Semantični analizator – razvoj programskega okolja za algoritmično obdelavo slovenskih besedil

Miha Jesenko[1], Miro Lozej[1], Karmen Kern Pipan[1], Primož Godec[2], Vesna Tanko[2], Lan Žagar[2], Ajda Pretnar Žagar[2], Nikola Đukić[2], Blaž Zupan[2]
[1] Ministrstvo RS za javno upravo (MJU), Tržaška 21, 1000 Ljubljana
[2] Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Večna pot 113, 1000 Ljubljana
Miha.Jesenko@gov.si, Miro.Lozej@gmail.com, Karmen.Kern-Pipan@gov.si, primoz.godec@fri.uni-lj.si, Vesna.Tanko@fri.uni-lj.si, lan.zagar@fri.uni-lj.si, ajda.pretnar@fri.uni-lj.si, nikoladjukic.djukic2@gmail.com, Blaz.Zupan@fri.uni-lj.si

Izvleček

Uslužbenci in funkcionarji v javni upravi se dnevno srečujejo s številnimi obsežnimi dokumenti, ki jih je treba pregledati in uporabiti glede na informacijske zahteve konkretne naloge. To velja pri pripravi odločitev, pripravi zakonodaje in politik, pregledovanju zakonodaje in politik, ocenjevanju učinkov zakonodaje in politik, pri raznih analizah, pri opisovanju podatkovnih virov in storitev ter pri številnih drugih nalogah. Ker pregledovanje množice dokumentov in izbor uporabnih dokumentov glede na naše potrebe pomeni velik časovni vložek, smo oblikovali pristop na podlagi umetne inteligence za vsebinsko pregledovanje velikih zbirk besedil. Pristop s semantično analizo besedil ter primerjavo vsebinske sorodnosti med posameznimi besedili v zbirki omogoča časovni prihranek in celovito analizo zbirk. V prispevku predstavimo prve rezultate 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 vključuje 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.

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

Semantic Analyser - Development of a software environment for algorithmic processing of slovenian texts

Abstract

Every day, civil servants and officials are confronted with a large number of voluminous documents that need to be reviewed and applied according to the information requirements of a specific task. This is the case when making decisions, drafting legislation and policies, reviewing legislation and policies, assessing the impact of legislation and policies, carrying out various analyses, describing data sources and services and many other tasks. Since reviewing many documents and selecting the most relevant ones for our needs is a time-consuming task, we have developed an AI-based approach for the content-based review of large collections of texts. The approach of semantic analysis of texts and the comparison of content relatedness between individual texts in a collection allows for time-saving and the comprehensive analysis of collections. In the paper, we present the results of the project to develop a general-purpose tool for analysing sets 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.

Keywords: Semantic analysis, data spaces, text mining, visual analytics, workflows

1 UVOD

Javna uprava je, podobno kot preostala področja človeške družbe, vedno bolj usmerjena v podatke – njihovo zbiranje, obdelavo, preverjanje in razumevanje. Z razvojem novih tehnologij ter znanosti pridobiva vedno več podatkov, za njihovo obdelavo pa mora ob tem razviti primerna orodja in nove pristope.

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 večjega števila besedil, razvrščanju po vsebini ter semantičnem preiskovanju, kjer iščemo dokumente, ki so vsebinsko povezani z izbranimi pojmi.

Delo z večjimi zbirkami besedil je zelo pogosto v javni upravi in je ena izmed nujnih sestavin pri različnih aktivnostih. Pogosto je treba pregledovati celotne zbirke in izbrati ter uporabiti primerna besedila glede na konkretno vsebino v postopkih priprave zakonodaje, politik in drugih strateških ter krovnih dokumentov. Podobno velja pri pripravi podlag v postopkih odločanja, pripravi odgovorov na razna vprašanja in pobude, v postopkih ocenjevanja učinkov zakonodaje, politik in strategij ter pri vrsti drugih nalog. Ena izmed takih nalog je vzpostavitev kataloga podatkov javne uprave (s poudarkom na katalogu temeljnih registrov in evidenc ter šifrantov) v okviru upravljanja semantične interoperabilnosti v javni upravi, pri čemer je večina pojmov in njihovih lastnosti opredeljena in razpršena v zakonodaji. V omenjenih nalogah na podlagi vsebine, ki izhaja iz vhodnega dokumenta (oziroma pogosteje dela dokumenta), iščemo vsebinsko sorodne dele v večji zbirki besedil, s čimer se lahko izognemo podvajanju rešitev ali celo ustvarjanju (ali ugotavljanju) nasprotnih rešitev, predvsem pa zmanjšamo nabor besedil, ki so relevantna za naše informacijske potrebe. Hkrati zajamemo besedila, ki na prvi pogled niso relevantna, se pa v delih vsebine »skrivajo« uporabne informacije. Tak primer uporabe v javnem sektorju je analiza zakonskih aktov in predlogov vladi. 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 kakovost 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 pomagajo tako zaposlenim v javni upravi kot državljanom.

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.

