Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.822
DOI: https://doi.org/10.24867/11BE17Ruvceski

SOFTVERSKI SISTEM ZA OBRADU PRAVNIH DOKUMENATA ZASNOVAN NA TEHNIKAMA SEMANTIČKOG VEBA I MAŠINSKOG UČENJA

SOFTWARE SYSTEM FOR LEGAL DOCUMENT PROCESSING BASED ON SEMANTIC WEB AND MACHINE LEARNING TECHNIQUES

Stefan Ruvčeski, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj –*U ovom radu predstavljen je softverski sistem za obradu pravnih dokumenata australijskog saveznog suda. Pomenuti sistem se sastoji iz tri logičke celine: semantički veb, sumarizacija i modelovanje tema. Prva celina predstavlja populisanje ontologije i ekstrakciju relevantnih semantičkih veza unutar iste. Druga se zasniva na sumiranju obimnih dokumenata u kratak tekst koji sadrži najbitnije iz originalnog. Treća celina sve dokumente svrstava u definisani broj tema kako bi se korisniku olakšala pretraga relevantnih. Predloženi sistem zasnovan na gore navedenim tehnikama obezbeđuje podršku za semantičku pretragu, kratak opis obimnog dokumenta i prikaz svih dokumenata iste teme.*

Ključne reči: *semantički veb, ontologija, sumarizacija, modelovanje teme*

Abstract – *This paper presents a software system for processing legal documents of the Australian Federal Court. The mentioned system consists of three logical segments: semantic web, summarization, and topic modeling. The first segment represents the popularization of the ontology and the extraction of relevant semantic connections within it. The second is based on summarizing extensive documents into a short text that contains the most important from the original. The third unit classifies all documents into a defined number of topics to make it easier for the user to search for relevant ones. The proposed system based on the above techniques provides support for semantic search, a brief description of an extensive document, and a display of all documents on the same topic.*

Keywords: *semantic web, ontology, summarization, topic modeling*

1. UVOD

Ljudska potreba da pretražuje internet što efikasnije, sa što manje utrošenog vremena i napora uz veliku pomoć samih računara koji bi to obavljali umesto njih i davali im preporuke i sažetke, je osnovna ideja rada.

Sa sve većim brojem informacija na vebu, korisnici nemaju vremena da čitaju obimne tekstove da bi pronašli informacije koje su im potrebne.

Sa sve većim brojem digitalnih podataka koji su obimni, kao što su pravna dokumenta, naučni radovi, edukativni materijali i slično, dolazi do potrebe za njihovom sumarizacijom i izvlačenjem konteksta kako bi ljudima bilo lakše da ih konzumiraju. Pored potrebe da se izvuče suština celokupnog dokumenta u par rečenica, javlja se potreba i da takvi dokumenti, koji se nalaze na vebu, budu mašinski čitljivi. Takođe pomenuti podaci ne pripadaju već predodređenim temama, bilo bi od koristi ljudima kada bi prilikom konzumiranja određenih dokumenata dobili preporuku semantički sličnih.

Za današnji (klasični) veb može se reći da je veb dokumenata (eng. *web of documents*), tj. da je sadržaj povezan na nivou reprezentacije. Žargonski rečeno, današnji veb je miljama širok, ali dubok svega nekoliko inča. Semantički veb je proširenje klasičnog veba kojim bi se omogućilo razumevanje i obrada podataka od strane računara (eng. web of data). Osnovna ideja semantičkog veba jeste da se postojeće informacije, razumljive za čoveka, prošire dodatnim mašinski čitljivim informacijama.

Važnost brzog pronalaska podatka koji je potreban se ogleda i u pogledu tema. U današnje vreme potreba za preporukama od strane sistema je veća nego ikada od postojanja veba. Korisnik, nakon pronalaska podatka koji zadovoljava kriterijume, ne želi ponovo da prolazi kroz ceo proces pronalaženja sličnog podatka. Bitno je predložiti mu relevantne podatke trenutnom prikazu kako bi mu se olakšala pretraga sličnih. Ovaj deo je osnova tehnike nazvane modelovanje tema koja je obuhvaćena u ovom radu.

