Interaktivno eksperimentiranje z besednimi vložitvami v platformi ClowdFlows

Interactive Experimentation with Word Embeddings in the ClowdFlows platform

Martin Žnidaršič
martin.znidarsic@ijs.si

Senja Pollak
senja.pollak@ijs.si

Vid Podpečan
vid.podpecan@ijs.si

Institut "Jožef Stefan"
Jamova cesta 39
1000 Ljubljana, Slovenija

POVZETEK

V članku predstavimo spletno platformo ClowdFlows, ki je namenjena analiziranju podatkov in strojnemu učenju in omogoča uporabo interaktivnih delotokov. Posebej predstavimo značilnosti platforme, ki lajšajo njeno uporabo programiranja neveščim uporabnikom in elemente platforme, ki omogočajo analizo teksta z najsodobnejšimi pristopi vektorskih vložitev. Poročamo tudi o praktičnem preizkusu uporabnosti platforme in njenih orodij z vektorskimi vložitvami za izbrane ciljne uporabnike s področij humanistike in družboslovja.

KLJUČNE BESEDE

procesiranje naravnega jezika, besedne vložitve, spletna aplikacija, delotoki

ABSTRACT

The paper presents the ClowdFlows web platform for machine learning and data analysis using interactive workflows. In particular, we highlight selected features that facilitate its use by non-programmers as well as selected elements of the platform that enable text analysis using state-of-the-art word embedding approaches. We also report on a hands-on evaluation of the usability of the platform and its word embedding components in a selected group of end users from the fields of humanities and social sciences.

KEYWORDS

natural language processing, word embeddings, web application, workflows

1 UVOD

Področja, povezana z metodami umetne inteligence, kot so rudarjenje podatkov, strojno učenje in avtomatska obdelava naravnega jezika, v zadnjih letih doživljajo razmah v praktični uporabi. Najnovejši metodološki dosežki so običajno najprej na voljo v obliki programskih knjižnic ali spletnih storitev (angl. *web services*), pozneje v platformah za razvijanje rešitev z udobnim uporabniškim vmesnikom in običajno še pozneje v namenskih orodjih, ki to metodologijo uporabljajo interno in omogočajo njeno uporabo brez ali z zelo omejenim vplivom na način delovanja tudi uporabnikom brez računalniškega predznanja. Slednjim samostojno rabo tovrstnih metod med drugim otežuje potrebno predznanje, ki je potrebno za njihovo smiselno uporabo, včasih pa tudi postopki namestitve in nastavitev programske opreme. Prototipno raziskovalno orodje ClowdFlows, ki ga razvijamo na Odseku za tehnologije znanja na Institutu "Jožef Stefan", naslavlja ti dve oviri in kaže potencial za praktično uporabo. V sklopu projekta EMBEDDIA [14, 13, 16] smo razširili nabor zmogljivosti tega orodja predvsem na področju analize naravnega jezika, zato se v tem prispevku osredotočamo na metode in končne uporabnike s tega področja. Natančneje, predstavimo primer učenja in uporabe modelov za besedne vektorske vložitve in izkušnje novih uporabnikov s področja humanistike in družboslovja.

V razdelku 2 predstavimo osnovno sorodno delo. Platforma ClowdFlows je opisana v razdelku 3. Razdelek 4 predstavi primer uporabe vektorskih vložitev in uporabniške izkušnje. Zaključki so podani v razdelku 5.

