SEMANTIČNI ANALIZATOR – RAZVOJ PROGRAMSKEGA OKOLJA ZA ALGORITMIČNO OBDELAVO SLOVENSKIH BESEDIL

Miha Jesenko, Miro Lozej, Karmen Kern Pipan, Primož Godec, Vesna Tanko, Lan Žagar, Ajda Pretnar Žagar, Nikola Đukić, Blaž Zupan
Ministrstvo RS za javno upravo (MJU), Tržaška 21, 1000 Ljubljana
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Večna pot 113, 1000 Ljubljana
Karmen.Kern-Pipan@gov.si, Blaz.Zupan@fri.uni-lj.si, Miha.Jesenko@gov.si, Miro.Lozej@gov.si

Povzetek

Poročamo o prvih rezultatih projekta, v katerem razvijamo splošno uporabno orodje za analizo množice besedilnih dokumentov. Cilj projekta je izbor in implementacija gradnikov semantične analize, s kombinacijo katerih lahko izvajamo poljubne tipe analiz dokumentov in gradimo analitične delotoke, ki bi bili lahko uporabni pri tipičnih nalogah, opravilih in storitvah javne uprave. Implementacija vključujeje gradnike za dostopanje do podatkovnih prostorov, vložitve dokumentov v vektorske prostore, iskanje podobnih dokumentov, vizualizacijo podatkovnih kart, iskanje karakterističnih pojmov, rangiranje dokumentov glede na semantično podobnost z izbranimi pojmi in urejanje pojmov v ontologije. V članku predstavimo primer uporabe semantičnega povezovanja predlogov vladi z zbirko zakonskih besedil.

Abstract

We report on the results of the project to develop a general-purpose tool for analyzing a set of textual documents. The project aims to select and implement semantic analysis building blocks that can be used to perform arbitrary types of document analyses and prototype analytical workflows that could support the tasks and decision-making in public administration. The building blocks we have developed include components to access data repositories, embed documents in vector spaces, search for similar documents, visualize document maps, search for characteristic terms, rank documents according to their semantic similarity to selected terms, and arrange concepts into ontologies. In the paper, we present a use case to semantically link the proposals to the government with a collection of laws.

Ključne besede semantična analiza podatkov, podatkovni prostori, analiza besedil, analitika z vizualizacijami, delotoki

Keywords
semantic analysis, data spaces, text mining, visual analytics, workflows

UVOD

Današnji čas in naša vizija prihodnosti sta vedno bolj odvisna ter navezana na podatke o čim več vidikih človeka, družbe, stvari, in pojavov, skratka, glede vsega, kar človek lahko zazna in obdela s svojimi čutili. Na tej podlagi in ob razvoju tehnologije ter znanosti pridobivamo vedno več podatkov in

razvijamo nove pristope za njihovo obdelavo. Z rastjo množice shranjenih podatkov nujno in neobhodno trčimo ob problem njihovega razumevanja. Poenostavljeno, ali lahko računalnik »razume« vsebino podatkov? Ali malce bolj prizemljeno, ali lahko uredimo podatke skladno z vsebino in ali lahko v podatkih poiščemo tiste dele, ki nas, uporabnike, vsebinsko najbolj zanimajo?

Podatki so nam na voljo v najrazličnejših oblikah in strukturah. V zadnjih letih se je opazno povečala tudi zmožnost za obdelavo in uporabo nestrukturiranih podatkov, do katerih lahko vse lažje dostopamo in za katere so v zadnjem času razviti tudi ustrezni analitični postopki. Primer nestrukturiranih podatkov so prosta besedila v klasičnem, esejskem zapisu. Večjih množic takih esejev posameznik ni več sposoben količinsko, kaj šele kakovostno pregledati, razumeti in med sabo primerjati. Za kakovost in učinkovitost dela v prihodnje je nujno oblikovati analitična orodja, ki nam bodo v pomoč pri razumevanju besedil, razvrščanju po vsebini ter semantičnem preiskovanju, kjer iščemo dokumente, ki so vsebinsko povezani z izbranimi pojmi. Našteti pristopi lahko podprejo posredovanje, predstavitev in razlago podatkov in sklepanj ter odločitve na njihovi podlagi, privarčujejo čas, omogočijo, da se osredotočimo le na pomembne zapise, in s tem izboljšajo kvaliteto odločitev.

Digitalna transformacija omogoča javnemu sektorju, da sodeluje z notranjimi in zunanjimi deležniki pri novih in učinkovitejših načinih za ustvarjanje javne vrednosti, delitvi virov in uporabi podatkov za večjo odzivnost na potrebe državljanov in podjetij. V javni upravi imamo bogat spekter podatkov, s katerimi upravljamo delovne procese in izvajamo storitve za državljane, podjetja in širšo družbo (Kern Pipan idr., 2020). Kot navaja OECD (OECD, 2019) so nekatere države v zadnjem času dosegle pomemben razvojni premik s strateško uporabo podatkov za boljše oblikovanje politik, izvedbo storitev ali poslovanja. OECD je v svojih pregledih vloge podatkov v podatkovni ekonomiji in javnem sektorju oblikoval model podatkovno spodbujenega javnega sektorja (angl. *data-driven public sector*), ki:

- prepoznava podatke kot ključno strateško vrednost (bogastvo),
- izpostavlja odstranjevanje ovir pri upravljanju, deljenju in ponovni uporabi podatkov,
- uporablja podatke za preobrazbo oblikovanja, izvedbe in nadzora javnih politik in storitev, in
- ceni prizadevanja za objavo podatkov na odprt način kot tudi uporabo podatkov znotraj organizacij ter znotraj javnega sektorja.

