***Pronalaženje sličnih presuda upotrebom metoda procesiranja prirodnog jezika***

Jelena Vlajkov, Ilija Brdar

Fakultet tehničkih nauka, Univerzitet u Novom Sadu

{[jelena.vlajkov, brdar.ilija}@uns.ac.rs](mailto:jelena.vlajkov@uns.ac.rs)

***Apstrakt* – U ovom radu implementirano je nekoliko tehnika procesiranja prirodnog jezika u cilju rešavanja problema pronalaženja sličnih presuda. U upotrebljene tehnike spadaju TF-IDF, kao i pretrenirani modeli zasnovani na neuronskim mrežama Word2Vec i GloVe. Takođe je isprobana kombinacija pretreniranih modela sa NER, POS i IDF vrednostima, kako bi se istakla semantika reči u presudama. Evaluacija je vršena poređenjem izlaza iz modela i većinskim glasanjem za slične presude. Korektnost izglasanih presuda je nemoguće utvrditi bez njihovog čitanja i asistencije domenskog eksperta.**

1. UVOD

Anglosaksonsko pravo (eng. *common law*) ili precedentno pravo jeste pravni sistem koji preovlađuje u brojnim državama na engleskom govornom području, gde spadaju i SAD, Australija, UK. Pored zakona, ovaj pravni sistem zasnovan je i na presudama izrečenim u prošlosti (precedentima). Ukoliko je sličan spor rešen u prošlosti, sud je dužan da sledi obrazloženje iz prethodne presude. Ukoliko se proceni da je slučaj različit od prethodnih, sudije imaju ovlašćenje i dužnost da reše pitanje spora [1].

Pretraga korpusa presuda u cilju nalaženja sličnih može predstavljati kompleksan i vremenski zahtevan postupak. Jedan model automatizacije postupka pretrage presuda je pretraživač (eng. *search engine*). Pretraživač kao ulaz prihvata korisnički upit (eng. *query*), a kao rezultat pretrage vraća presude koje odgovaraju postavljenom upitu. U slučaju primene pretraživača, korisnici moraju biti upoznati sa upitnim jezikom (eng. *query language*), što može predstavljati dodatno opterećenje. U cilju dostizanja veće efikasnoti pretrage, sam upit se može apstrahovati, tako da predstavlja ceo dokument. U ovom slučaju, pretraživač na ulazu dobija celu presudu, a kao rezultat vraća njoj slične presude iz celog korpusa presuda. Ovakav model pretraživača biće implementiran u ovom radu. Za implementaciju biće korišćene tehnike računarske inteligencije, konkretno one iz oblasti procesiranja prirodnog jezika (eng. *Natural Language Processing*, NLP).

Metode procesiranja prirodnog jezika imaju za cilj izvlačenje svojstava (eng. *features*) iz teksta koji je napisan u slobodnoj formi. Na ovaj način, svaki tekst se može predstaviti kao jedinstven vektor karakteristika. Nakon što postoje *fature* vektori, možemo ih uporediti u cilju kvantifikacije njihove sličnosti. U ovom radu, sličnost se posmatra u kontekstu samog sadržaja presude. Kao tehnika za kvantifikaciju sličnosti, odabrano je kosinusno rastojanje.

Skup podataka na kojim se izvode eksperimenti preuzet je sa *Court Listener* veb sajta[[1]](#footnote-1) i čine ga prvostepene presude. Eksperimenti podrazumevaju primenu tehnika TF-IDF, Word2Vec, GloVe, TextRank, kao i njihovo kombinovanje.

Evaluacija performansi u kontekstu problema koji ovaj rad rešava dosta je otežana s obzirom na nedostatak anotiranih podataka i variabilnosti značenja pojma *sličnost*. Zbog toga ne postoji klasičan vid evaluacije, već će ona izvodi poređenjem izlaza različitih NLP tehnika.

Drugo poglavlje ovog rada bavi se pregledom dosadašnje literature na temu sličnih problema. Treće poglavlje opisuje metodologiju izvođenja eksperimenata. Četvrto poglavlje izlaže dobijene rezultate, dok peto iznosi opšte zaključke ovog rada.