2. TEORIJSKE OSNOVE

Jasno je da ontologije imaju široku mogućnost primene u pravnom domenu. Jedna od mogućih primena ontologija u pravu jeste semantičko indeksiranje i pretraživanje.

Ideja ovakve primene jeste predstavljanje semantike sadržane u dokumentima kako bi se olakšalo pretraživanje.

Ontologija za predstavljanje pravnih slučajeva opisana u ovom radu implementirana je upravo sa ciljem da se koristi za semantičko pretraživanje. Još jedna od mogućih primena ontologija je semantičko integrisanje (interoperabilnost), u kom ontologija ima ulogu zajedničkog jezika prilikom razmene informacija između različitih aplikacija. Pored navedenih primena, ontologije pravnog domena moguće je koristiti i za organizaciju i strukturiranje

---

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Milan Segedinac, vanr. prof.

informacija, razumevanje domena, kao i podršku rasuđivanju i rešavanju problema.

Postoji veliki broj različitih podela sumarizacije, čiji je pregled dat u radu [1]. Zavisno od načina na koji se generiše sažetak originalnog teksta, postoje dva različita pristupa automatskoj sumarizaciji teksta: ekstraktivni i apstraktivni. Sumarizacijom zasnovanom na ekstrakciji dobija se sažetak koji se sastoji iz najbitnijih delova teksta, izdvojenih iz izvornog dokumenta [2].

Prethodno pomenuti način analogan je podvlačenju (označavanju) bitnih delova teksta u knjizi prilikom učenja.

Sumarizacijom zasnovanom na apstrakciji generišu se nove fraze i rečenice, koje nisu deo izvornog teksta, i koje sadrže najznačajnije informacije iz originalnog teksta. Ovakav pristup automatskoj sumarizaciji znatno je komplikovaniji, jer zahteva model koji pored prepoznavanja bitnih delova teksta treba da uči i gramatiku prisutnu u dokumentima, da bi bio u stanju da generiše gramatički ispravne rečenice sažetka.

Pregled različitih tehnika apstraktivne sumarizacije dat je u radu [3].

Obrada prirodnog teksta (*NLP*) predstavlja izazovno istraživanje iz računarske nauke za upravljanje informacijama i omogućavanje računarima da izvuku semantički značajne informacije iz tekstualnih dokumenata. Metode modelovanja tema su snažne inteligentne tehnike koje se široko primenjuju u obradi prirodnog jezika za otkrivanje tema i semantičkog značenja iz neuređenih dokumenata.

3. SISTEM ZA OBRADU PRAVNIH DOKUMENATA

Rezultat rada je LCR (*Legal Case Reports*) aplikacija namenjena pravnim licima. Detaljniji prikaz iste biće dat u narednom poglavlju. U ovom detaljnije će biti objašnjena implementacija segmenata koji čine rad celinom.

3.1. Implementacija segmenta semantičkog veba

Ontologije su često u osnovi sistema koji podržavaju odgovore na pitanja, izvlačenje informacija i modelovanje znanja. Koriste se za modelovanje domena znanja za koju je razvijen sistem i osnovne konceptualne strukture. Dizajn sistema zasnovanih na ontologiji obično se dodeljuje računarima kojima je pored tehničkog znanja potrebno dodatno znanje o domenu za koji se sistem razvija. Posebno izazovan domen je zakon, gde se koriste koncepti povećanja složenosti i koji su povezani jedni sa drugima [4]. U tom kontekstu, postoji potreba za definisanjem alata koji mogu da podrže i programere i krajnje korisnike u pravcu boljeg razumevanja pravnih koncepata izraženim u pravnim ontologijama, tako da se izabere informisana odluka o izboru najbolje ontologije, zavisno od cilja sistema.