OECD še poudarja, da države lahko uporabijo podatke za oblikovanje javne vrednosti s tremi tipi aktivnosti:

- predvidevanje in planiranje: uporaba podatkov pri oblikovanju politik, načrtovanje posredovanj, predvidevanje možnih sprememb in napovedovanje potreb,
- izvedba storitev: uporaba podatkov za informiranje in izboljšanje vpeljave politik, odzivnosti vlad in aktivnosti pri izvedbi storitev,
- ocenjevanje in spremljanje: uporaba podatkov pri merjenju vpliva, revizijske odločitve ter spremljanje uspešnosti poslovanja (OECD, 2019).

Vse zgoraj zapisano seveda predvideva, da so podatki v javni upravi zbrani, urejeni, dostopni, da so tehnologije za njihovo uporabo nared in da so vključene v praktične informacijske sisteme, ki s pridom pomagajo tako zaposlenim v javni upravi kot državljanom. A to je le cilj. Preobrazba v podatkovnospodbujeni javni sektor je pot, ki jo je potrebno prehoditi in katere dolžina bo odvisna predvsem razumevanja uporabnosti novih tehnologij, praktičnosti rešitev in sodelovanja domenskih ekspertov iz javne uprave z oblikovalci novih pristopov in rešitev.

Prav v namene spodbujanja sodelovanja in iskanja rešitev na področju uvajanja pristopov umetne inteligence v javni upravi smo ob koncu leta 2020 avtorji tega prispevka pričeli delo na projektu, katerega cilj je razviti in raziskati uporabnost pristopov semantične analize podatkovnih prostorov dokumentov, ki se tipično skladiščijo in uporabljajo v javni upravi. Naš cilj je razvoj orodij, ki uporabnikom omogočijo enostavno snovanje analitičnih delotokov, in prototipni razvoj aplikacij, ki jih lahko ovrednotimo s stališča uporabnosti in možnosti integracije v obstoječe informacijske sisteme. Spodaj poročamo o začetnih rezultatih projekta, identifikaciji osnovnih gradnikov analitičnih delotokov za semantično analizo dokumentov in o primeru uporabe na področju semantičnega povezovanja državljanskih predlogov vladi in zakonov, ki so povezani s področjem izbranega predloga.

PODATKI, UMETNA INTELIGENCA IN ODLOČANJE

Nove tehnologije, predvsem pa umetna inteligenca s hitro razvijajočo se podatkovno znanostjo, odpirajo nova obzorja in do sedaj neslutene možnosti uporabe podatkov praktično vsakomur. Pri tem je pomembno razmisliti tudi o priložnostih in izzivih, ki jih nove podatkovne tehnologije prinašajo na vseh ravneh. V strokovni literaturi (OECD, 2019; Provost in Fawcett, 2013) najdemo izraze kot so denimo »podatkovno usmerjeno delovanje« in »odločanje na podlagi podatkov«. Slednje med drugim zahteva zavedanje o pomenu podatkov v kar najširšem družbenem obsegu, zlasti seveda pri organih strateškega odločanja, ter nova znanja in veščine pri uporabi algoritmov in orodij za obdelavo podatkov (Kern Pipan idr., 2020). Vse to zahteva tudi nove načine organiziranja in upravljanja podatkov tako na mikro kot makro ravni. Tako je EU lansko leto napovedala vzpostavitev podatkovnih prostorov za področje javne uprave (EC, 2020). Namen je organizirati tudi podatke javnih uprav tako, da jih je možno obdelovati s sodobnimi tehnikami. Poleg tega, da bo javna uprava sistematično strukturirala in usklajevala svoje podatke, bo omogočena primerjava stanja z drugimi državami članicami EU na tem področju. Podatkovni prostori bodo omogočali učinkovitejšo uporabo podatkov z novimi podatkovnimi tehnikami v podporo odločanju (različna enostavna razvrščanja, oblikovanje in priprava kriterijev ter vzorčenja). Prav odločanje na podlagi dejstev in podatkov je ideal, h kateremu stremi vsaka napredna organizacija, tudi javna uprava. V zadnjih letih smo priča premiku usmerjenosti razvoja informacijskih sistemov od aplikacij k uporabi podatkov za pridobivanje informacij. Podatkovna orodja so tista, ki omogočajo hitro pridobivanje informacij iz večjih količin nepovezanih podatkov, ki so sicer običajnemu uporabniku težje dostopni.

