Semantični analizator – Pametni iskalnik besedil

(*razvoj prototipa je končan; celovita rešitev je v razvoju, projekt se zaključi konec julija 2022*)

Semantični analizator je orodje, s katerim lahko analiziramo slovenska besedila (na primer zakonske dokumente), ugotavljamo, kateri pojmi so ključni za razumevanje vsebine in kateri od teh manjkajo v nekem besednjaku oziroma slovarju (npr. besednjaku temeljnih pojmov javne uprave). Hkrati pa lahko v množici obdelanih dokumentov izberemo tiste, v katerih so uporabljeni (izbrani) pojmi iz besednjaka. Pojmi v besednjaku tako postanejo vezni člen med dokumenti, s čimer odkrivamo soodvisnosti in/ali vsebinske povezave med dokumenti. Tako lahko na primer poiščemo tiste zakonske dokumente, ki se sklicujejo na nek vir podatkov (register, evidenca, seznam).

Orodje je odprtokodno in poleg funkcionalnosti opisanih v tem projektu vključuje tudi standardne funkcije s področja analitike besedil, vizualizacije in strojnega učenja. Vhodni vir meta podatkov o pojmih naj bodo zapisi v obliki datotek rdf (sintaksa Turtle).

**Namen projekta**: Prototip programske rešitve, ki bo iz množice dokumentov vsak dokument opisala z zbirko izrazov in na podlagi tako pridobljenih dokumentnih profilov poročala o podobnosti med dokumenti in podobnosti povzela v grafičnih prikazih dokumentne množice. Orodje bo analiziralo tudi izbrani besednjak ali slovar in tako dobljeno zbirko izrazov primerjalo ter povezalo z dokumentnimi profili (npr. pojmi, ki jih ni v besednjaku; pojmi iz besednjaka s pripadajočimi dokumenti).

Besednjak temeljnih pojmov javne uprave (Centralni besednjak javne uprave) pokriva širok nabor vsebinskih področij sistemov državne uprave. Za opisovanje pojmov in njihovih definicij so odgovorni skrbniki posameznih vsebinskih področij. Pričakujemo, da bo količina opisanih pojmov kmalu prevelika, kar bi zelo otežilo skrbnikom vpogled v celoto in s tem ugotovitev, ali je bil nek pojem že opisan, ali je njegov opis primeren, ali pa gre za sinonim. Ker take povezave med pojmi bistveno olajšajo iskanja, je pomembno, da jih ne spregledamo in jih ustrezno zabeležimo v besednjaku.

**Cilj**: Glavne funkcionalne zahteve so: (1) prepoznava ključnih besed v dokumentih, (2) analiza dokumentov po bližini ključnih besed in glede na kontekst posameznih dokumentov (tudi na podlagi naročnikovih že izdelanih ontologij), (3) samodejna izdelava in/ali dopolnitev ontologij na podlagi analize dokumentov ter (4) vizualizacija rezultatov posameznih analiz.

1, 2, 4 – ustvarjanje konteksta oziroma skupin kontekstov v množici besedil

1 – ustvarjanje konteksta posameznega besedila

2, 4 – povezave med dokumenti glede na vsebino, torej prek analize pojmov, ki so značilni za posamezna besedila (povezave med dokumenti glede na izbrane pojme iz ontologij, slovarjev, stvarnih kazal, ključnih besed)

3 – pomoč pri izdelavi ontologij, slovarjev, stvarnih kazal, ključnih besed

**Obseg projekta razvoja prototipa**: omejen nabor zakonov s PISRS na temo registrov (iskalni izraz: register OR registr\*)

**Testni rezultati razvoja prototipa**: Preberemo tekstovne datoteke, v katerih so zakonodajna besedila s področja registrov, in jih ustrezno obdelamo za nadaljnjo analizo (npr. poenotenje malih in velikih črk, deljenje besedil na besede, poenotenje sklonov in spregatev, filtriranje besed, ki odstrani za pomen besedila nepomembne besede in števila ter izbriše tiste besede, ki se pojavijo preveč ali premalo pogosto, saj te tipično niso informativne za pomen besedila).

