D08 – Databáza slovenských a anglických právnych pojmov (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 KPB2- 4 bol vypracovaný výstup Databáza Slovenských a Anglických právnych pojmov (dataset). Tento dokument obsahuje popis štruktúry a metodológie vytvárania databázy právnych pojmov ako aj ich definícii v slovenskom a anglickom jazyku vo formátoch .csv, .xlsx a .json.

1. Slovník slovenských a anglických právnych pojmov

Databáza je uložená vo formátoch .csv, .xlsx a .json. Obsahuje právne pojmy zozbierané z online zdrojov, ako sú **Slov-Lex** a **Najpravo.sk**, ako aj z vybraných nadpisov paragrafov Zbierky zákonov Slovenskej republiky. Nakoľko nie každý nadpis paragrafu je právnym pojmom, tieto nadpisy boli následne odfiltrované kombináciou štatistických a sémantických metrík, ako aj pomocou promptovania veľkých jazykových modelov.

Databáza obsahuje aj definície právnych pojmov v slovenskom jazyku, pričom pri niektorých pojmoch sú dostupné definície z viacerých zdrojov, či už z online databáz alebo priamo z textov paragrafov, ku ktorým dané pojmy vystupujú ako nadpisy. Súčasťou databázy sú aj anglické právne pojmy a ich definície, ktoré boli získané prekladom pomocou veľkého jazykového modelu **LLaMA 3.3**.

2. Vlastnosti

Databázy pravne_pojmy_sk_en.csv a pravne_pojmy_sk_en.xlsx obsahujú výlučne zoznamy pojmov, kde stĺpec **SK** predstavuje slovenské právne pojmy a stĺpec **EN** ich anglické ekvivalenty. Tieto súbory neobsahujú definície ani ďalšie metadáta.

Databáza legal_terms_sk_en.json obsahuje rozšírené informácie o pojmoch a zahŕňa nasledujúce polia:

- **_id**: samotný právny pojem v slovenskom jazyku (napr. „premlčacia doba“)

- **definitions**: zoznam definícií prislúchajúcich k danému pojmu, pričom každá definícia obsahuje:
    - **source**: zdroj definície (napr. „najpravo“)
    - **definition**: samotná definícia v slovenskom jazyku (napr. „Premlčacia doba je ...“)

V prípade, že hodnota poľa **source** je „zz“ (Zbierka zákonov) a pojem bol získaný z nadpisu paragrafu, pole **definition** neobsahuje iba jednoduchý text, ale má štruktúrovanejší charakter:

    - **law**: identifikácia zákona (napr. „61/1952“)

    - **title_tag_id**: identifikátor nadpisu v štruktúre zákona (napr. „predpis.hlava-10.nadpis“)

- **definition**: definícia pojmu extrahovaná z textu paragrafu v slovenskom jazyku

- **english**: informácie o anglickom ekvivalente pojmu získanom pomocou veľkého jazykového modelu:
  - **term**: anglický ekvivalent právneho pojmu (napr. „statute of limitations“)
  - **definition**: anglická definícia pojmu vytvorená prekladom a sumarizáciou slovenských definícií

3. Požiadavky

Na základné používanie databázy postačuje nástroj na čítanie a úpravu súborov vo formátoch .csv a .xlsx, ako sú textové editory alebo tabuľkové procesory (napr. Microsoft Excel).

Na prácu s rozšírenou databázou obsahujúcou definície vo formáte .json je vhodné použiť softvér určený na prácu s databázami alebo štruktúrovanými dátami, napríklad **MongoDB**, **CouchDB**, **PostgreSQL**, prípadne aj pokročilé textové editory.

4. Zdroje

Zdroje právnych pojmov tvoria verejne dostupné online repozitáre, najmä **Slov-Lex** a **Najpravo.sk**, ktoré poskytujú aj definície pojmov. Táto množina pojmov bola následne rozšírená o vybrané pojmy získané z nadpisov paragrafov zákonov v Zbierke zákonov.
Tento krok bol motivovaný jednak nedostatočným pokrytím právnej terminológie v dostupných online zdrojoch, ako aj pozorovaním, že mnohé paragrafy majú definičný charakter. Keďže nie všetky nadpisy paragrafov sú vhodné ako právne pojmy, boli podrobené filtrovaniu na základe analýzy ich znenia, obsahu samotného paragrafu a všeobecného právneho charakteru, a to aj s využitím promptovania veľkých jazykových modelov.
V prípadoch, kde bol text paragrafu identifikovaný ako definičný, bola definícia priamo extrahovaná a priradená k príslušnému pojmu.
Z dôvodu nedostatku kvalitných zdrojov právnych pojmov v anglickom jazyku boli anglické ekvivalenty získané prekladom pomocou veľkého jazykového modelu. Príslušné anglické definície vznikli prekladom a sumarizáciou všetkých dostupných slovenských definícií.
Hoci tieto definície nemusia byť vždy terminologicky úplne presné, keďže ide o generované dáta, ich primárne využitie spočíva v tvorbe vektorových reprezentácií pojmov, kde je dôležitý najmä kontext a tematická konzistentnosť, nie absolútna jazyková presnosť.