V javni upravi najdemo veliko besedilnih dokumentov, med katerimi moramo poiskati tiste, ki govorijo o neki vsebini in jih je potrebno pregledati, da bi dobro utemeljili obrazložitve svojih predlogov ali celo odločitev. Tipičen primer so denimo zakonska besedila oz. iskanje zakonskih dokumentov, ki po vsebini obravnavajo želeno vsebino. Iskanje po vsebini bi nam predstavilo kratek seznam dokumentov, ki bi jih

bilo vredno podrobneje preučiti. Če se pri tem lahko zanesemo, da je predlagani nabor dokumentov popoln (t. j. da pomembni dokumenti niso izpuščeni) in relevanten (t. j. da v naboru ni dokumentov, ki se ne nanašajo na iskano vsebino) lahko uporabnikom bistveno poenostavimo in skrajšamo delo. Uporabniki se ne bi ukvarjali z nerelevantnimi dokumenti in bi bili prepričani, da so upoštevali vsa pomembna gradiva. Običajno iskanje po ključnih besedah takim potrebam ne more zadostiti, potrebno je poznavanje vseh besedil, da smo lahko prepričani, da ničesar nismo spregledali in da zadošča pregled predlaganih besedil, ne da bi skrbeli, da česa nismo spregledali.

Uporabo semantičnih tehnologij na področju javne uprave pa moramo še raziskati. Ugotoviti moramo, ali so podatki, ki jih imamo na voljo, primerni, odkriti zahteve uporabnikov in preskusiti, kako in ali jim je moč zadostiti s trenutno znanimi tehnologijami. Ker je področje novo, je do uporabniških zahtev najenostavneje priti z gradnjo pilotnih aplikacij. Za podatkovno analitiko je te najenostavneje graditi v sistemih, ki podpirajo vizualno gradnjo analitičnih delotokov iz osnovnih gradnikov oziroma analitičnih komponent. Aplikacije te vrste lahko gradimo v okoljih, kot sta komercialna KNIME (https://www.knime.com) in RapidMiner (http://rapidminer.com) ter prostodostopni in odprti Orange (http://orangedatamining.com). To je mogoče le, če so v teh okoljih na voljo ustrezni gradniki za gradnjo prototipov. Tekom projekta smo tako ugotovili, da med temi potrebujemo na primer gradnike za dostop do podatkovnih prostorov dokumentov, gradnike za pripravo iskalnih pojmov in gradnjo ontologij in gradnike za iskanje karakterističnih izrazov v dokumentih. Pomembno je, kako ti gradniki predstavijo rezultate analize in ali je z njimi moč zgraditi aplikacije, ki lahko služijo različnim namenom in lahko obdelujejo vrsto različnih tipov dokumentov, naslovijo večino potreb uporabnikov in je z njimi moč na razložljiv način prikazati uporabnost novih tehnologij.