2 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, z namenom da se uporablja podatke na čimbolj enoten, varen in zaupanja vreden način (da se lahko zaneseš na celovitost, pravilnost, verodostojnost in ažurnost podatkov) za ustvarjanje novih dodanih vrednosti na njihovi podlagi. Standardizirani in kakovostni procesi upravljanja zagotavljajo tudi hitrejšo ter bolj učinkovito uporabo podatkov v novih tehnologijah in podatkov iz novih tehnologij. Iz standardiziranih procesov upravljanja hitro pridemo do ideje skupnih prostorov za podatke, ki se urejajo skozi take procese. Podatkovni prostori predstavljajo idejo pospešitve podatkovne ekonomije (OECD, 2019, 2. Data governance in the public sector; British Academy and the Royal Society, 2017; Centre for International Governance Innovation, 2018; Micheli, Ponti, Craglia, Berti Suman, 2020). Tako je EU lansko leto napovedala vzpostavitev podatkovnih prostorov za področje javne uprave poleg preostalih osem začetnih področij (industrija, zeleni dogovor, mobilnost, zdravje, finance, energija, kmetijstvo in veščine) (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 omogočajo učinkovitejšo uporabo podatkov z novimi podatkovnimi tehnikami v podporo odločanju (različna enostavna razvrščanja, oblikovanje in priprava kriterijev ter vzorčenja). Kako natančno naj bi podatkovni prostori izgledali oziroma kaj vse naj bi vključevali podatkovni prostori, še niso v celoti odgovorjena vprašanja. Glede bistvenih funkcionalnostih pa se je izoblikoval določen osnovni pogled na logično arhitekturo in načela, ki naj bi jo zasledovali podatkovni prostori. Vsekakor ne gre samo za podatke, ampak tudi za varnost, različne vloge udeležencev, pravice za dostop do pravno varovanih podatkov, upravljanje (procesi upravljanja) za zagotavljanje celovitosti in kakovosti podatkov ter samih prostorov, standarde za semantično opisovanje podatkov, kataloge podatkov in storitev, servise za dostop, obdelavo in izmenjavo podatkov ipd. (Matranga, 2021; Nagel, 2021; OPEN DEI Task Force 1, 2021). Podatkovni prostori morajo sloneti na načelih FAIR[1]. Zagotavljati morajo interoperabilnost v vseh pogledih, tako tehnično in semantično kot tudi organizacijsko ter pravno. Prav odločanje na podlagi dejstev in podatkov je ideal, h kateremu stremi vsaka napredna organizacija, tudi javna uprava (OECD, 2018, 3. The application of data in the public sector to generate public value). V za-

[1] https://www.go-fair.org/fair-principles/.

dnjih letih smo priča premiku usmerjenosti razvoja informacijskih sistemov od aplikacij k uporabi podatkov za pridobivanje informacij. Podatkovna orodja so tista (Stavrianou, Andritsos in Nicoloyannis 2007), ki omogočajo hitro pridobivanje informacij iz večjih količin nepovezanih podatkov, ki so sicer običajnemu uporabniku težje dostopni.

Javna uprava v delovnem procesu obdela veliko število besedilnih dokumentov, (Hollibaugh 2019) med katerimi moramo poiskati tiste, ki govorijo o neki vsebini in jih je treba pregledati, da bi dobro utemeljili predloge ali celo odločitve. Tipičen primer so denimo zakonska besedila oz. iskanje zakonskih dokumentov, ki po vsebini obravnavajo želeno vsebino. Iskanje po vsebini bi nam dostavilo 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.

Za uspešno preiskovanje besedil so potrebni novi načini predstavitve znanj in luščenja informacij. Peppes idr. (2020) na primer predlagajo orodje za podporo raziskovanju kriminala. Njihov semantični iskalnik (Semantic Engine) omogoča primerjavo oseb iz podatkovne baze na podlagi podobnosti ter prikaz ontologij. Jain, Seeja in Jindal (2020) uporabijo pristop latentne semantične analize za izračun semantične podobnosti besed. Za luščenje kandidatnih sosednjih besed podani besedi pa uporabijo mehko analizo formalnih konceptov. Konceptualno je uporaba semantičnih tehnologij v javni upravi podobna problemu iskanja ekspertnih odgovorov v nalogah skupnostnega odgovarjanja na vprašanja (CQA). Liu idr. (2022) uporabijo večnivojsko semantično analizo za iskanje domenskih ekspertov, ki bi lahko odgovorili na uporabniška vprašanja. Večnivojska analiza je sestavljena iz grobega tematskega modeliranja ter finega BERT modela, s čimer zajamejo domenske informacije vprašanj in uporabnikov. Na ta način točneje lahko predlagajo kandidatne eksperte, podobno pa bi lahko v javni upravi predalagali kandidatne odgovore ali direktive.

Za hitro pregledovanje besedil in iskanje sorodnih besedil je pomembno tudi luščenje ključnih pojmov iz besedila ali množice besedil. Luščenje ključnih pojmov je naloga, pri kateri ključne besede ali ključne fraze izberemo med besedami in besednimi zvezami v dokumentu (Campos idr., 2020). Za naš namen se osredotočamo na nenadzorovane pristope, ki so neodvisni od domene in ne potrebujejo označenega nabora podatkov. Med temi je osnovni pristop TF-IDF (Jones, 1972), ki ocenjuje pomembnost besed v dokumentu glede na celoten korpus. Izračuna pogostost izraza v dokumentu uteženo s frekvenco v celotnem korpusu. Naprednejši pristop YAKE! (Campos idr., 2020) uporablja statistične značilnosti, kot sta položaj in pogostost besed, informacije o kontekstu in razširjenost izraza v dokumentu. V nasprotju s TF-IDF ta lušči ključne pojme na podlagi enega dokumenta in ne potrebuje velikega korpusa. Metode, ki temeljijo na grafih, zgradijo graf sosednjih pojmov v dokumentu in uporabljajo metode za točkovanje pojmov v grafu. RAKE (Rose idr., 2010) zgradi graf sosednjih pojmov in za točkovanje uporablja pogostost izrazov, stopnjo izraza v grafu ali kombinacijo obojega. Pristopi globokega učenja temeljijo na vektorskih vložitvah pojmov. Ti vložijo pojme in dokument v isti vektorski prostor in kot ključne pojme izberejo tiste z največjo podobnostjo vložitvi dokumenta (Bennani-Smires idr., 2018).