1. SLIČNA ISTRAŽIVANJA

U ovom poglavlju biće opisane metodologije i tehnike korišćene u dosadašnjim radovima koji su se bavili rešavanjem sličnih problema. Istaknute metode možemo podeliti u dve grupe: metode zasnovane na grafu (eng. *network-based*) i metode zasnovane na tekstu (eng. *text-based*) [2].

Metode zasnovane na grafu u osnovi imaju graf gde čvorovi predstavljaju dokumente, a grane označavaju citate. Ukoliko dokument *A* citira dokument *B*, u grafu će postojati grana sa početnim čvorom *A*, i terminalnim čvorom *B*. Razlog citiranja ne igra nikakvu ulogu. Prema tome, graf će biti isti ukoliko je presuda *B* precendent presude *A*, ili ukoliko se presuda *B* navodi kao suprotnost presudi *A*. Kumar i ostali [3] definišu dva načina za pronalaženje sličnih dokumenata upotrebom grafa: bibliografsko sparivanje (eng. *bibliographic coupling*) i kocitiranje (eng. *co-citation*). Bibliografsko sparivanje meri sličnost dva dokumenta po broju dokumenata istovremeno citiranih iz oba dokumenta koji su predmet poređenja. Suprotno, kocitiranje meri broj dokumenata koji istovremeno citiraju oba dokumenta koja su predmet poređenja.

Iako grafovske metode daju potvrđeno dobre rezultate, oni najviše zavise od strukture samog grafa [2]. Prilikom izrade pravnih dokumenata, često se ne navedu svi prethodni dokumenti relevantni za posmatrani. Zbog toga, graf citiranja sadrži više od 80% izolovanih čvorova [2]. Ovo se negativno može odraziti na performanse, pa se u ovakvim situacijama češće koriste metode zasnovane na tekstu.

Kumar i ostali [3] predlažu korišćenje TF-IDF mere i kosinusne sličnosti, kao primere metoda zanovanih na tekstu. Naime, dokumenti se reprezentuju kao vektori veličine jednakoj broju jedinstvenih reči u rečniku (celom korpusu). Svaka koordinata vektora sadrži TF-IDF meru korespodentne reči iz rečnika. Kada su formirani vektori za svaki dokument, odredi se kosinusno rastojanje, kako bi se zaključilo koji dokumenti su slični.

Sugathadasa i ostali [4] koriste *TextRank* algoritam kako bi izdvojili najbitnije rečenice. Nad najbitnijim rečenicama se dalje izvode tehnike zasnovane na tekstu, računaju TF-IDF mere i kosinusna sličnosti. Deo ovog eksperimenta je preuzet i u ovom radu, te će u Sekciji 3 biti detaljnije opisan.

Almuslim i Inkpen [1] koriste metode zasnovane na neuronskim mrežama za vektorsku reprezentaciju teksta. Tu spadaju GloVe i Doc2Vec (modifikacija Word2Vec metode). Za kvantifikaciju sličnosti koriste kosinusno rastojanje između vektora. Dobijeni rezultati prikazuju da GloVe i Doc2Vec daju dosta lošije performanse, od standardne TF-IDF metode. To su pripisali nedostatku dovoljne količine podataka potrebnih za trening. Takođe, korišćenje pretreniranih modela se takođe nije pokazalo dobro, s obzirom da je pravni domen specifičan, i da se na njemu slabo mogu primeniti opšti modeli.

Thenmozhi i ostali [5] koriste TF-IDF i Word2Vec tehnike ali primenjene na posebno odabrane reči. U tekstu dokumenta izdvajaju koncepte i relacije. Koncepti predstavljaju imenice, a relacije glagole. Za njihovo izdvajanje koristili su POS (*Part of Speech*) tehniku. Metod koji podrazumeva primenu TF-IDF tehnike na relacije i koncepte daje najbolje rezultate, dok Word2Vec daje najslabije. Loše rezultate takođe pripisuju nedostatku trening podataka i opštosti modela i navode da bi Doc2Vec tehnika potencijalno mogla da unapredi rezultate.