PRISTOPI Z VIZUALNIM PROGRAMIRANJEM IN VIZUALNO ANALITIKO TER NAPREDNE TEHNIKE ANALIZE BESEDIL

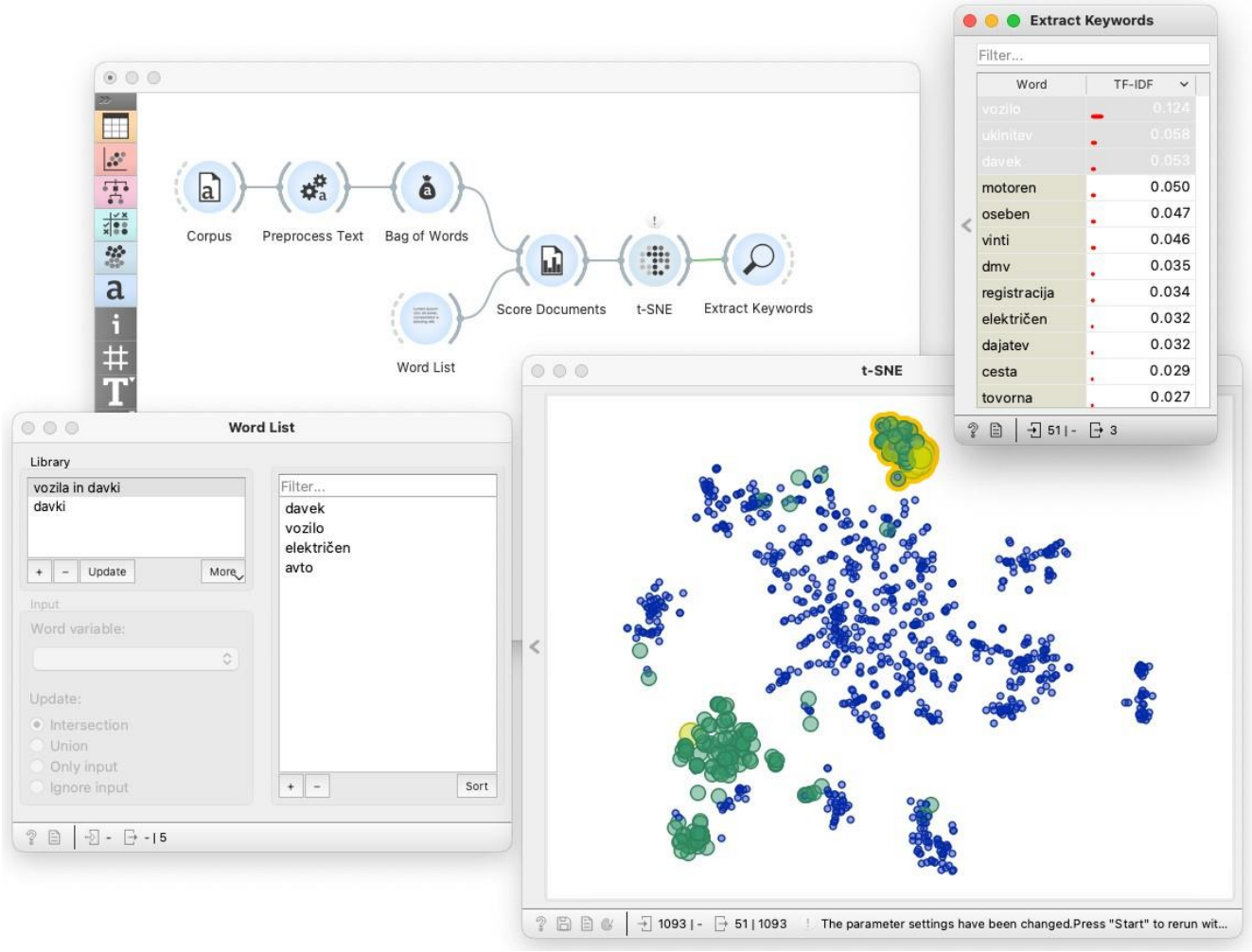
V projektu orodje za podatkovno analitiko Orange razširjamo z gradniki, ki služijo dostopu do podatkovnih prostorov besedilnih dokumentov, in z gradniki za semantično analizo besedil. Orodje Orange (Demšar idr., 2013) gradi na kombinaciji vizualnega programiranja in interaktivne vizualne analitike (Sacha idr., 2017). Z vizualnim programiranjem gradimo analitične delotoke tako, da kombiniramo gradnike in jih povezujemo v smiselne in uporabne analitične postopke. Gradniki v Orangeu izvedejo branje, predobdelavo, vizualizacijo in gradnjo opisnih in napovednih modelov. Posebnost programa Orange je, da so vsi gradniki interaktivni in da se vsaka sprememba v izboru podatkov ali nastavitvi parametrov metod odrazi na spremembi izhoda iz gradnika, ta pa nadalje na vsebini, ki je posredovana vsem nižjeležečim gradnikom delotoka. Na primer, v delotoku na sliki 1 bo vsaka sprememba v seznamu besed gradnika *Word List* sprožila spremembo v vizualizaciji t-SNE oziroma kot odziv na spremembo izpostavila dokumente, ki bodo semantično ustrezali novemu seznamu pojmov. Podobno vsaka sprememba v izboru dokumentov, prikazanih z gradnikom t-SNE, sproži ponoven izračun ključnih besed in njihov prikaz v gradniku *Extract Keywords*. Vizualno programiranje in interaktivni gradniki Orangea omogočajo hitro snovanje analitičnih aplikacij in preizkus njihovega delovanja na poljubnih podatkovnih zbirkah.

Gradniki, ki so prikazani na sliki 1, so seveda samo podmnožica teh, ki jih razvijamo v projektu. V splošnem se projekt osredotoča na gradnike za dostop in branje podatkov, predobdelavo podatkov in njihove vložitve v vektorske prostore, gradnike za gradnjo seznamov zanimivih pojmov in gradnjo ter uporabo pojmovnih ontologij, gradnike za ocenjevanje in rangiranje dokumentov z ozirom na semantično podobnost z izbranimi pojmi, gradnike za vizualizacijo dokumentnih prostorov in gradnike za opise izbranih skupin dokumentov (Godec idr., 2021). Uporaba teh gradnikov seveda ni vnaprej določena. Gradniki v Orangeu so nekakšne LEGO kocke podatkovne analitike in z njimi lahko oblikujemo poljubne analitične procese.

Semantično analizo v Orangeu izvedemo predvsem z vložitvijo dokumentov in pojmov v vektorske prostore. Pri tem uporabljamo vnaprej zgrajene in naučene globoke mreže, podobnosti med dokumenti in pojmi pa potem ocenjujemo v prostorih vložitve. V sami implementaciji gradnikov so te vložitve sicer lahko vidne, a jih gradniki, če ni potrebno, ne izpostavljajo in lahko prikažejo le vsebine, ki so pomembne za uporabnika.