2 OZADJE IN SORODNO DELO

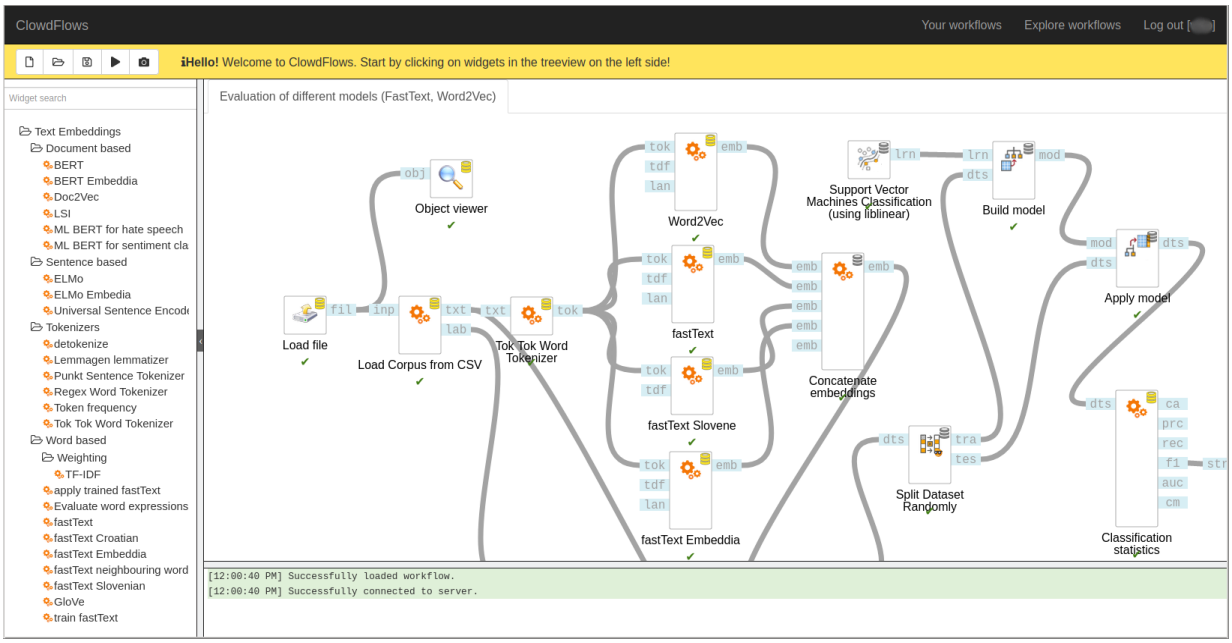
2.1 Platforme za vizualno programiranje in deljenje rešitev

Programsko orodje ClowdFlows, ki je predstavljeno in uporabljeno v tem prispevku, je podobno nekaterim drugim orodjem za upravljanje delotokov podatkovnega rudarjenja. Slovenskim uporabnikom je verjetno najbolj poznano orodje Orange [2], podobni pa sta orodji tudi Weka [18] in RapidMiner [8, 5]. Vsa ta orodja omogočajo vizualno programiranje s programskimi gradniki in upravljanje tako izdelanih programov. Manj razširjene so rešitve za skupno rabo delovnih tokov. To recimo ponuja portal myExperiment [15] ali spletna stran pobude OpenML [17]. Je pa uporabnost teh rešitev omejena predvsem na dobro podprto javno deljenje rešitev, za izvajanje ali urejanje delovnih tokov pa mora uporabnik še vedno namestiti posebno programsko opremo, v kateri so bili le-ti zasnovani. ClowdFlows, po drugi strani, omogoča tako izdelavo kot tudi deljenje in izvajanje delotokov.

2.2 Besedne vložitve

Besedne vektorske vložitve, ki so strojno naučene z uporabo nevronskih mrež, so predstavitve besed v prostoru, kjer vsako besedo opisuje vektor z veliko dimenzijami (tipično od nekaj deset do nekaj sto). Besede, ki so si blizu v vektorskem prostoru (kar lahko merimo s kosinusno razdaljo), so si tudi semantično podobne. Med vektorskimi vložitvami je mogoče računati tudi odnose, ki presegajo enostavno sorodnost besed, npr. preko analogij. Na primer, odnos *Madrid:Španija* je podoben odnosu *Pariz:Francija* [10]. Pri statičnih vložitvah, kot so modeli word2vec [9] in fastText [1], je posamezna beseda v korpusu predstavljena z enim vektorjem. Pri metodi fastText je vsaka

Information Society 2022, 10–14 October 2022, Ljubljana, Slovenia



Slika 1: Glavni pogled v ClowdFlows.

beseda predstavljena kot vsota vektorskih vložitev znakovnih n-gramov, ki jih beseda vsebuje. V praksi to pomeni, da metoda pri modeliranju semantične bližine upošteva tudi morfološko podobnost besed, zaradi česar je ta metoda še posebej uporabna za izračun besednih vložitev v morfološko bogatih jezikih, kot je slovenščina. Za razliko od statičnih vložitev pa pri kontekstualnih vložitvah, kot sta na primer modela ELMo [12] in BERT [3], vsako pojavitev besede opisuje svoj vektor. To je pomembno predvsem z vidika večpomenskih besed pa tudi v primerih, kjer analiziramo razlike med besedami v različnih kontekstih. Za veliko jezikov obstajajo prednaučeni modeli na velikih jezikovnih korpusih [4, 3], ki jih je mogoče priučiti za posamezne domene in naloge.