Uporaba semantičnih tehnologij na področju javne uprave je še dokaj novo področje (Eggers, Gracie, Malik idr., 2018), zlasti v Sloveniji. Ob pričetku projekta smo morali preveriti, ali so podatki, ki so na voljo, primerni, odkriti zahteve uporabnikov in preskusiti kako in ali jim je moč zadostiti s trenutno znanimi tehnologijami. Uporabniške zahteve smo identificirali 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. Taki sistemi so na primer komercialna KNIME (https://www.knime.com) in RapidMiner (http://rapidminer.com) ter prosto dostopni in odprti Orange (http://orangedatamining.com).

Projekt predvideva izdelavo gradnikov za dostop do podatkovnih prostorov dokumentov, gradnike za pripravo iskalnih pojmov in gradnjo ontologij ter gradnike za iskanje karakterističnih izrazov v doku-

mentih. Pomembno je, kako ti gradniki predstavijo rezultate analize in ali je z njimi moč zgraditi delotoke, 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.

3 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) temelji 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 odraža v spremembi izhoda iz gradnika, ta pa nadalje na vsebini, ki je posredovana vsem nižje-ležečim gradnikom delotoka. Na primer v delotoku na sliki 1 (poglavje 4, str. 9) 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 Orange-a omogočajo hitro snovanje analitičnih delotokov 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 Orange-u izvedemo predvsem z vložitvijo dokumentov in pojmov v vektorske prostore. Vložitev dokumentov v vektorske prostore pomeni matematični opis dokumenta glede na besede, ki se v njem pojavljajo. Pri tem uporabljamo vnaprej zgrajene in naučene globoke mreže, podobnosti med dokumenti in pojmi pa potem ocenjujemo v prostorih vložitve. Globoke mreže, ki jih uporabimo v rešitvi, so modeli vložitev besed za slovenščino *fastText* (Joulin idr. 2016) ter BERT (Devlin idr. 2018). *fastText* model je bil naučen na *Common Crawl* in Wikipedia besedilih, BERT pa na strojno prevedenih korpusih *BooksCorpus* in Wikipedia. Vsak model nabor pojmov iz korpusa predstavi v svojem semantičnem vektorskem prostoru, nato pa za povprečno vložitev dokumenta poiščemo nabor najbližjih pojmov. S tem identificiramo tiste pojme, ki so izbranemu dokumentu semantično najbližje.

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.

4 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. Razvoj in vzdrževanje centralnega besednjaka je del projekta Tehnične in semantične prenove temeljnih registrov in evidenc v javni upravi, s katerim se želi v javno upravo vpeljati semantično interoperabilnost na bolj standardiziran in metodološki način. Semantična interoperabilnost koristi različnim uporabnikom, da poznajo ali poiščejo definicije in opise pojmov z namenom čimbolj učinkovite in nedvoumne medsebojne komunikacije. Po drugi strani pa je zelo pomembna tudi za informacijske sisteme, da lahko samodejno komunicirajo med sabo na podlagi semantičnih oznak, ki vsebinsko povezujejo sorodne podatke. Centralni