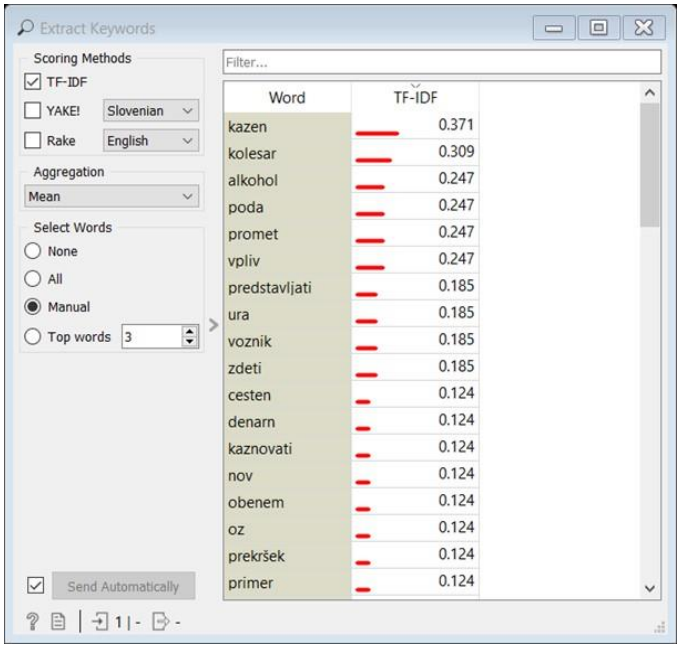
SEMANTIČNI ANALIZATOR S PRIMERI UPORABE

Semantični analizator je torej skupek gradnikov programskega sistema Orange, kot smo ga opisali zgoraj in s katerimi je z vizualnim programiranjem moč graditi poljubne aplikacije za analizo zbirk dokumentov. Analizator služi kot pripomoček za razvoj in vzdrževanje centralnega besednjaka, ki ga razvijamo in vzdržujemo na MJU. Centralni besednjak enolično in jasno določa ključno terminologijo, ki se uporablja v javni upravi. Vsi pojmi v centralnem besednjaku imajo jasno, nedvoumno in neredundantno definicijo. V centralnem besednjaku so pojmi organizirani v hierarhično strukturo. Vsak pojem je lahko v enem ali več odnosih nadrejenosti ali podrejenosti do drugih pojmov. Odnosi med pojmi vključujejo tudi asociativne (nehierarhične) odnose. Centralni besednjak vsebuje tudi druge metapodatke (npr. skrbnike pojmov, dovoljene vrednosti v obliki šifrantov, pripadajoče vire, kot so spletne storitve, ki izpostavljajo določene podatke). V besednjaku so opisane tudi podatkovne strukture ključnih registrov, slednji pa so podprti z ustreznimi zakonskimi dokumenti. Semantični analizator poskusno uporabljamo kot orodje za obdelavo teh dokumentov. Tako na primer iščemo, katere strukture v slovarju še niso opisane, ter najdemo dobre definicije in opise pojmov. Hkrati pa bi lahko pregledali, kateri dokumenti se sklicujejo na te registre in pri tem iskali morebitna neskladja.



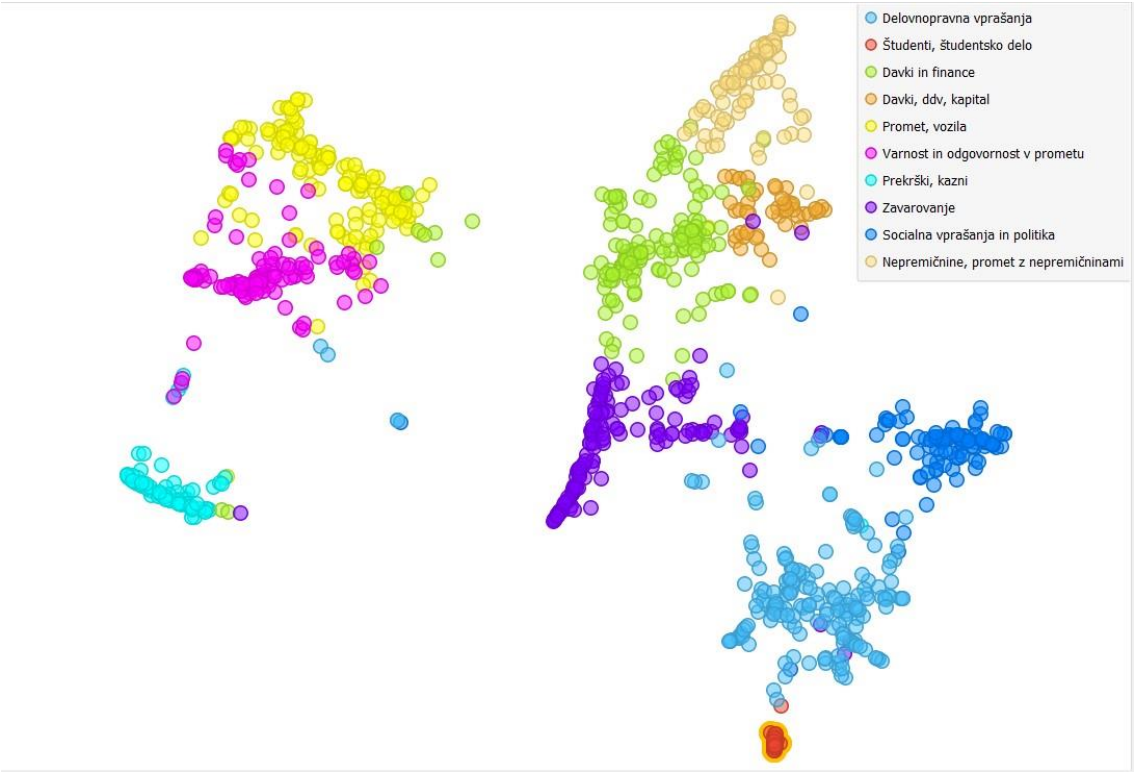
Slika 1. Primer analitičnega delotoka v Orangeu. Delotok določajo gradniki (levo zgoraj). Prikazani delotok prebere dokumente z nekaj več kot tisoč predlogi vladi RS (gradnik *Corpus*), jih predobdela (gradnika *Preprocess Text* in *Bag of Words*) ter dokumente v zbirki oceni (gradnik *Score* Documents) glede na prisotnost pojmov, ki smo jih našteli v gradniku *Word List*. Za prikaz podobnosti med dokumenti smo uporabili vizualizacijo t-SNE, kjer je vsak predlog vladi označen s točko in so predlogi, ki semantično ustrezajo naštetim pojmom iz gradnika *Word List*, izpostavljeni barvno in z velikostjo oznake. Opazimo lahko, da imamo vsaj tri skupine takih predlogov. Med njimi smo izbrali skupino zgoraj desno (točke so obrobljene z rumeno barvo) in te posredovali gradniku *Extract Keywords*, ki nam za izbrano množico dokumentov izlušči karakteristične besede.

