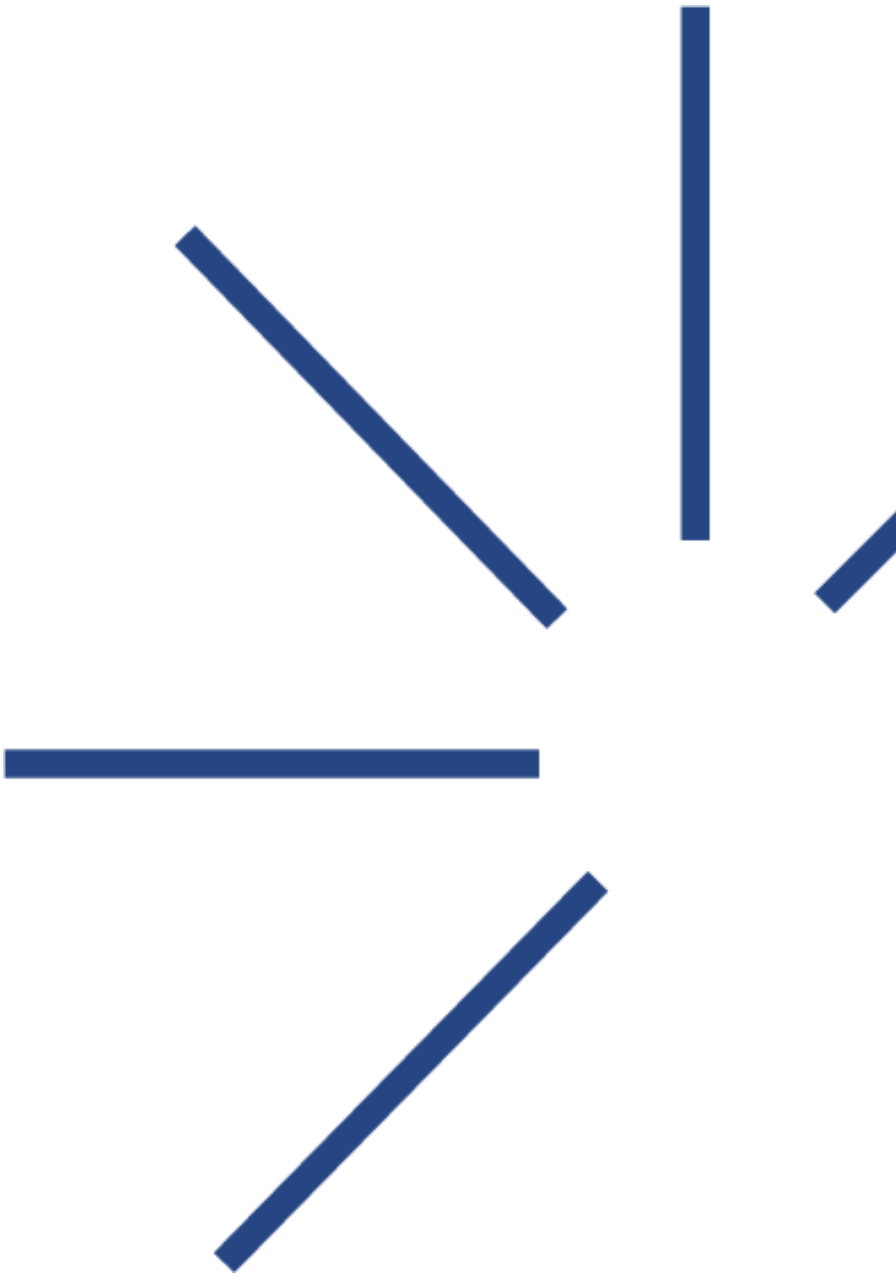
D06 – Metóda na extrakciu rozsudkov(SK)



Projekt financovaný Európskou úniou Next Generation EU 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 na extrakciu rozsudkov. Navrhnutý systém predstavuje hybridný extraktor údajov z trestnoprávnych rozhodnutí, ktorý kombinuje pravidlové spracovanie textu, heuristické metódy spracovania prirodzeného jazyka a veľký jazykový model lokálne nasadený prostredníctvom platformy Ollama. Systém je navrhnutý na automatizovanú analýzu slovenských trestných rozhodnutí a identifikáciu štruktúrovaných právnych údajov vrátane skutkovej vety, uložených trestov, poľahčujúcich a priťažujúcich okolností, dôkazných prostriedkov a identifikácie relevantných trestných činov.

Súčasťou systému je viacúrovňová architektúra, ktorá kombinuje deterministické pravidlá, štatistické predspracovanie textu a jazykový model veľkého rozsahu (LLM). Extrakcia trestných činov je obmedzená na vopred definovanú databázu slovenských trestných činov, čím sa zabezpečuje terminologická konzistentnosť a minimalizuje sa riziko generovania neexistujúcich alebo právne nesprávnych označení. Stručnejší popis systému je uvedený nižšie.

Hybridný extraktor procesných údajov a trestných činov zo súdnych rozhodnutí

Extraktor je softvérový modul implementovaný v jazyku Python, určený na automatickú štrukturalizáciu údajov zo slovenských trestných rozhodnutí. Systém využíva kombináciu pravidlových metód, heuristickej analýzy textu a veľkého jazykového modelu LLaMA 3.3 na presnú extrakciu právne relevantných informácií.

Funkcionalita systému

Hlavná funkcionalita systému pozostáva z nasledujúcich krokov:

1. Príprava vstupných dát

Systém pracuje s korpusom trestnoprávnych rozhodnutí vo formáte JSON. Rozhodnutia obsahujú plné textové znenie dokumentu vrátane skutkového stavu, právneho posúdenia, dôkaznej situácie a výroku súdu.