Pregledovanje dokumentnih kart pri analizi besedil je zelo pomembno, saj dobimo globalno sliko o podobnosti besedil. Dokumentna karta je prikaz besedil, ki so si med seboj podobna, kot točk (kot skupine točk-skupine oziroma gruče podobnih dokumentov oziroma pomensko povezave med dokumenti) v dvodimenzionalnem prostoru, ki ležijo blizu (bolj ali manj) ena drugi. (gornja veja delokroga prototipa prikaže značilnosti celotnega nabora besedil)

Na karti je vsako besedilo predstavljeno z eno točko. Točke so obarvane glede na gruče, ki jih je določil algoritem. Na dokumentni karti lahko izberemo točke (besedila), ki nas zanimajo, in opazujemo pogoste besede v izbranih dokumentih. (srednja veja delokroga prototipa prikaže značilnosti izbora besedil glede na celoten nabor)

Funkcija »Obogatenost besed« prikaže seznam besed, ki so bolj značilne za izbrana besedila v primerjavi s celotnim naborom. Če želimo videti, katere besede so bolj značilne za izbrana besedila, jih razvrstimo glede na p-vrednost (p-value). Nižja p-vrednost pomeni bolj obogatene (značilne) besede v izbranih besedilih.

Zadnja funkcionalnost prototipa pa prebere datoteko v obliki RDF (pojme z njihovimi lastnostmi in povezavami z drugimi pojmi), ki jo je v našem primeru predstavljala ontologija oz. semantični jedrni podatkovni model osrednjega besednjaka javne uprave. Nabrane pojme iz besednjaka nato aplikacija poišče v izbranih besedilih in pove, kateri pojmi ter v kakšnem obsegu so zastopani v posameznih besedilih. (tretja veja delokroga je namenjena namenski nalogi, ki uporabi rezultate prejšnjih analiz; prototip poišče pojme iz besednjaka v izbranih besedilih)

**Nauki iz projekta razvoja prototipa**:

Iz zakonskih besedil, izbranih prek iskalnega izraza »register OR registr\*«, smo dobili okoli 240 dokumentov. To je predstavljalo delovno področje, korpus, na katerem se je opravila semantična analiza.

Na tej podlagi so nastale 4 glavne skupine dokumentov. Zaradi omejenega nabora z določenim iskalnim nizom so skupine sicer nekoliko manj izrazite, vseeno pa imajo nekaj značilnosti. Tako skupina C1 bolj poudarja varstvo okolja, varstvo pri delu in socialno varstvo ter povezano s tem tudi o (javno) finančne posledice. Skupina C2 npr. bolj poudarja javne in zasebne pravne osebe (organizacijski vidik). Skupina C3 poudarja dejavnosti (gospodarske in negospodarske) in pogoje, zahteve, dolžnosti v zvezi z njimi. Skupina C4 pa npr. poudarja obravnavo podatkov, evidence in registre.

Glede na testne rezultate vidimo možnost uporabe npr. pri delu z zakonodajo in sorodnih dokumentih (pripravljalni in dopolnilni akti, predlogi z obrazložitvami, analize …), pri katerem bi dobili iz množice besedil nekaj vsebinskih skupin, s čimer bi izbrali skupino besedil, ki nas zanima, ali bi razdelili skupine besedil med več ljudi in si razdelili delo oziroma bi lahko nekatera besedila uvrstili kot prednostna ali po drugi strani kot nepomembna...

Možnost uporabe se vidi tudi glede iskanja vsebinske povezanosti med dokumenti iz sicer različnih vsebinskih področij. V katerih besedilih najdemo npr. iste pojme, na tej podlagi pa potem ugotavljamo, ali gre za podvajanje ali za dopolnjevanje ali morda celo protislovja…

Nadalje bi bilo orodje lahko uporabno za iskanje vsebinskih povezav med različnimi vrstami besedil, kot npr. zakonodajo, sodno prakso, članki, drugimi besedili … Tako bi lahko naredili delovne pakete besedil po področjih.

Orodje omogoča tudi iskanje po posameznih predizbranih pojmih, po katerih nam potem poišče sorodne dokumente, s čimer si zopet zožimo nabor besedil, ki jih moramo (želimo) pregledati, oziroma tudi razširimo nabor, saj orodje lahko najde sorodnosti tudi v besedilih, ki sicer (na prvi pogled, torej glede na ime dokumenta (zakona oz. zakonodajnega ali drugega besedila)) najverjetneje ne bi prišla v vsebinsko obravnavo glede na temeljni namen besedila.

Orodje analizira posamezne dokumente po vsebinski zastopanosti posameznih pojmov, kar lahko prikaže potem skupno za celoten nabor besedil, torej o čem največ govorijo vsa besedila, potem nam prikaže skupine glede na najbolj značilne pojme za skupino (posebnosti skupine glede na vsa besedila). Orodje pa omogoča tudi prikaz posebnosti posameznega besedila glede na posamezno skupino, v kateri se nahaja. Pri tem se vidi tudi možnost opisovanja besedil s ključnimi besedami, uvrščanjem besedil v kazala, slovarje ipd.