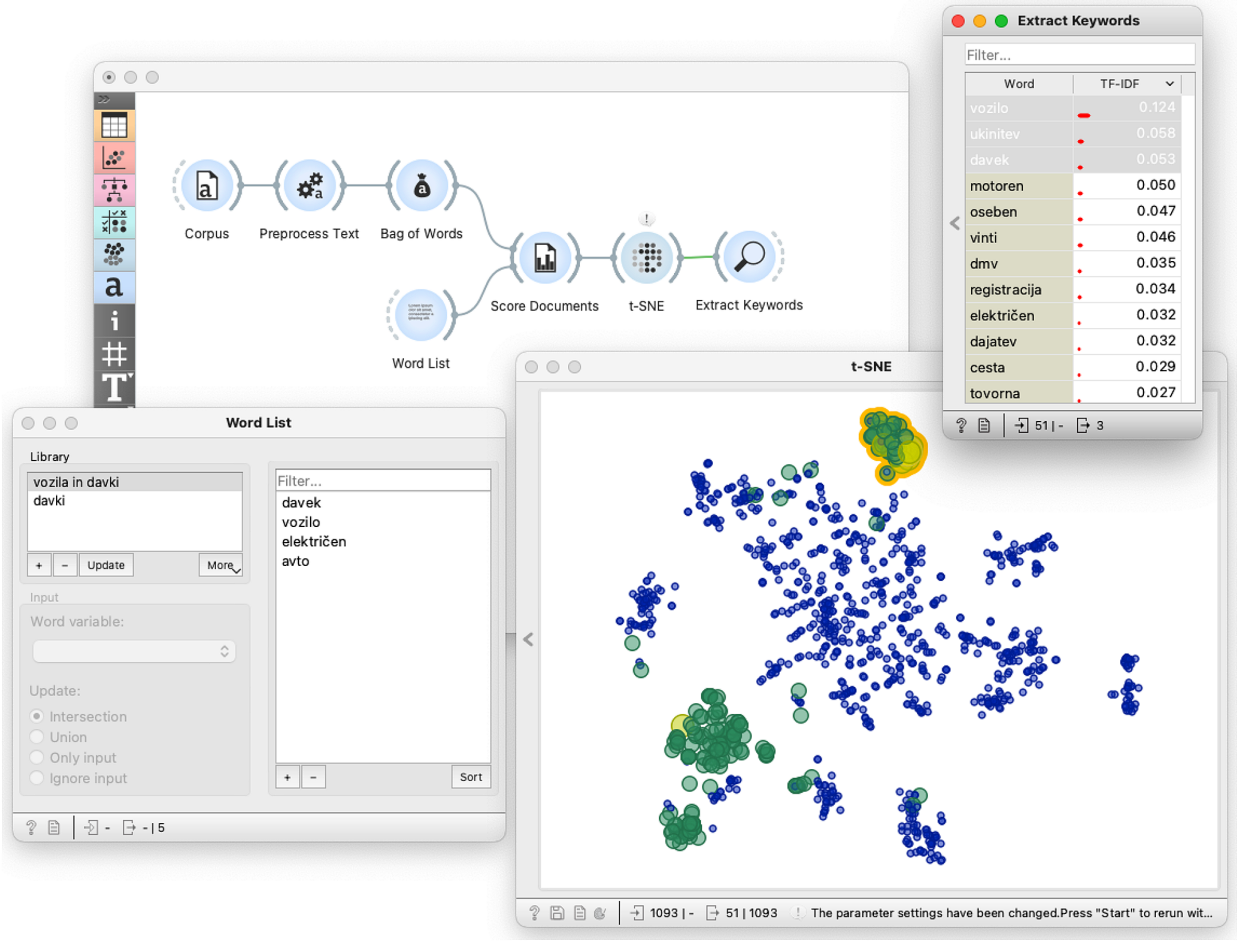
besednjak v formatu semantičnega spleta OWL (na osnovi RDF) enolično in jasno določa ključno terminologijo, ki se uporablja v javni upravi. Vsi pojmi v centralnem besednjaku imajo jasno, nedvoumno in neredundantno definicijo. V centralnem besednjaku so pojmi organizirani v hierarhično strukturo. Vsak 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. Registri in evidence ter druge zlasti formalne podatkovne zbirke so seznami subjektov ali stvari, ki so opisani z določenimi podatki in vpisanemu dajejo določene pravice. Vzpostavitev, oblika, način upravljanja in uporabe registra so določeni z zakonom (npr. Zakon o matičnem registru), medtem ko je vsebina, delovno področje registra (torej dejanske kategorije vpisov v registru-pojmi in njihovi opisi/značilnosti) prav tako urejeno z zakonom oziroma več zakoni (npr. Družinski zakonik (v razmerju do matičnega registra), za definicije določenih pojmov iz registra tudi npr. Zakon o osebnem imenu itd.). Semantični analizator poskusno uporabljamo kot orodje za obdelavo zakonskih dokumentov. Tako na primer iščemo po zakonih, katere strukture v slovarju še niso opisane, ter najdemo dobre definicije in opise pojmov. Hkrati bi lahko pregledali, kateri dokumenti se sklicujejo na te registre oziroma urejajo sorodno vsebino in pri tem iskali morebitna neskladja. S tega vidika je bila izbira zakonov kot prve zbirke za analizo v semantičnem analizatorju najbolj primerna, saj vsebuje največ informacij, ki jih moramo prenesti, zmodelirati v besednjakih (ontologijah), ki opisujejo določeno področje.

Med razvojem semantičnega analizatorja se je izkazalo, da lahko takšno orodje zaradi njegove večopravilnosti uporabimo še na veliko drugih zanimivih in koristnih načinov, kar smo želeli tudi praktično preizkusiti. Glede na že izbrano zbirko zakonov, smo poiskali drugo primerno zbirko, ki se vsebinsko močno povezuje z zakonodajo in jo je razmeroma enostavno pripraviti za semantično analizo. Na ta način smo prišli do zbirke Predlogi vladi, ki vsebuje predloge posameznikov vladi, da reši določen problem, ki so ga zaznali. Uslužbenec, ki se ukvarja s temi predlogi, mora nanje odgovoriti, pri čemer mora najprej ugotoviti, ali v zbirki že obstajajo sorodni predlogi in če na kakšno vprašanje že obstaja odgovor. V nasprotnem primeru ga najprej čaka naloga, da poišče vsebinsko relevantne zakone, na podlagi in v skladu s katerimi potem napiše odgovor. S semantičnim analizatorjem 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.

V primeru s slik 1, 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« , ki na dan 15. 9. 2021 vsebuje 11.471 dokumentov oz. predlogov državljanov in drugih subjektov ter 3.528 odzivov nanje. Zraven smo dodali vzorec 353 zakonskih besedil kot vir črpanja možnih podlag za odgovore na vprašanja oziroma problematiko iz predlogov.

Prikazani delotok na sliki 1 (spodaj) prebere dokumente z nekaj več kot tisoč predlogi vladi RS (gradnik *Corpus*, pri čemer lahko namesto zbirke predlogi vladi enostavno izberemo zbirko zakoni), 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.