1. METODOLOGIJA

U ovom radu, rešavanje problema pronalaženja sličnih presuda sastoji se iz nekoliko eksperimenata. Prvo se skup preuzetih presuda podelio na trening i test skup. Potom je vršeno niz eksperimenata. Prvi eksperiment predstavlja pronalaženje najbitnijih rečenica u dokumentu korišćenjem *TextRank* algoritma [7], a zatim predstava tih rečenica kao vektor karakteristika (eng. *feature vector*) putem TF-IDF mere. Potom drugi eksperimenti se zasnivaju na predstavi dokumenta kao vektor karakteristika korišćenjem Word2Vec i GloVe metoda. Takođe u ovom eksperimentu su korišćene i IDF (*Inverse Document Frequency*), POS (*Part Of Speech*) i NER (*Named Entity Recognition)* tehnike kako bi se odredila važnost reči u rečenici. Implementacija eksperimenata izvršena je u programskom jeziku Python.

1. *Prikupljanje podataka*

Kao skup podataka korišćene su američke prvostepene presude sa *Court Listener* veb sajta. Za trening skup izdvojeno je 1399 presuda, a za test skup 434 presude. Presude preuzete sa veb sajta dolaze u *JSON* formatu. Za svrhe eksperimenata korišćen je čist tekst presude, bez ostalih metapodataka, kako je cilj eksperimenta čisto procesiranje prirodnog jezika.

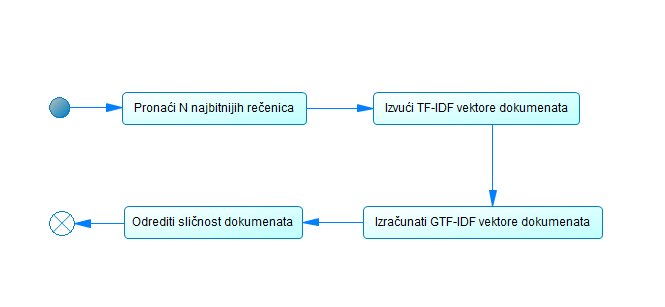
1. *TextRank i TF-IDF*

*TextRank* [7] je algoritam zasnovan na *PageRank* algoritmu. Često je korišćen za ekstrakciju ključnih reči kao i za sumarizaciju teksta. U ovom eksperimentu korišćen je *TextRank* algoritam za ekstrakciju N najvažnijih rečenica, gde u našem eksperimentu N predstavlja 50. Algoritam je zasnovan na grafovima gde svaka rečenica u dokumentu predstavlja jedan čvor u grafu, a potom se računaju sličnosti između rečenica. Kada se dobije matrica sličnosti čvorova u grafu, pokreće se *PageRank* algoritam nad njom.

Kada smo izvukli 50 najbitnijih rečenica u presudi, takav dokument je predstavljen TF-IDF vektorom. Dodatno, pored TF-IDF vektora, izračunata je GTF (*Global Term Frequency*) vektor koji predstavlja pojavu reči na nivou svih dokumenata u trening skupu. Konačni vektor karakteristika jednog dokumenta izračunat je formulom:

gde je t reč za koju se vrednost traži, D dokument u kom se računa, a n broj dokumenata u trening skupu.

Ovakav postupak izvršen je i na trening i na test skupu. Sličnost dokumenata određena je kosinusnim rastojanjem. Tok eksperimenta predstavljen je slici 1.



**Slika 1** Tok prvog eksperimenta

1. *Word2Vec i Glove*

Za drugi deo eksperimenta korišćene su metode Word2Vec i GloVe. Modeli koji su korišćeni su pretrenirani 300-dimenzioni *Google*-ov Word2Vec model i GloVe model treniran na podacima sa Wikipedie. Dodatno treniranje je izvršeno nad trening skupom, te je finalni vektor predstavljao sumu vektora svih reči, odnosno srednju vrednost svih reči u dokumentu.