3 CLOWDFLOWS

ClowdFlows [6, 7] je spletna platforma za analiziranje podatkov in strojno učenje z grafičnim uporabniškim vmesnikom, ki omogoča izvajanje v brskalniku brez zahtev po lokalni namestitvi programske opreme, ponuja pa tudi preprosto javno deljenje izdelanih rešitev. Gre za odprtokodno raziskovalno orodje, katerega zadnja stabilna različica ClowdFlows 3 je na voljo na naslovu: `https://cf3.ijs.si/`.

Grafičen način sestave delovnih tokov in uporaba javno deljenih rešitev brez nameščanja dodatne programske opreme sta značilnosti, ki lajšata uporabo tudi uporabnikom, ki nimajo programerskega predznanja, imajo pa zanimive podatke in raziskovalne probleme, pri katerih bi jim prav prišle metode, ki so na voljo v ClowdFlows. Za raziskovalce je poleg tega pomembno tudi preprosto deljenje in preprostost ponavljanja ali nadgrajevanja obstoječih eksperimentov.

Elementi v ClowdFlows 3 vsebujejo vrsto programskih gradnikov, ki ponujajo delo z vektorskimi vložitvami. Vsebujejo prednaučene statične in kontekstualne modele za več jezikov kakor tudi nekaj orodij, ki na njih temeljijo, kot so na primer klasifikatorji za analizo sentimenta novic [11] in prepoznavanje sovražnega govora [11].

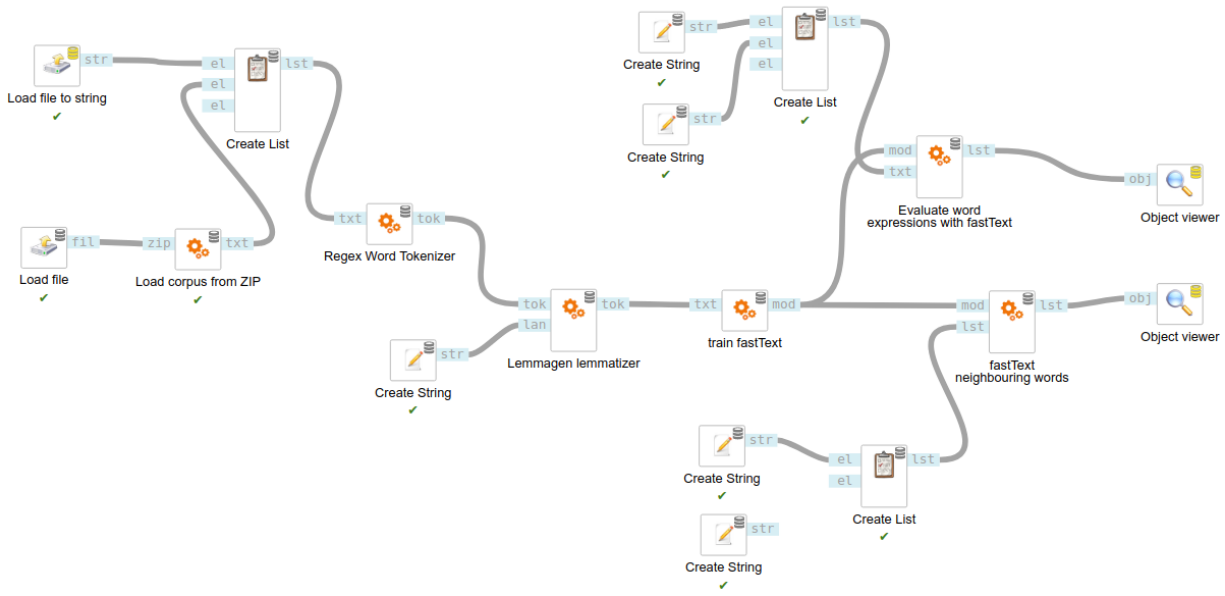
Po prijavi v ClowdFlows imamo na voljo kratek tečaj o osnovah, izdelavo novega delotoka ali pregled javno dostopnih rešitev. Glavni pogled je namenjen izdelavi, pregledu in poganjanju delotokov. Prikazan je na sliki 1. Večji del tega pogleda predstavlja delovna površina, na katero lahko potegnemo (ali uvrstimo z dvoklikom) željeni programski gradnik (angl. *widget*) iz seznama razpoložljivih gradnikov na levi strani pogleda. Smiselno povezani gradniki predstavljajo delotok, ki ga lahko poženemo z nadzornim gumbom *Play*.