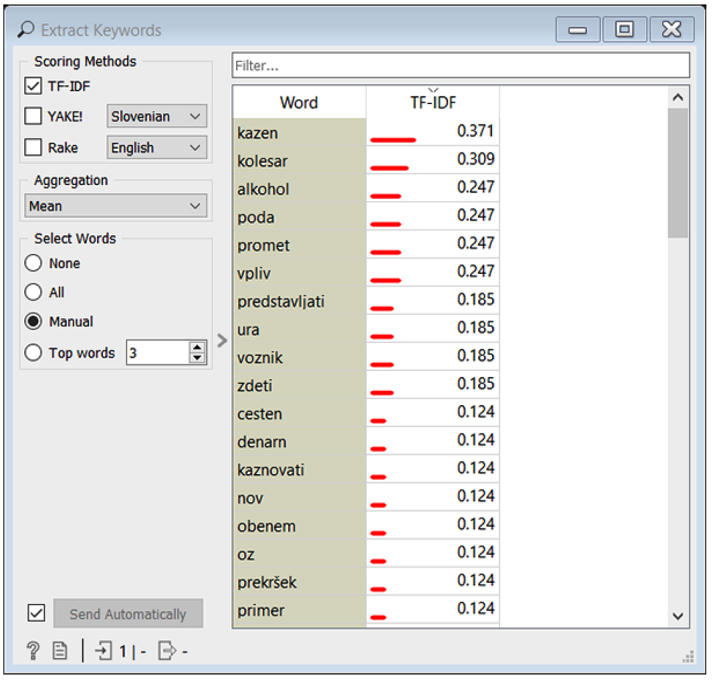
Sistem vsakemu besedilu poišče nabor ključnih pojmov. Sorodnost med besedili sistem avtomatsko pripravi glede na to, kateri in koliko pomembnih pojmov se pojavlja v več besedilih, ter tvori seznam najdenih ključnih pojmov, razvrščen po oceni pomembnosti pojma za posamezne skupine besedil. Primer iz slike 2 prikazuje ključne pojme predloga z naslovom Milejše kaznovanje kolesarjev pod vplivom alkoho-



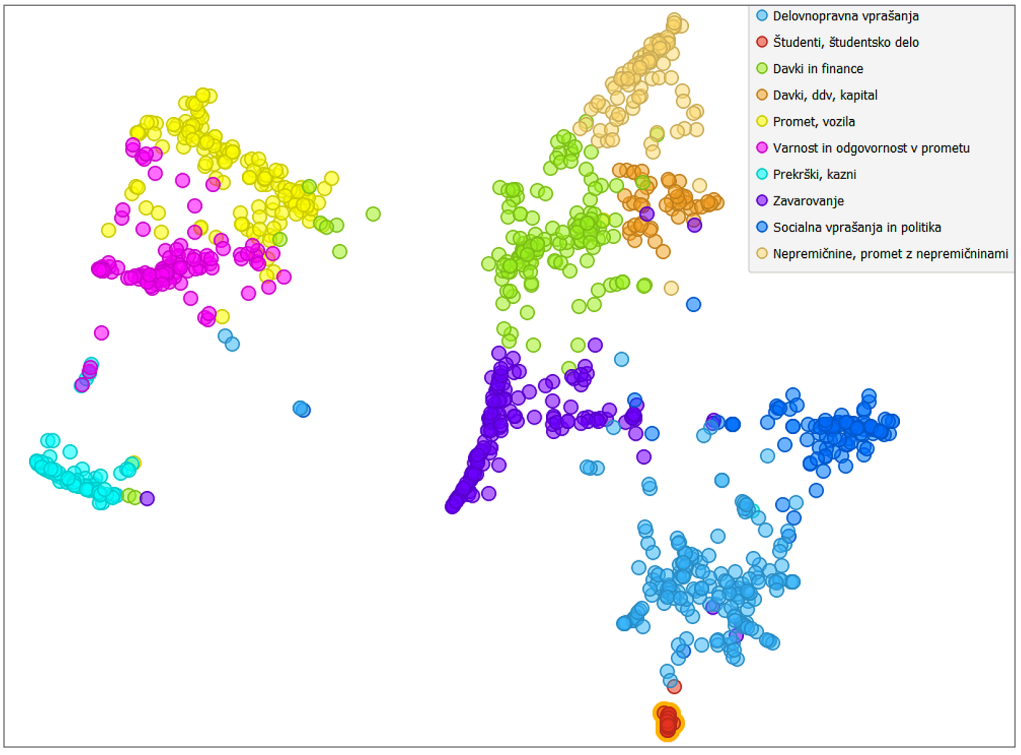
Slika 1: **Primer analitičnega delotoka v Orange-u. Delotok na Sliki 1 določajo gradniki (levo zgoraj).**

la. Z izbranimi ključnimi besedami predloga lahko semantični analizator v zbirki zakonov poišče dokumente, ki so po vsebini najbolj sorodni izbranemu predlogu, kar bo opisano v nadaljevanju poglavja.

Tako lahko z ustreznimi algoritmi besedila razvrstimo po vsebini in poiščemo značilne skupine. Zgovoren prikaz sorodnosti med besedili je denimo karta besedil zbirke Predlogi vladi, 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.



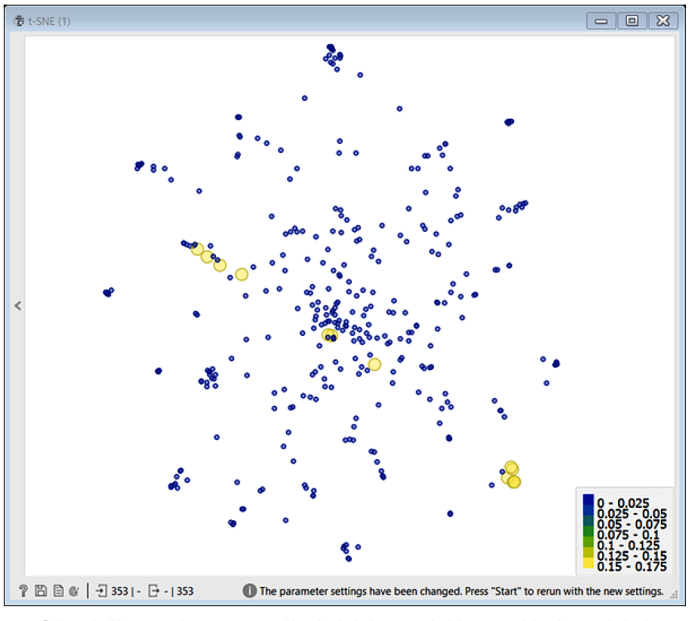
Slika 2: **Seznam najdenih ključnih pojmov za izbran predlog vladi z naslovom Milejše kaznovanje kolesarjev pod vplivom alkohola.**