Pored osnovnih vektora, sume i srednje vrednosti vektora, isprobani su i eksperimenti korišćenjem tehnika IDF (*Inverse Document Frequency*), POS (*Part Of Speech*) i NER (*Named Entity Recognition*). Ideja ovog pristupa je bila da se određeni reči više vrednuju od drugih u dokumentu. POS, NER i IDF matrice kreirane su na osnovu trening skupa.

Kod IDF pristupa, svaka reč pored vrednosti vraćene iz Word2Vec ili GloVe modela, ta vrednost se množila dodatno sa vrednošću iz IDF matrice za tu reč. IDF se računao na nivou svih dokumenata, odnosno koliko česta je neka reč bila u celom skupu za treniranje. Time možemo odrediti koliko je neka reč česta ili retka u našem skupu za treniranje.

Za POS pristup, za određene kategorije reči davali su se koeficijenti. Npr. imenice su se vrednovale koeficijentom 0,7, vlastite imenice 0,9, glagoli 0,7 itd. Time se davala prednost određenim kategorijama reči, pretpostavljajući da najveći uticaj imaju imenice, pridevi i glagoli.

Za NER pristup, slično kao i kod POS pristupa, različitim entitetima pridavali su se koeficijenti. Kako je eksperiment u domenu prava, najveći prioritet pridavao se rečima koji pripadaju entitetima osoba, lokacija, zakon, novac, umetnička dela.

Na kraju, eksperimenti koji su izvršeni, uključujući oba pristupa u predstavi vektora, tj. suma i srednja vrednost vektora reči:

1. Word2Vec
2. Word2Vec + IDF
3. Word2Vec + POS
4. Word2Vec + NER
5. Word2Vec + IDF + POS
6. Word2Vec + IDF + NER
7. Word2Vec + POS + NER
8. GloVe
9. GloVe + IDF
10. GloVe + POS
11. GloVe + NER
12. GloVe + IDF + POS
13. GloVe + IDF + NER
14. GloVe + POS + NER

Kao mera sličnosti između dokumenata koristilo se kosinusno rastojanje.

1. REZULTATI

U ovom radu predstavljeno je više eksperimenata za pronalaženje sličnih presuda. Isproban je *TextRank* algoritam uz TF-IDF predstavu vektora karakteristika, Word2Vec i GloVe model uz kombinacije IDF, POS i NER vrednosti.

Rezultate ovih eksperimenata nemoguće je tačno odrediti u ovom trenutku zbog nedostatka domenskog znanja ili stručnjaka i varijabilnosti značenja pojma *sličnosti presud*a. Zbog toga će se porediti izlazi modela i uočavati sličnost između navodno sličnih dokumenata.

Za svaku presudu iz testnog skupa uzeto je 100 najsličnijih presuda iz trening skupa. Nakon toga je izvršen presek skupova iz svakog eksperimenta, te će se smatrati da ako većina modela vrati iste presude, da su one slične. Kako su modeli koji su trenirani preko sume i srednje vrednosti vektora davali skoro pa identične rezultate, u nastavku biće analizirani podaci koji su dobijeni iz modela treniranih preko sume vrednosti vektora.

Pretrenirani modeli, Word2Vec i GloVe, rade na sličan način, presek skupova sličnih presuda vršiće se na sledeći način: ako više od 5 eksperimenata sa pretreniranim vektorima vrati da su presude slične i ako je eksperiment sa *TextRank* algoritmom takođe vratio takve rezultate, smatraće se da su te presude slične.

U Tabeli 1 predstavljen je presek sličnih presuda iz svih eksperimenata za 3 presude.