Povezave med gradniki vzpostavimo s klikom na izhod enega gradnika in vhod drugega. Vhodi so predstavljeni kot svetlo modri pravokotniki na levi strani gradnika in izhodi kot tovrstni pravokotniki na desni. Povezave lahko odstranimo tako, da z desno tipko miške kliknemo povezavo in izberemo možnost *Remove*. Delotoki se shranjujejo samodejno, lahko pa jih tudi eksplicitno shranimo s pritiskom na kontrolnik za shranjevanje, kar nam omogoča tudi lastno poimenovanje shranjenega dela. Shranjene delotoke lahko pregledujemo, kopiramo, brišemo, izvažamo ali javno delimo na pogledu, ki se pokaže ob izbiri *Your workflows*. Javno objavljeni delotoki dobijo nespremenljiv URL naslov, ki ga lahko delimo in vsakemu uporabniku ClowdFlows omogoča, da ustvari svojo kopijo tako deljenega dela.

4 UPORABA VEKTORSKIH VLOŽITEV

4.1 Učenje modela vložitev v ClowdFlows

Za pridobitev predstavitve teksta v obliki vektorskih vložitev lahko uporabimo predpripravljene modele ali pa take modele sami strojno naučimo. Tovrstno strojno učenje je običajno računsko zelo zahtevno in za smiselne rezultate potrebuje velike količine podatkov. V praksi se zato pogosto uporablja predhodno naučene modele. ClowdFlows ponuja več prednaučenih modelov, naprimer ELMo, Word2Vec in razne modele pristopa BERT in fastText, tudi za slovenščino.



Slika 2: Delotok, ki je bil uporabljen na delavnici s ciljnimi uporabniki. Dostopen je na: https://cf3.ijs.si/workflow/283

.

Učenje lastnih modelov je smiselno, ko gre za posebna besedila ali naloge, pri katerih jih želimo uporabljati. Velika računska zahtevnost, velike količine podatkov in z njima povezani daljši časi obdelave namreč niso združljivi z interaktivno uporabo, ki je značilna za ClowdFlows. Pri uporabnikih digitalne humanistike in družboslovja smo zaznali potrebo za učenje vložitev na majhnih, specifičnih korpusih, kot so pesniške zbirke, specializirani novičarski članki ipd. Takšni korpusi so pogosto bistveno manjši od tipičnih korpusov, ki se uporabljajo za učenje vektorskih vložitvev. Glede na potrebe uporabnikov in zmogljivosti platforme smo se odločili za implementacijo gradnika za učenje modelov *train fastText*, saj je algoritem *fastText* eden najučinkovitejših in najmanj računsko zahtevnih. Implementacija gradnika v ClowdFlows vsebuje tudi namige, kako prilagoditi privzete parametre za učenje na majhnih korpusih. Za sprejemljivo hitro interaktivno delo vseeno priporočamo, da vhodni korpus ne presega dveh milijonov besed ali približno 10 MB neobdelanega besedila.

Gradnik *train fastText* z uporabo algoritma fastText nauči nov vektorski model na vhodnem korpusu. Tak model lahko nato posredujemo drugim gradnikom. Vhod v *train fastText* je besedilni korpus, kot je na primer izhod gradnika *Load Corpus from CSV*. Korpus je mogoče tokenizirati, lematizirati ali pa uporabiti tudi brez tovrstne predobdelave.

train fastText uporabniku ponuja nastavljanje sledečih parametrov:

bucket - število skupin (značilke besednih in znakovnih *n*-gramov so zgoščene v fiksno število skupin);
epoch - število epoh učenja;
lr - hitrost učenja;
dimension - velikost besednih vektorjev;
window - velikost kontekstnega okna;
model - vrsta nenadzorovanega fastText modela (cbow ali skipgram) ter
min_count - najmanjše število pojavitev besede, pri katerem se beseda še upošteva.

Kjer je primerno, opis parametra vključuje namig, ali je v primeru majhnih učnih podatkov priporočljivo povečati oz. zmanjšati vrednost parametra.

4.2 Izkušnje uporabnikov

Uporabnost platforme ClowdFlows in najpomembnejših komponent za analizo naravnega jezika z vidika ciljnih končnih uporabnikov smo preverjali v okviru enodnevne delavnice, ki je potekala (na daljavo) 27. januarja 2022. Delavnica je bila namenjena eni od naših primarnih ciljnih skupin: raziskovalcem z različnih področij humanistike in družboslovja, ki (predvidoma) niso vešči programiranja.