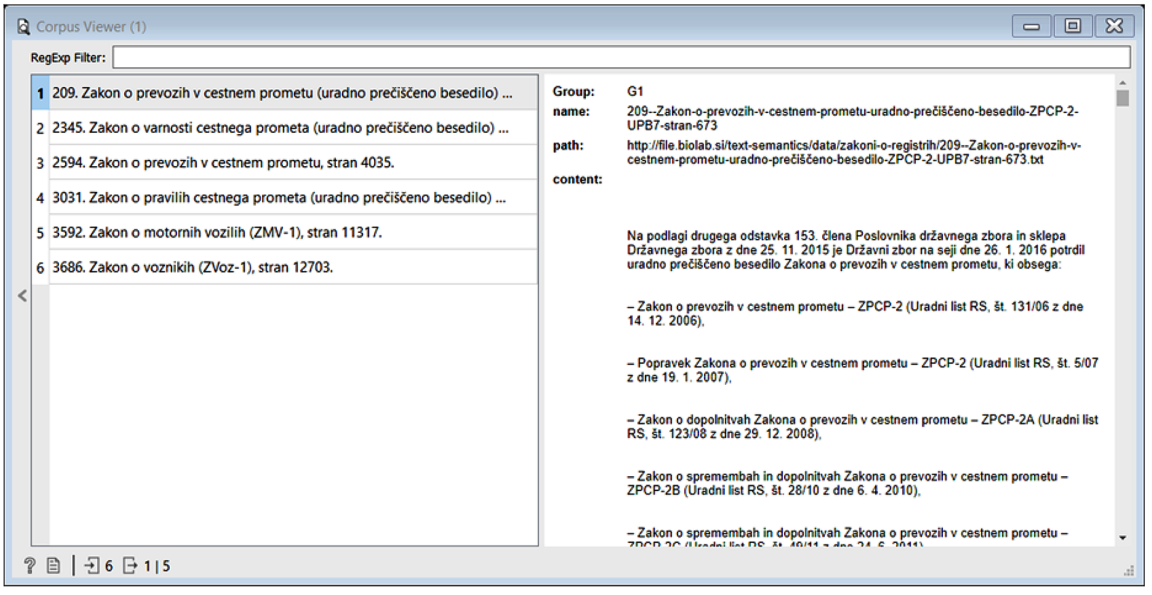
Slika 3: **Zemljevid besedil dokumentov s predlogi vladi RS s prikazom skupin**

Lahko rečemo, da besedila znotraj iste skupine 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 skupaj. Tako si lahko učinkovito pomagamo pri iskanju besedil, ki govorijo o isti vsebini kot dano besedilo. Poiskati moramo le, v katero skupino sodi izhodiščno besedilo.

Z orodjem najprej v novo prejetem predlogu, ki prispe v javno zbirko »Predlagam vladi«, poiščemo ključne pojme, ki dovolj dobro opredeljujejo vsebino 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 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 izbranimi besedili zbirke Predlagam vladi). Na sliki 4 so z rumeno barvo označeni tisti zakoni, ki imajo vsebino najbolj sorodno predlogu z naslovom Milejše kazno-



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



Slika 5: **Podrobnejši vpogled v izbrana zakonska besedila**

vanje kolesarjev pod vplivom alkohola v javni zbirki »Predlagam vladi« in s katerimi 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, ta 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.

Nadalje orodje generira seznam ključnih besed novega predloga v zbirki zakonskih besedil, kjer poiščemo tista, ki 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 najustreznejše potencialne vsebine.

Orodje omogoča, da lahko z izbranim naborom ključnih besed 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, posodobljenost 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. Vsaka zbirka potrebuje skrbnika in postopek za vzdrževanje zbirke. V praksi imamo tudi povečini povezane zbirke in ne zgolj ene, ki jih uporabljamo za določene naloge. Tako že npr. pri uporabi zakonov takoj trčimo ob vprašanje, kje so podzakonski predpisi. Govorimo o uporabi zakonodaje, pri čemer se nam nadalje takoj odprejo vprašanja, ali ne bi bilo smiselno vključiti tudi evropske zakonodaje in mogoče tudi pripravljalnih aktov zakonodaje ipd. Z vsebinskim preizkušanjem orodja je treba preveriti kakovost analize, še posebej kadar gre za dokumente, ki opisujejo različna vsebinska področja, če so zajeti značilni pojmi vseh področij in da se dejansko poiščejo vsi sorodni dokumenti. Glavni namen orodja je zmanjšati količino časa, ki ga porabimo za iskanje vseh dokumentov, ki nam lahko ponujajo odgovore na določena vprašanja. Poleg tega je treba paziti pri dokumentih z malo vsebine, saj je analiza boljša, če je na voljo več besedila. Na podlagi zadovoljstva z rezultati bomo lahko takim orodjem vedno bolj zaupali.