Med razvojem semantičnega analizatorja se je izkazalo, da lahko takšno orodje zaradi njegove večopravilnosti uporabimo še na veliko drugih zanimivih in koristnih načinov. Z njim bi lahko na hiter in enostaven način v obsežnih zakonskih dokumentih iskali različne pojme, besedne zveze in podobno, saj lahko velik nabor besedil razvrščamo oz. združujemo po vsebini. Sistem vsakemu besedilu poišče nabor ključnih pojmov. Sorodnosti med besedili sistem avtomatsko pripravi glede na to, kateri in koliko pomembnih pojmov se pojavlja v več besedilih, ter tvori seznam najdenih ključnih pojmov, razvrščen po oceni pomembnosti pojma za posamezne skupine besedil (kot prikazuje slika 2).



Slika 2: Seznam najdenih ključnih pojmov za izbran nabor zakonov.

Tako lahko z ustreznimi algoritmi besedila razvrstimo po vsebini in poiščemo značilne skupine. Zgovoren prikaz sorodnosti med besedili je denimo karta besedil, kot prikazuje slika 3. Sorodna besedila so na karti prikazana s točkami v bližnji soseščini. Skupine, ki smo jih sicer v delotoku pred to vizualizacijo odkrili z algoritmi razvrščanja, so prikazane z različnimi barvami. Orodje dopolnjujejo algoritmi, ki ključne pojme razširjajo na bolj povedne konstrukte, s katerimi lahko vsebino besedila natančneje opredelimo.



Slika 3: Zemljevid besedil dokumentov s predlogi vladi RS s prikazom skupin

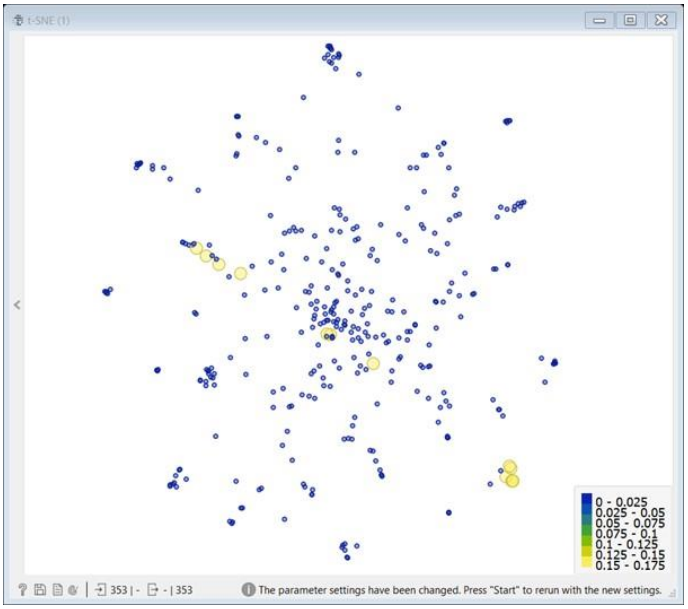
Lahko rečemo, da besedila znotraj iste skupino govorijo o sorodnih vsebinah. Če med besedili iščemo tista, ki govorijo o neki vsebini, zadošča, da pregledamo le besedila v ustrezni skupini in tako močno skrčimo število in obseg besedil, ki bi jih sicer moral uporabnik v celoti natančno pregledati. Med izbranimi se je smiselno osredotočiti na tista besedila, ki so na karti narisana bolj skupaj. Tako si lahko učinkovito pomagamo, ko iščemo besedila, ki govorijo o isti vsebini kot neko dano besedilo. Poiskati moramo le, v katero skupino sodi iskano besedilo.