Za potrebe delavnice smo pripravili primer delotoka za analiziranje besedil z vektorskimi vložitvami. Prikazan je na sliki 2. Delotok se začne z dvema možnima načinoma vnosa vhodnih podatkov, nadaljuje z opcijsko uporabo tokenizatorja in lematizatorja (ta kot vhodni podatek sprejema tudi oznako jezika), čemur sledi učenje modela fastText. Naučeni model nato v delotoku uporabimo na dva načina: v gradniku *fastText neighboring words* pregledujemo okolico (sosednje besede) izbranih besed, v gradniku *Evaluate word expressions with fastText* pa na modelu preizkušamo uporabo izrazov (seštevanje, odštevanje) na vektorskih predstavitvah besed. Ogled rezultatov v obeh primerih omogočimo z gradnikom *Object viewer*.

Delavnica se je začela s skupno uvodno predstavitvijo platforme ClowdFlows in primera delotoka s slike 2, ki je trajala 20 minut in v kateri smo izbrane primere prikazali z uporabo besedila novele *Deseti brat* Josipa Jurčiča.

Temu je sledilo osem 20-minutnih sej, v katerih je vsak uporabnik ustvaril svoj primerek delotoka, naložil svoj korpus in preizkusil izbrane komponente ClowdFlows. Ena seja je bila namenjena enemu uporabniku in njegovim podatkom, drugi uporabniki pa

so lahko prisostvovali kot opazovalci. Uporabnikom smo pri njihovem delu pomagali, če so imeli težave pri uporabi platforme ali pri pripravi svojih vhodnih podatkov. Udeležba na delavnici je bila na povabilo. Udeleženci, ki so bili povabljeni na delavnico, so raziskovalci s področij literarnih ved, sociologije, socialnega dela in sorodnih področij. Pripravili so lastne korpuse s svojih področij, kot so na primer tematski korpusi migracij, korpus del slovenskih literatov, korpus francoske poezije, LGBT, novice, ki govorijo o socialnem delu in podobno. Nekateri udeleženci so bili vabljeni v okviru interdisciplinarnih projektov SOVRAG in CANDAS. Nihče od udeležencev pa ni imel predhodnih izkušenj s ClowdFlows. Zaradi velikega zanimanja smo število sej povečali s predvidenih 8 na 10.

Uporabljeni korpusi so bili zelo raznoliki, udeležence pa so zanimali različni vidiki obdelave besedil. V večini primerov so bili začetnemu delotoku dodani dodatni gradniki, da bi rešili določeno težavo ali zadovoljili posebne interese. Udeleženci so na primer iskali podobnosti in razlike v sosedstvu besed na podlagi korpusov iz različnih obdobij ali od različnih avtorjev. Zanimale so jih tudi osnovne značilnosti takih korpusov, kot so recimo najpogosteje uporabljene besede, s čimer so bile povezane tudi druge osnovne operacije, kot je na primer filtriranje besed. Med delavnico sta bili odkriti dve specifični tehnični težavi: (I) napake so se pojavile v primeru vnosa besedila s posebnimi znaki, ki ni bilo kodirano v kodni tabeli UTF, in (II) nekateri gradniki, ki vsebujejo klice na spletne storitve, so poročali o preseženi časovni omejitvi.

Za udeležence je bil pripravljen anonimen vprašalnik, povezava do vprašalnika pa je bila posredovana po delavnici. Večini udeležencev se je prikazani potek dela zdel zelo uporaben. Velika večina (80%) še nikoli ni poskusila uporabljati vektorskih vložitev. O uporabniškem vmesniku ClowdFlows so večinoma poročali kot o preprostem za uporabo (preprost: 60%, zelo preprost: 30%), le enemu udeležencu pa se je zdel zapleten. Te rezultate je sicer treba upoštevati v kontekstu dejstva, da so odzivi zbrani kmalu po uporabi ClowdFlows, pri čemer je bila na voljo pomoč. Brez uvoda in pomoči odgovori morda ne bi bili tako pozitivni, vendar tega še nismo preizkusili. Večina udeležencev je menila, da bi ponovno uporabili ClowdFlows, če bi jim zagotovili vnaprej pripravljen delotok za njihov problem.

5 ZAKLJUČEK

Predstavili smo spletno platformo ClowdFlows, izbrane elemente, ki omogočajo napredno uporabo pristopov za učenje besednih vektorskih vložitev, in izkušnje nekaterih od naših ciljnih uporabnikov teh orodij.

