami.

Keď všetko pôjde podľa plánu, tak ako možný parciálny cieľ do budúcnosti, by sme mohli vytvoriť aj rôzne šablóny pre rôzne typy súdnych rozhodnutí. Toto závisí od úspešnosti základného cieľa, ale je dobrou možnosťou pre vylepšenie.

Náš cieľ môžeme dosiahnúť viacerými spôsobmi alebo ich rôznymi kombináciami. Takými spôsobmi sú napr. hľadanie štandardnosti/unikátnosti citovaných zákonov, sémantickej excentricity alebo sémantického porovnania so všeobecným právnym textom. Taktiež môžeme využiť pološtrukturovanosť textov súdnych rozhodnutí, aby sa na členenie mohli využiť štandardizované frázy.

Keďže budeme pracovať so slovenčinou a presnejšie s právnou slovenčinou, ktorá používa odbornejšie slová, tak nebudeme mať ľahkú úlohu, lebo neexistuje veľa jazykových modelov na tento jazyk. V práci použijeme transformerové jazykové modely, ktoré nám umožnia spracovať texty. Priamo na slovenčinu je optimalizovaný model SlovakBERT (dostupný napr. na stránke o machine learningu, Hugging Face), ktorý nám poskytne presnejšiu analýzu textu. So SlovakSentenceBERT môžeme rozčleniť vety a odseky. Najnovšie jazykové modely ako Mistral-7B a LLaMa sú veľkými jazykovými modelmi a tiež vyskúšame ich využitie počas práce.

**Literatúra:**

1. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge: Cambridge University Press.
2. Jurafsky, Daniel, and James H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models. 3rd ed., 2024.
3. Lattisi, T., Farina, D., & Ronchetti, M. (2022). Semantic Segmentation of Text Using Deep Learning. COMPUTING AND INFORMATICS, 41(1), 78–97. https://doi.org/10.31577/cai\_2022\_1\_78
4. Šavelka, Jaromír and Kevin D. Ashley. “Segmenting U.S. Court Decisions into Functional and Issue Specific Parts.” International Conference on Legal Knowledge and Information Systems (2018).