**Tabela 1** Primer preseka izlaza modela

|  |  |
| --- | --- |
| ID presude iz testnog skupa | ID presude iz trening skupa |
| 199040.json | 195731.json, 195810.json, 195914.json, 196012.json, 196162.json, 196352.json, 196395.json |
| 199046.json | 195672.json, 195724.json, 195889.json |
| 199070.json | 194977.json, 195537.json |

Tačnost ovih izlaza moguće je utvrditi samo čitanjem presuda, nakon jasno uspostavljene definicije *sličnosti presuda*. U ovom trenutku, tačna provera izlaza nije moguća.

1. ZAKLJUČAK

Ovaj rad nastao je sa zadatkom rešavanja problema efikasne pretrage prethodnih sudskih presuda koje su po sadržaju slične posmatranoj presudi. Potreba za rešavanjem opisanog problema javlja se u zemljama koje implementiraju pravni sistem gde na presudu utiču ishodni prethodnih sličnih presuda.

Prvi deo eksperimenta podrazumevao je primenu *TextRank* algoritma za dobijanje najvažnihi rečenica, a zatim primenu TF-IDF i GTF metoda na te rečenice. Na taj način dobijen je jedinstven vektor karakteristika za svaki dokument. Sledeći korak podrazumeva primenu kosinusne sličnosti, za izdvajanje najsličnijih presuda. Drugi deo eksperimenta vektor karakteristika dokumenta pronalazi upotrebom Word2Vec i GloVe modela. Kombinacijom izlaza ovih modela sa vektorima dobijenim primenom *IDF*, *POS* i *NER* tehnika, obogaćena je semantika teksta, tako što se određenim rečima daje viši prioritet. Kombinovanjem modela i navedenih tehnika, uočeno je 14 eksperimenata čije smo rezultate beležili.

Evaluacija performansi modela i primenjenih tehnika predstavlja veliki izazov s obzirom na nedostatak anotiranih podataka i domenskog (pravog) znanja. Takođe, sama definicija pojma *sličnosti presuda* može biti interpretirana na različite načine. Zbog toga je odlučeno da se evaluacija vrši poređenjem izlaza modela. Naime, prvi uslov za valjan rad modela je da se u određenom procentu poklapaju njihovi izlazi. Zatim je potrebno ručno pregledati presude označene kao slične, da bi odluka modela bila potvrđena.

Buduća angažovanja na ovom radu bi trebalo koncentrisati na problem evaluacije. Tu spada angažovanje domenskih eksperata u interpretaciji izlaza modela ili masovno anotiranje skupa podataka. Na ovaj način evaluacija bi dala pouzdanije i egzaktnije rezultate.

LITERATURA

|  |  |
| --- | --- |
| [1] | "Common law," 2 March 2022. [Online]. Available: https://en.wikipedia.org/wiki/Common\_law. |
| [2] | A. Mandal, R. Chaki, S. Saha, K. Ghosh, A. Pal i S. Ghosh, „Measuring Similarity among Legal Court Case Documents,“ *Mandal, A., Chaki, R., Saha, S., Ghosh, K., Pal, A., & Ghosh, S. (2017). Measuring Similarity among Legal Court Case Documents. Proceedings of the 10th Annual ACM India Compute Conference on ZZZ - Compute ’17. doi:10.1145/3140107.3140119 ,* 2017. |
| [3] | S. Kumar, K. Reddy, B. Reddy i M. Suri, „Finding Similar Legal Judgements under Common Law System,“ 2013. |
| [4] | K. Sugathadasa, B. Ayesha, N. de Silva, A. S. Perera, V. Jayawardana, D. Lakmal i M. Perera, „Legal Document Retrieval using Document Vector,“ *Computing Conference,* 2018. |
| [5] | I. Almuslim i D. Inkpen, „Document Level Embeddings for Identifying Similar,“ *AILA,* 2020. |
| [6] | D. Thenmozhi, K. Kannan i C. Aravindan, „A Text Similarity Approach for Precedence Retrieval,“ *FIRE,* 2017. |
| [7] | R. Mihalcea and . P. Tarau, "TextRank: Bringing Order into Texts," *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing,* p. 404–411, 2004. |

1. https://www.courtlistener.com/api/bulk-info/ [↑](#footnote-ref-1)