Na efektívne spracovanie rozsiahlych dátových súborov je použitý streamingový parser ijson, ktorý umožňuje sekvenčné načítavanie rozhodnutí bez potreby načítania celého datasetu do operačnej pamäte.

Texty rozhodnutí sú normalizované jednoduchým čistením, ktoré zahŕňa:

- odstránenie nadbytočných medzier,
- normalizáciu interpunkcie,
- odstránenie špeciálnych znakov,
- transformáciu textu do jednotného formátu.

2. Predbežná analýza rozhodnutia

Pred samotnou extrakciou systém vykonáva kontrolu počtu obžalovaných alebo páchateľov v rozhodnutí. Táto úloha je realizovaná pomocou jazykového modelu, ktorý analyzuje úvodnú časť rozhodnutia a určuje, či ide o konanie proti jednému alebo viacerým páchateľom.

Rozhodnutia obsahujúce viac páchateľov sú automaticky vyradené zo spracovania, čím sa znižuje riziko nesprávnej atribúcie údajov jednotlivým osobám.

3. Extrakcia procesných údajov

Hlavná extrakcia údajov je realizovaná pomocou veľkého jazykového modelu riadeného striktne definovaným promptom formulovaným v slovenskom jazyku.

Systém extrahuje najmä:

- skutkovú vetu rozhodnutia,
- informáciu o oslobodení spod obžaloby,
- uložené tresty,
- poľahčujúce okolnosti (§ 36 TZ)
- priťažujúce okolnosti (§ 37 TZ),
- dôkazné prostriedky uvedené v rozhodnutí.

Skutková veta je extrahovaná v čo najvernejšej podobe priamo z textu rozhodnutia bez sumarizácie alebo preformulovania.

4. Extrakcia trestných činov

Identifikácia trestných činov prebieha v samostatnej viacstupňovej fáze spracovania:

4.1. Databáza trestných činov

Systém využíva externý JSON súbor obsahujúci databázu slovenských trestných činov.

Každý záznam obsahuje:

- názov trestného činu,
- príslušný paragraf,
- textový popis skutkovej podstaty.

Táto databáza predstavuje uzavretú množinu kandidátnych trestných činov, z ktorej môže systém vyberať.

4.2. Heuristický výber kandidátov

Pred odoslaním textu jazykovému modelu systém vytvára množinu kandidátnych trestných činov na základe lexikálneho prekryvu medzi skutkovou vetou a popismi trestných činov.

Pri výpočte kandidátov sa využíva:

- tokenizácia textu,
- odstránenie slovenských stopslov,
- normalizácia textu,
- heuristické zvýhodnenie vybraných právnych indikátorov.

Systém obsahuje špecializované heuristiky pre často sa vyskytujúce skutkové situácie, napríklad:

- neplatenie výživného,
- nebezpečné vyhrážanie,
- poškodzovanie cudzej veci,
- skutky pod vplyvom alkoholu,
- podvodné konania

.

Výsledkom tejto fázy je obmedzený počet najrelevantnejších kandidátov, ktoré postupujú do finálneho rozhodovania jazykového modelu.

5. Reranking pomocou jazykového modelu

Finálna identifikácia trestných činov je realizovaná pomocou lokálne nasadeného jazykového modelu LLaMA 3.3 prostredníctvom platformy Ollama.

Model je riadený striktne definovaným promptom, ktorý zabezpečuje:

- výber výhradne zo zadaných kandidátov,
- zákaz generovania nových názvov trestných činov,
- možnosť identifikácie viacerých trestných činov v súbehu,
- deterministický výstup vo forme JSON štruktúry.

V prípade zlyhania jazykového modelu alebo prázdneho výsledku systém automaticky aktivuje pravidlový fallback mechanizmus založený na heuristických pravidlách a kľúčových slovách.

6. Extrakcia dôkazov a okolností

Systém identifikuje a kategorizuje dôkazy uvedené v rozhodnutí do viacerých skupín:

- výsluchy,
- listinné dôkazy,
- znalecké posudky,
- vecné dôkazy a záznamy,
- ostatné dôkazy.

Poľahčujúce a priťažujúce okolnosti sú mapované na ich plné zákonné znenie podľa ustanovení Trestného zákona. Pri identifikácii odkazu na konkrétne písmeno zákonného ustanovenia systém automaticky doplní celé textové znenie okolnosti.

7. Výstup systému

Výstupom systému je JSON štruktúra obsahujúca:

- identifikátor rozhodnutia,
- informáciu o prípadnom preskočení rozhodnutia,
- skutkovú vetu,
- identifikované trestné činy,
- uložené tresty,

- poľahčujúce a priťažujúce okolnosti,
- kategorizované dôkazy.

Výsledky sú ukladané priebežne počas spracovania, čím sa minimalizuje riziko straty dát pri spracovaní rozsiahlych korpusov.

8. Požiadavky a technológie

Systém je implementovaný v jazyku Python a využíva nasledujúce technológie:
- Ollama pre lokálne spúšťanie jazykového modelu,
- model LLaMA 3.3 pre právnu extrakciu údajov,
- ijson pre streamingové spracovanie rozsiahlych JSON datasetov,
- requests pre komunikáciu s LLM API,
- regulárne výrazy a heuristické pravidlá pre predspracovanie textu.

Odporúča sa minimálne 32 GB RAM. Podpora GPU (CUDA) výrazne zrýchľuje inferenciu jazykového modelu pri spracovaní rozsiahlych právnych korpusov.

9. Zdrojové súbory

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