Za implementaciju ontologije opisane u ovom radu korišćene su već postojeće ontologije za pravni domen*: LKIF Core i JudO*. LKIF Core (*Legal Knowledge Interchange Format*) je biblioteka ontologija razvijanih za pravni domen i deo je arhitekture sistema pravnog znanja [5]. Sastavljena je iz više modula koji opisuju skupove usko povezanih koncepata iz pravnog domena.

Opisana ontologija populisana je podacima iz 120 sudskih slučajeva. Korišćeni skup podataka sastavljen je od slučajeva australijskog saveznog suda (*Federal Court of Australia*). Podaci su javno dostupni i mogu se preuzeti sa [6] gde postoji i detaljniji opis podataka. Za potrebe populisanja ontologije, iz sirovih tekstova slučajeva ručno su ekstrahovani sledeći podaci prikazani u tabli ispod.

Tabela 1. *Izvučeni podaci radi populisanja*

case_id	Identifikator slučaja
case_name	Naziv slučaja
judgement_date	Datum presude
hearing_date	Datum saslušanja
judgement_registry	Registar
jurisdiction_court	Nadležni sud
jurisdiction_court_city	Grad
judge	Sudija
side1_counsel	Branioci prve strane
side1_solicitor	Advokati prve strane
side2_counsel	Branioci druge strane
side2_solicitor	Advokati druge strane
side1_role	Uloga prve strane
side1	Imena učesnika prve str.
side2_role	Uloga druge strane
side2	Imena učesnika druge str.
legal_rules	Akti
referenced_cases	Referencirani slučajevi
ref_names	Imena slučajeva

3.2. Implementacija segmenta sumarizacije

Kao što je već rečeno, u okviru ovog rada implementiran je sistem za automatsku sumarizaciju sudskih presuda, zasnovan na ekstrakciji i nenadgledanom mašinskom učenju.

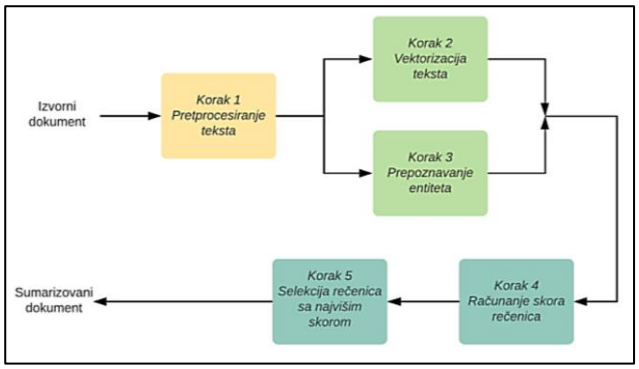
Implementacija sistema podrazumeva više koraka, koji obuhvataju: pretprocesiranje izvornog teksta, vektorizaciju teksta, identifikovanje entiteta u rečenicama, formiranje svojstava i računanje skora za svaku rečenicu i ekstrakciju rečenica sa najvišim skorom. Sve navedene funkcionalnosti objedinjene su u poseban modul za sumarizaciju teksta.

Korišćeni skup podataka sastavljen je od pravnih dokumenata australijskog saveznog suda (Federal Court of Australia), prikupljanih od 2006. do 2009. godine. Podaci se mogu preuzeti u *xml* formatu sa (Legal Case Reports, 2020), gde postoji i detaljniji opis formata podataka i mogućih pravaca istraživanja.

Slika 1. predstavlja dijagram toka aktivnosti u okviru sumarizacije teksta. Različitim bojama predstavljeni su različiti tipovi zadataka.

Žutom bojom označeno je pretprocesiranje teksta izvornog dokumenta, zelena boja predstavlja korake procesiranja, tj. ekstrakcije svojstava

za svaku rečenicu, i plavom bojom predstavljeni su završni koraci formiranja sumarizovanog teksta.