V primeru iz slik 2 in 3 smo uporabili vzorčni nabor besedil iz javne zbirke "Predlagam vladi" v povezavi z vzorčnim naborom zakonskih besedil, ki vsebujejo besedo "register". Preveriti smo hoteli, ali lahko orodje pomaga pri iskanju zakonov, ki so povezani z izbranim predlogom vladi. Kot primer smo uporabili vzorec podatkov iz javne zbirke "Predlagam vladi"[1], ki na dan 15. 9. 2021 vsebuje 11.471 dokumentov oz. predlogov državljanov in drugih subjektov ter 3528 odzivov nanje. Zraven smo dodali vzorec 353 zakonskih besedil kot vir črpanja možnih podlag za odgovore na vprašanja oziroma problematiko iz predlogov.

Z orodjem najprej v novo prejetem predlogu, ki prispe v javno zbirko "Predlagam vladi", poiščemo ključne pojme, ki dovolj dobro opredeljujejo vsebino tega predloga. Na osnovi primerjave ključnih pojmov v ostalih, že prejetih predlogih, ki jih javna zbirka "Predlagam vladi" vsebuje, lahko hitro pregledamo, ali smo že kdaj obravnavali primere s sorodno vsebino in denimo uporabimo odzive, ki so

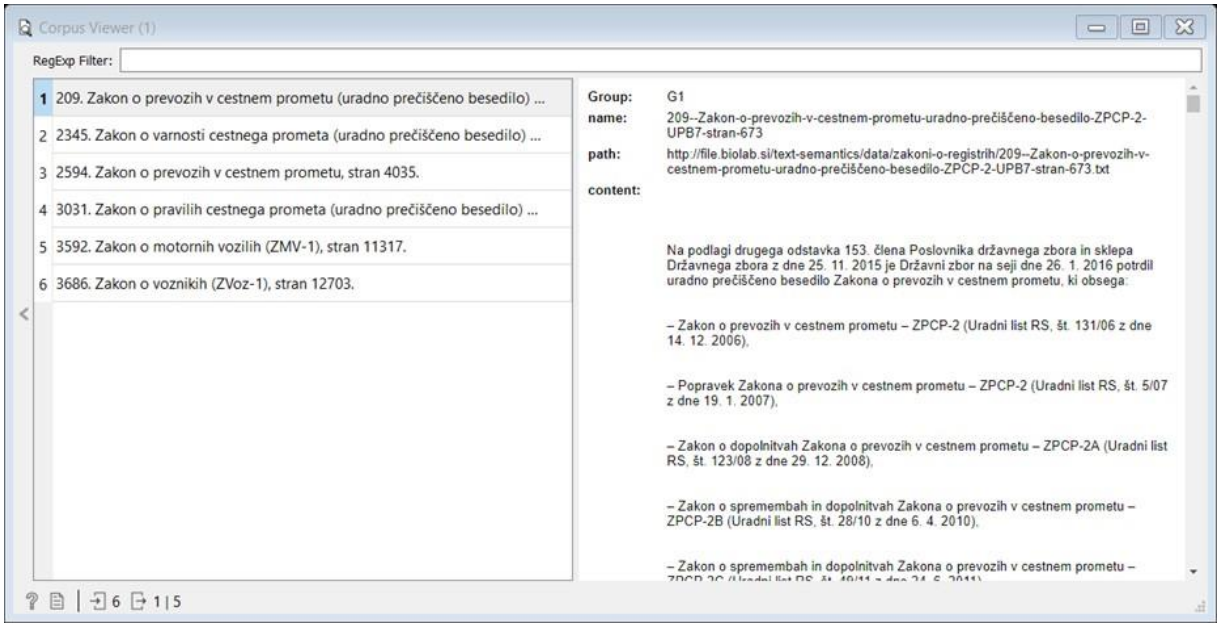
[1] https://predlagam.vladi.si/

že bili pripravljeni nanje. To razberemo iz karte besedil tako, da nov predlog pripada eni od skupin, pri čemer se umesti blizu drugim besedilom, ki se že nahajajo v tej skupini kot prikazujeta slika 3 (zbirka Predlagam vladi) in slika 4 (zbirka Zakoni z označenimi območji največje vsebinske sorodnosti z izbranim(i) besedili zbirke Predlagam vladi). Na sliki 4 so z rumeno barvo označeni tisti zakoni, ki imajo najbolj sorodno vsebino z novo prispelim predlogom v javni zbirki "Predlagam vladi", na katerih bi lahko po vsebini utemeljili odziv na prejeti predlog. V primeru, da je novo prejeti predlog izviren, bi se na zemljevidu, kot ga prikazuje slika 3, pokazal odmaknjen od drugih predlogov in ne bi pripadal nobeni od skupin. V tem primeru bi takoj vedeli, da bo potrebno pripraviti nov odziv in zanj poiskati ustrezno zakonsko podlago, ker se takšna vsebina v javni zbirki "Predlagam vladi" še ne nahaja.



Slika 4: Karta zakonov z označenimi dokumenti, ki so vsebinsko podobni izbranemu predlogu vladi RS

Nadalje orodje generira seznam ključnih besed novega predloga v zbirki zakonskih besedil, kjer poiščemo tista, ki temu seznamu najbolj ustrezajo, kot prikazuje slika 5. Na podoben način orodje generira seznam ključnih besed predloga v javni zbirki "Predlagam vladi", s čimer lahko vidimo najbolj ustrezne potencialne vsebine.



