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

Jelena Vlajkov

Fakultet tehničkih nauka

Univerzitet u Novom Sadu

[jelena.vlajkov@uns.ac.rs](mailto:jelena.vlajkov@uns.ac.rs)

Ilija Brdar

Fakultet tehničkih nauka

Univerzitet u Novom Sadu

[brdar.ilija@uns.ac.rs](mailto:brdar.ilija@uns.ac.rs)

Abstrakt –

....

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 *feature* vektor (eng. t*ext embedding*). 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 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. PRETHODNI RADOVI

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 ukuliko se presuda *B* navodi kao suprotnost presudi *A*. Kumar i ostali [3] definišu dva načina za pronalaženje slučnih 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 trenutni. 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 [5] 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 [6] 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

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* 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:

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



1. *Word2Vec i Glove*

Za drugi deo eksperimenta korišćene su metode Word2Vec i GloVe.

1. REZULTATI
2. ZAKLJUČAK
3. 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. |

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