Eden od ciljev platforme ClowdFlows je približanje uporabe najnovejših metod analize podatkov in strojnega učenja tudi uporabnikom, ki niso vešči programiranja. Izkušnje naše delavnice z nekaterimi od potencialnih uporabnikov so pokazale, da je to vsekakor smiselno, saj so na podlagi predpripravljenih delotokov uporabniki (sicer strokovnjaki na drugih področjih) lahko opravili analize na lastnih podatkih in tudi že iskali in predlagali nadaljnje postopke analiz, ki so smiselni in uporabni pri njihovem delu. Poleg metodoloških razširitev in tehničnih izboljšav platforme bomo zato v bodoče več pozornosti namenjali tudi razvoju primerov rešitev za ciljne uporabnike, v prvi vrsti za raziskovalce s področij, ki niso povezana z računalništvom.

ZAHVALA

Prispevek je rezultat raziskovalnih projektov *Računalniško podprta večjezična analiza novičarskega diskurza s kontekstualnimi besednimi vložitvami* (št. J6-2581), *Sovražni govor v sodobnih konceptualizacijah nacionalizma, rasizma, spola in migracij* (št. J5-3102) in programa *Tehnologije znanja* (št. P2-0103), ki jih je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna, ter evropskega projekta EMBEDDIA (št. 825153), ki ga v okviru okvirnega programa za raziskave in inovacije Obzorje 2020 financira EU.

LITERATURA

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin in Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135–146.

[2] Janez Demšar in sod. 2013. Orange: data mining toolbox in python. *Journal of Machine Learning Research*, 14, 1, 2349–2353.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee in Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. V *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

[4] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin in Tomas Mikolov. 2018. Learning word vectors for 157 languages. V *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

[5] Markus Hofmann in Ralf Klinkenberg. 2016. *RapidMiner: Data mining use cases and business analytics applications*. CRC Press.

[6] Janez Kranjc. 2017. *Web Workflows for Data Mining in the Cloud*. Doktorska disertacija. Jožef Stefan International Postgraduate School.

[7] Janez Kranjc, Vid Podpečan in Nada Lavrač. 2012. ClowdFlows: a cloud based scientific workflow platform. V *Machine Learning and Knowledge Discovery in Databases*. Lecture Notes in Computer Science. Zv. 7524. Peter A. Flach, Tijl Bie in Nello Cristianini, uredniki. Springer Berlin Heidelberg, 816–819.

[8] Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz in Timm Euler. 2006. YALE: rapid prototyping for complex data mining tasks. V *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 935–940.

[9] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado in Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. V *Advances in Neural Information Processing Systems*, 3111–3119.

[10] Tomas Mikolov, Wen-tau Yih in Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. V *Proceedings of the 2013 Conference of the North American Chapter of the ACL: Human Language Technologies*. ACL, 746–751.

[11] Andraž Pelicon, Ravi Shekhar, Matej Martinc, Blaž Škrlj, Matthew Purver in Senja Pollak. 2021. Zero-shot cross-lingual content filtering: offensive language and hate speech detection. V *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. Association for Computational Linguistics, Online, (apr. 2021), 30–34. https://aclanthology.org/2021.hackashop-1.5.

[12] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee in Luke Zettlemoyer. 2018. Deep contextualized word representations. V *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237.

[13] Senja Pollak in Andraž Pelicon. 2022. EMBEDDIA project: cross-lingual embeddings for less- represented languages in European news media. V *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation, Ghent, Belgium, (jun. 2022), 293–294.

[14] Senja Pollak in sod. 2021. EMBEDDIA tools, datasets and challenges: resources and hackathon contributions. V *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. Association for Computational Linguistics, Online, (apr. 2021), 99–109. https://aclanthology.org/2021.hackashop-1.14.

[15] David De Roure, Carole Goble in Robert Stevens. 2009. The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows. *Future Generation Computer Systems*, 25, (feb. 2009), 561–567.

[16] Matej Ulčar, Aleš Žagar, Carlos S. Armendariz, Andraž Repar, Senja Pollak, Matthew Purver in Marko Robnik-Šikonja. 2021. Evaluation of contextual embeddings on less-resourced languages. (2021). https://arxiv.org/abs/2107.10614.

[17] Joaquin Vanschoren, Jan N Van Rijn, Bernd Bischl in Luis Torgo. 2014. OpenML: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15, 2, 49–60.

[18] Ian H Witten in Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.