Slika 5: Podrobnejši vpogled v izbrana zakonska besedila

Orodje omogoča, da lahko z izbranim naborom ključnih besed po isti vsebini pregledujemo različne nabore/zbirke besedil. Tako lahko isto vsebino osvetlimo z različnih področij. Praktična vrednost orodja narašča s količino besedil, ki jih moramo upoštevati pri reševanju naloge oz. problema. Zato so za uporabo orodja pomembni vsebinska in oblikovna celovitost ter verodostojnost besedil in zbirke, pri čemer je treba spodbujati in nuditi ustrezno podporo upravljavcem zbirk, da jih opremijo skladno s potrebami digitalne vizije na podlagi podatkovne ekonomije. Uporabnik naj bi se ukvarjal samo z dobljeno analizo besedil in vsebinskim reševanjem problema. Kot že izhaja iz prikazanih možnosti uporabe, uporabnost orodja narašča tudi s povezovanjem različnih zbirk, torej z iskanjem vsebinskih sorodnosti med različnimi zbirkami, s čimer se srečujemo pri reševanju vsakodnevnih nalog in problemov.

ZAKLJUČKI

V projektu izgradnje semantičnega analizatorja razvijamo zbirko analtičnih gradnikov, s katero je moč razviti prototipe aplikacij za razvoj preglednih in strokovnih postopkov upravljanja z besedili, ki tipično nastopajo v javni upravi. Gradniki, ki smo jih razvili, so uporabni tako pri analizi zbirk slovenskih kot tujih besedil. V prvih primerih uporabe se izkaže, da tudi za relativno kompleksna opravila zadošča manjša skupina analitičnih gradnikov, namenjenih dostopu do besedilnih dokumentov, vnosu gesel in njihovi organizaciji v ontologiji, iskanju podobnosti med dokumenti in gesli in vizualizacijam dokumentov in dokumentnih prostorov.

Javna uprava shranjuje in ustvarja velike količine besedil in dokumentov, zato so semantični analizator in druga podobna orodja, ki znajo na enostaven način obdelovati velike količine besedil in dokumentov iz različni virov, korak v smer poenostavitve, optimizacije in avtomatizacije razumevanja besedil in obladovanja procesov, ki ta besedila obravnavajo. Razvoj orodij, kot smo jih predstavili v pričujočem prispevku, je potreben za nadaljnji razvoj analitičnih tehnik na področju analize besedil in razvoj

uporabniških vmesnikov, ki domenskim ekspertom omogočajo dostop do analitike. Orodja, kot je semantični analizator, podpirajo razvoj podatkovne ekonomije in digitalizacije v širšem smislu ter ciljajo na demokratizacijo umetne inteligence (Godec idr., 2019).

VIRI IN LITERATURA

[1] Demšar J, Curk T, Erjavec A, …, Zupan B (2013) Orange: data mining toolbox in Python, *Journal of Machine Learning Research* 14: 2349-2353.

[2] European Comission (2020) Communication from the Commission to the European Parliament, the Council, the

[3] European Economic and Social Committee and the Committee of the Regions. A European Strategy for Data (COM(2020) 66 final) z dne 19. februarja 2020, str. 22-23, https://eur-lex.europa.eu/legalcontent/EN/ALL/?uri=CELEX%3A52012DC0673.

[4] Godec P, Đukić N, Pretnar A, Tanko V, Žagar L, Zupan B (2021) Explainable Point-Based Document Visualizations. *International Workshop on eXplainable Artificial Intelligence in Healthcare*, AIME 2021.

[5] Godec P, Pančur M, Ilenič N, Čopar A, Stražar M, Erjavec A, Pretnar A, Demšar J, Starič A, Toplak M, Žagar

[6] L, Hartman J, Wang H, Bellazzi R, Petrovič U, Garagna S, Zuccotti M, Park D, Shaulsky G, Zupan B (2019) Democratized image analytics by visual programming through integration of deep models and small-scale machine learning, *Nature Communications* 10(1): 4551.

[7] Kern Pipan K, Jesenko M, Lozej M, Jesenko P (2020). Izzivi in perspektiva upravljanja podatkov v javni upravi z vidika uporabe naprednih tehnologij, *Informatika v javni upravi* 2020, Zbornik konference.

[8] OECD (2019) The Path to Becoming a Data Driven Public Sector. https://www.oecd.org/gov/the-path-tobecoming-a-data-driven-public-sector-059814a7-en.htm (dostop 13. 09. 2021).

[9] Pedregosa F, Varoquaux G, Gramfort A, ..., Duchesnay É (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12: 2825–2830.

[10] Provost F, Fawcett FT (2013) Data Science and its Relationship to Big Data. *Big Data* 1(1).

[11] Sacha D, Sedlmair M, Zhang L, ..., Keim DA (2017) What you see is what you can change: human-centered machine learning by interactive visualization. *Neurocomputing* 268: 164–175.