Pri več zbirkah se srečujemo s potrebami, da pri novem vhodnem dokumentu preverimo, ali v zbirki že obstaja dokument z enako vsebino, ali obstajajo dokumenti s podobno vsebino in določena informacija, kako podobni so si (npr. zelo podobni, podobni, malo podobni) dokumenti. Iskanje duplikatov, sorodnih ali popolnoma novih dokumentov je zagotovo pogosto željena funkcionalnost pri različnih uporabnikih. Podobna pogosta funkcionalnost je tudi oženje nabora relevantnih dokumentov iz velikih zbirk na zgolj tiste, ki so vsebinsko povezani z našo informacijsko potrebo. Omenjene in podobne funkcionalno-

sti bi bile uporabne pri določenih vsakodnevnih oz. obdobnih opravilih v javni upravi, pri čemer take naloge opravljajo javni uslužbenci, ki so bolj ali manj usposobljeni na informacijskem področju, vsekakor pa ne gre za podatkovne znanstvenike (ali zelo redko) ali osebe, ki se ukvarjajo z analizo naravnega jezika oz. analizo podatkov in besedil.

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.

5 ZAKLJUČKI

V projektu izgradnje semantičnega analizatorja razvijamo zbirko analtičnih gradnikov, s katero je moč razviti prototipe delotokov 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] Bennani-Smires, K., Musat, C., Hossmann, A., Baeriswyl, M., & Jaggi, M. (2018). Simple Unsupervised Keyphrase Extraction using Sentence Embeddings. *Proceedings of the 22nd Conference on Computational Natural Language Learning, 221–229.*

[2] British Academy and the Royal Society (2017). Data management and use: Governance in the 21st century. *https://royalsociety.org/~/media/policy/projects/data-governance/data-management-governance.pdf* (dostop 5.5.2022).

[3] Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences, 509, 257-289*.

[4] Centre for International Governance Innovation (2018). Data Governance in the Digital Age. *https://www.cigionline.org/static/documents/documents/Data%20Series%20Special%20Reportweb.pdf* (dostop 5.5.2022).

[5] Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinović, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M. & Zupan, B. (2013). Orange: data mining toolbox in Python, *Journal of Machine Learning Research, 14*, 2349–2353.

[6] Devlin, J., Chang, M.W., Lee, K. & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*. arXiv:1810.04805.

[7] Eggers, W.D., Gracie, M., Malik, N. idr. (2018). Using AI to unleash the power of unstructured government data. *Center for Government Insights, Deloitte Service LP*. https://www2.deloitte.com/content/dam/Deloitte/lu/Documents/public-sector/lu-ai-unstructured-government-data.pdf.

[8] European Comission. (2020). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. A European Strategy for Data (COM(2020) 66 final), 19. februar 2020, str. 22-23, https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A52012DC0673.

[9] 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.

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

[11] Hollibaugh, Jr., G.E. (2019). The Use of Text as Data Methods in Public Administration: A Review and an Application to Agency Priorities. *Journal of Public Administration Research and Theory, 29*(3): 474–490. https://doi.org/10.1093/jopart/muy045.

[12] Jain, S., Seeja, K.R. & Jindal, R. (2020). A New Methodology for Computing Semantic Relatedness: Modified Latent Semantic Analysis by Fuzzy Formal Concept Analysis. *Procedia Computer Science, 167*: 1102-1109. https://doi.org/10.1016/j.procs.2020.03.412.

[13] Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation 28(1), 11–21 (1972)*Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H. & Mikolov, T. (2016). Fasttext. zip: Compressing text classification models. *arXiv preprint.* arXiv:1612.03651.

[14] 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.

[15] Liu, Y., Tang, W., Liu, Z., Ding, L. & Tang, A. (2022). High-quality domain expert finding method in CQA based on

multi-granularity semantic analysis and interest drift. *Information Sciences, 596*: 395–413. https://doi.org/10.1016/j.ins.2022.02.039.

[16] Marina Micheli, Marisa Ponti, Max Craglia and Anna Berti Suman (2020). Emerging models of data governance in the age of datafication. *Big Data & Society, July–December, p. 1–15,* https://ec.europa.eu/jrc/communities/sites/default/files/2053951720948087.pdf (dostop 5.5.2022).

[17] Isabel Matranga (2021). Five questions to ... https://www.eng.it/en/interviews/5-domande-a-Isabel-Matranga (dostop 5.5.2022).

[18] Lars Nagel (2021). The Magic of Data Spaces Now. https://internationaldataspaces.org/the-magic-of-data-spaces-now/ (dostop 5.5.2022).

[19] OPEN DEI Task Force 1 (2021). Design Principles for Data Spaces. *International Data Spaces Association,* https://design-principles-for-data-spaces.org/ (dostop 5.5.2022).

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

[21] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J.. Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res., 12*: 2825–2830.

[22] Peppes, N., Alexakis, T., Adamopoulou, E., Remoundou, K. & Demestichas, K. (2020). A semantic engine and an ontology visualization tool for advanced crime analysis. *Procedia Computer Science, 176*: 1829–1838.

[23] Provost, F. & Fawcett, F.T. (2013). Data Science and its Relationship to Big Data. *Big Data, 1*(1).

[24] Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text mining: applications and theory, 1, 1-20.*

[25] Sacha, D., Sedlmair, M., Zhang, L., Lee, J. A., Peltonen, J., Weiskopf, D., North, S. C. & Keim, D.A. (2017). What you see is what you can change: human-centered machine learning by interactive visualization. *Neurocomputing, 268*: 164–175.

[26] Stavrianou, A., Andritsos, P. & Nicoloyannis, N. (2007). Overview and Semantic Issues of Text Mining. *SIGMOD Record, 36*(3). https://sigmodrecord.org/publications/sigmodRecord/0709/p23.cesar-andritsos.pdf (dostop 5.5.2022).