Slika 1. *Dijagram toka obrade teksta*

Skor rečenice računa se na osnovu ukupnog TF-IDF skora i identifikovanih entiteta rečenice. TF-IDF skor rečenice predstavlja sumu skorova pojedinačnih reči, normalizovan na dužinu rečenice. Korak normalizacije je važan, kako bi se izbeglo favorizovanje dugačkih rečenica. Tabela 1 prikazuje vrednosti svojstava za nasumično odabranih nekoliko uzastopnih rečenica jednog pravnog dokumenta. Atribute rečenica formiraju TF-IDF skor i brojevi svih identifikovanih entiteta od strane spaCy NER modela. Dodatno, za svaku rečenicu beleži se i njena dužina, tj. broj reči, da bi se izvršila normalizacija TF-IDF skora prilikom računanja konačnog ranga rečenice. Bitno je napomenuti da se dužina rečenice određuje nakon izbacivanja stop reči, jer samo preostale reči formiraju TF-IDF skor rečenice.

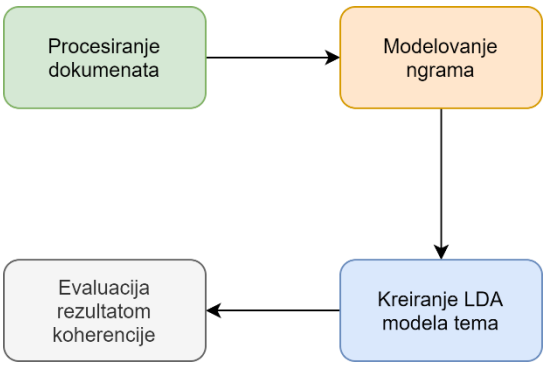
Tabela 2. *Svojstva dela rečenica sudske presude*

Svojstva	Rečenice						
TF-IDF	4.3	3.6	3.3	3.9	4.1	2.8	3.7
Person	0	0	0	0	1	1	0
Norp	0	0	0	0	0	0	0
Fac	0	0	0	0	0	0	0
Org	0	2	1	0	0	0	0
Gpe	0	0	1	0	0	0	0
Loc	1	0	0	0	0	0	0
Product	0	0	0	0	1	0	0
Date	1	0	0	0	0	0	0
Time	0	0	1	0	0	0	0
Percent	0	0	0	0	0	0	0
Money	0	0	0	0	0	0	0
Quantity	0	0	0	0	0	0	0
Ordinal	0	0	0	0	1	0	0
Cardinal	0	0	0	0	0	0	0
Length	20	15	12	16	18	8	15

2.3. Implementacija segmenta modelovanja teme

Poslednja, od tri, celina je posvećena pronalasku tema dokumenata iz korpusa istih. Da bismo korisniku sistema omogućili vizuelnu preporuku sličnih dokumenata za konzumiranje prvenstveno ih je potrebno svrstati u određen broj tema.

U ovom radu je korišćen LDA model tema, koji je opisan u prethodnom poglavlju. Da bi se njime pronašao optimalan broj tema i dokumenti razvrstali u njih potrebno je ispratiti tok operacija prikazan na slici 2.



Slika 2. *Tok modelovanja tema*

Podrazumevani koraci ovog segmenta su preprocesiranje dokumenata, koji transformaciju dokumenata u tekstualni oblik lakši za dalji rad. Nakon transformacije, izvršava se tokenizacija nad tekstom i uklanjaju se stop reči. Time za svaki dokument se kreira lista tokena kojima je isti definisan. Kreiranje n-grama, talnije bigrama i trigrama. U oblasti računarske lingvistike i verovatnoće, n-gram je neprekidni niz od n stavki iz datog uzorka teksta ili govora.