Miha Jesenko je diplomiral leta 2004 na Pravni fakulteti Univerze v Ljubljani in leta 2008 magistriral na Pravni Fakulteti Univerze v Stockholmu s področja pravo in informacijska tehnologija. Večino poklicne poti je posvetil delu s podatki, zlasti informacijskimi sistemi s pravnimi in poslovnimi informacijami. Od skrbništva raznih podatkovnih baz do skrbništva evropskih pravnih vsebin. Opravljal je tudi naloge odgovornega urednika revije Pravna praksa. Trenutno je zaposlen kot podsekretar na Ministrstvu za javno upravo, Direktorat za informatiko, Urad za razvoj informacijskih rešitev, Sektor za upravljanje podatkov.

▪

Miro Lozej je diplomiral na Fakulteti za matematiko in fiziko Univerze v Ljubljani in se je udejstvoval v veči dejavnostih. Od pregleda metod sistemske analize, predavanj na tedaj Višji pomorski in prometni šoli, programiranja, vodenja razvojnih projektov do vodje Službe za informacijsko tehnologijo Ministrstva za kmetijstvo, gozdarstvo in prehrano in službe na Ministrstvo za javno upravo je bilo računalništvo skupni imenovalec njegovih dejavnosti. Kot diplomant na matematiki je imel pri tem dobro podlago. Vendar pa ga je razmišljanje o vplivih tehnoloških novostih na kakovost našega življenja vodila od navdušenega uporabnika prvih osebnih računalnikov k resno zadržanemu skeptiku.

▪

Karmen Kern Pipan je diplomirala leta 1998 na Univerzi v Mariboru, Fakulteti za organizacijske vede na področju informatike, kjer je leta 2001 tudi magistrirala in leta 2010 doktorirala na področju managementa kakovosti. Ima bogate izkušnje iz kakovosti, strateškega načrtovanja, razvoja informacijskih rešitev ter upravljanja podatkov. V svoji karieri je vodila sektor za kakovost in poslovno odličnost na Uradu RS za meroslovje, delovala kot visokošolska predavateljica in vodila medresorsko skupino za pripravo Strategije razvoja javne uprave 2020. Deset let je delovala kot mednarodna ocenjevalka EFQM v Bruslju. Zadnja leta je zaposlena na Ministrstvu za javno upravo, Direktoratu za informatiko, Uradu za razvoj informacijskih rešitev kot vodja Sektorja za upravljanje podatkov.«

▪

Primož Godec je asistent in raziskovalec v Laboratoriju za bioinformatiko na Univerzi v Ljubljani, Fakulteti za računalništvo in informatiko. Po izobrazbi je magister računalništva in informatike. Aktiven je na področju interaktivne analize podatkov, predvsem se osredotoča na analizo slik in besedil. Aktiven je tudi pri razvoju metod za odprtokodno orodje Orange.

Vesna Tanko je leta 2009 diplomirala na Fakulteti za računalništvo in informatiko (smer informatika), in sicer po starem, pet letnem univerzitetnem programu. Že v času študija je sodelovala pri razvoju informacijskih sistemov v podjetju IxtlanTeam d.o.o. Sodelovala je pri razvoju aplikacij za Davčno upravo RS. Pri razvoju je uporabljala programska jezika PowerBuilder in .NET ter Oracle podatkovno bazo. Leta 2015 se je zaposlila v laboratoriju za Bioinformatiko na Fakulteti za računalništvo in informatiko, kjer sodeluje pri razvoju orodja Orange. Poleg tega je zaposlena še pri podjetju Revelo d. o. o., kjer se ukvarja z informacijskimi rešitvami in podatkovno analitiko. Pri podjetju deluje kot samostojna razvijalka in podatkovna znanstvenica.

▪

Lan Žagar je asistent na Fakulteti za računalništvo in informatiko ter član Laboratorija za bioinformatiko. Raziskovalno se ukvarja z učenjem rangiranja in povezavo slednjega s hkratnim učenjem več nalog.

▪

Ajda Pretnar Žagar je raziskovalka v laboratoriju za bioinformatiko na Fakulteti za računalništvo in informatiko Univerze v Ljubljani ter na Inštitutu za novejšo zgodovino. Ukvarja se metodologijo interdisciplinarnih in multidisciplinarnih raziskav ter uporabo strojnega učenja in podatkovnega rudarjenja v družboslovju in humanistiki.

▪

Nikola Đukić je študent magistrskega programa Podatkovne vede na Fakulteti za računalništvo in informatiko Univerze v Ljubljani. Raziskovalno deluje predvsem na področjih računalniškega vida in obdelave naravnega jezika.

▪

Blaž Zupan raziskuje in poučuje umetno inteligenco in strojno učenje na Univerzi v Ljubljani in na Baylor College of Medicine v Houstonu. Na Fakulteti za računalništvo in informatiko (FRI) v Ljubljani vodi laboratorij za bioinformatiko, ki med drugim razvija svetovno znano programsko orodje za strojno učenje Orange (https://orange.biolab.si). Svoja dela je objavil v več kot sto člankih, ki so skupaj prejeli več kot deset tisoč citatov. Je prejemnik Zoisovega priznanja (2010), dveh Zlatih plaket Univerze v Ljubljani (2011, 2019), Fulbrightove štipendije (2013) in šestkratni prejemnik naziva naj-učitelj, ki ga podeljujejo študenti FRI (2008-2017). Po izboru častnika Financial Times in podjetja Google je bil uvrščen na seznam sto najvplivnejših inovatorjev srednje in vzhodne Evrope (2016).