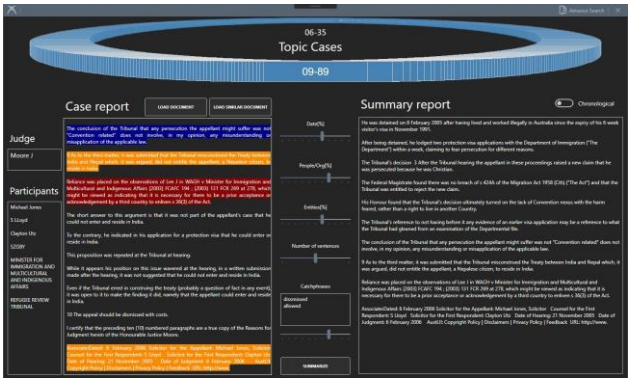
Nakon kreiranja n-grama, dolazi treći korak, modelovanje tema pomoću LDA modela. Prilikom kreiranja modela potrebno je definisati parametre, od kojih su najbitniji, korpus dokumenata nad kojim se trenira, rečnik i broj tema u koje treba podeliti dokumente. Rezultat ove faze je set podataka prikazan na slici 3.

dominant_topic	perc_contribution	topic_keywords	filename
4.0	0.4499	motion, relief, plead, trial, paragraph, actio...	06_1.xml
2.0	0.4290	contravention, penalty, charge, offence, crimi...	06_100.xml
0.0	0.4153	client, letter, email, advice, produce, record...	06_1001.xml
12.0	0.2248	agreement, contract, business, lease, clause, ...	06_1004.xml
16.0	0.3211	income, payment, trustee, assessment, bankrupt...	06_1005.xml

Slika 3. *Rezultati modelovanja tema*

4. PRIKAZ SLUČAJEVA

Pravna lica, konstantno pretražuju obimne dokumente, pokušavajući da pronađu informacije koje će im pomoći u trenutnim slučajevima. To je glavna motivacija, da korisnik lakše dođe do suštine dokumenta, sažetkom istog, da pretraži ostale dokumente iste teme i upitima kroz interfejs dođe do specifičnih informacija koje nudi aplikacija.



Slika 4. *LCR Aplikacija*

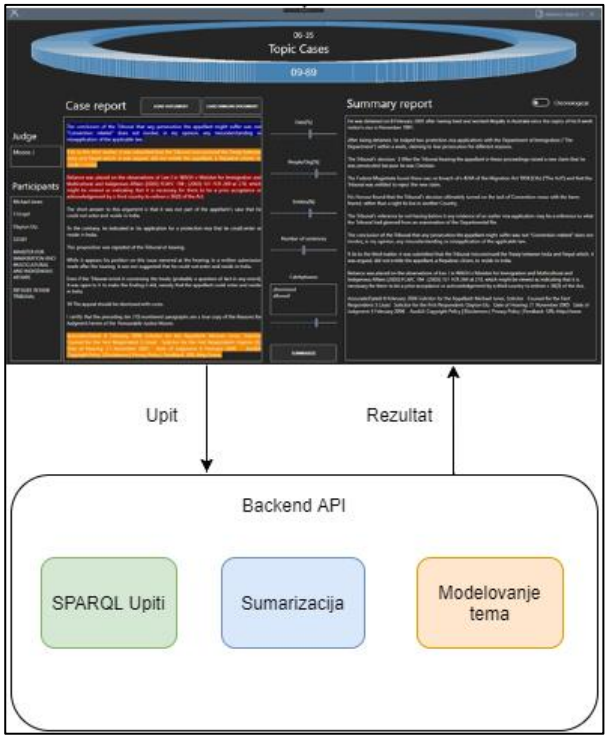
Na slici 4. je prikazan interfejs aplikacije. Mogu se uočiti tri celine koje su opisane kroz prethodna poglavlja ovog rada. Na samom vrhu korisnik aplikacije može da prođe kroz sve nazive dokumenata koji su relevantni, tačnije nalaze se u istoj temi kao trenutni dokument. Ispod relevantnih dokumenata se nalazi sekcija posvećena sumarizaciji. Sa leve strane korisnik može učitati dokument koji će mu se prikazati u levom polju, zatim u sredini može namestiti parametre entiteta, koji su mu bitniji koji ne u trenutnoj sumarizaciji.

Nakon što pritisne dugme za sumarizaciju original tekst sa leve strane će podvući najbitnije rečenice dok će se te iste rečenice, izdvojene, prikazati na desnoj strani.

Način prikaza zavisi od toga kako je korisnik izabrao da mu se rečenice prikazu, hronološki ili po bitnosti istih u tekstu. Poslednji segment interfejsa zasnovan je na ekstrakciji informacija iz ontologije putem SPARQL upita. Na levoj strani interfejsa može se videti sudija slučaja koji trenutni dokument opisuje, kao i učesnike u slučaju.

Aplikacija opisana u prethodnom pasusu je jedan deo, većeg sistema koji se sastoji od prikazane aplikacije (*frontend*) i API (*application programming interface*) dela gde se nalazi kompletna logika i implementacija svakog segmenta (*backend*).

Ove dve aplikacije komuniciraju, tačnije *frontend* aplikacija poziva akcije backend API-ja putem HTTP (*Hypertext Transfer Protocol*) protokola .Nakon što je opisan celokupan sistem, u nastavku poglavlja biće objašnjen a implementacija svakog segmenta, u vidu dijagrama slučaja. Na slici 5. se može videti ta komunikacija između dva dela sistema.



Slika 5. *Interakcija dva dela sistema*

5. ZAKLJUČAK

U ovom radu predstavljen je softverski sistem, nazvan *Legal Case Report*, za lakšu pretragu obimnih digitalnih dokumenata. Sistem nudi sumarizaciju teksta radi bržeg shvatanja konteksta celokupnog teksta. Nudi ekstrakciju određenih informacija putem SPARQL upita u pozadini. Pored navedenih funkcionalnosti, poslednja je prikaz relevantnih dokumenata, tačnije dokumenata koji pripadaju istoj temi.

Softverski sistem sastoji se iz dve glavne komponente, prototip aplikacije za pravna lica, koja realizuje interakciju sa korisnicima sistema putem interfejsa, i softverska biblioteka koja obavlja svu logiku predstavljenu u ovom radu, izloženu putem API-ja koji poziva prvi deo istog sistema putem HTTP protokola.

Jedan od mogućih pravaca budućeg rada jeste poboljšavanje dela sistema koji se bavi prepoznavanjem imenovanih entiteta. Zbog specifičnosti teksta i entiteta koji se pojavljuju u pravnim dokumentima, bilo bi dobro probati sa podešavanjem (eng. *tuning*) i dodatnim treningom postojećih NER modela u okviru spaCy biblioteke.

Budući da pravnici većinu posla za računarom provode koristeći Microsoft Word, jedan od pravaca budućeg rada mogao bi biti kreiranje *plugin*-a za Word, upotrebom implementiranog modula za automatsku sumarizaciju.

6. LITERATURA

[1] M. Ramezani, Ontology-Based Automated Text Summarization Using FarsNet.Tabriz, 2015.

[2] V. Gupta & G. Lehal, A Survey of Text Summarization, 2010.

[3] S. Gupta & S. K. Gupta, Abstractive Summarization: An Overview of the State of the Art.Expert Systems With Applications, 2018.

[4] V. Leone, L. Di Caro & S. Villata, Legal Ontologies and How to Choose Them, 2020.

[5] R. Hoekstra, J. Breuker, M. Di Bello & A. Boer, The LKIF Core Ontology of Basic Legal Concepts, 2007.

[6] Preuzeto sa: https://archive.ics.uci.edu/ml/datasets

Kratka biografija:

Stefan Ruvčeski rođen je u Novom Sadu 1996. god. Master rad na Fakultetu tehničkih nauka iz oblasti Računarstva i automatike – Semantički veb odbranio je 2020.god.

kontakt: stefanruvceski@uns.ac.rs