TakeLab

Laboratorij za analizu teksta i inženjerstvo znanja
Text Analysis and Knowledge Engineering Lab
Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
Unska 3, 10000 Zagreb, Hrvatska

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 3797

Primjena semantičkih jezgrenih funkcija u klasifikaciji teksta

Dino Radaković

Zagreb, lipanj 2014.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA
ODBOR ZA ZAVRŠNI RAD MODULA

Zagreb, 13. ožujka 2014.

ZAVRŠNI ZADATAK br. 3797

Pristupnik: **Dino Radaković**
Studij: Računarstvo
Modul: Računarska znanost

Zadatak: **Primjena semantičkih jezgrenih funkcija u klasifikaciji teksta**

Opis zadatka:

Sadržajna klasifikacija teksta jedan je od osnovnih zadatka dubinske analize teksta. Uobičajeno se u tu svrhu koriste modeli strojnog učenja temeljeni na vektorskome prikazu dokumenta kao vreće riječi. Premda jednostavan i učinkovit, takav prikaz ne modelira semantiku dokumenta na konceptualnoj razini, stoga su u literaturi predložena razna proširenja. Jedno je od takvih proširenja model temeljen na semantičkoj jezgrenoj funkciji, koji obogaćuje prikaz dokumenta znanjem izvedenom iz ontologije.
U okviru završnoga rada potrebno je proučiti osnovne postupke za klasifikaciju teksta s naglaskom na postupke strojnog učenja s jezgrenim funkcijama. Proučiti semantičke jezgrene funkcije temeljene na Wikipediji predložene u radu Wanga i Domeniconi (2008). Razraditi postupak za izgradnju semantičkih jezgrenih funkcija za dokumente na hrvatskome jeziku korištenjem hrvatske Wikipedije. Razviti odgovarajuće programsko rješenje te ga primijeniti na klasifikaciju dokumenata na hrvatskome jeziku. Provesti iscrpno vrednovanje točnosti klasifikacije na zbirkama dokumenata na hrvatskome jeziku. Radu priložiti izvorni i izvršni kod razvijenog sustava, skupove podataka i programsku dokumentaciju te citirati korištenu literaturu.

Zadatak uručen pristupniku: 10. ožujka 2014.
Rok za predaju rada: 13. lipnja 2014.

Mentor:

Doc. dr.sc. Jan Šnajder

Djelovođa:

Doc. dr.sc. Tomislav Hrkać

Predsjednik odbora za završni rad modula:

Prof. dr.sc. Siniša Srbljić

SADRŽAJ

1. Uvod

Klasifikacija teksta (engl. *text classification, TC*) postupak je pridjeljivanja oznaka tekstnim dokumentima, s ciljem raspodjele istih po prethodno definiranim kategorijama.

Zadatak strojne klasifikacije teksta je izrada računalnog sredstva koje bi, umjesto čovjeka, ulazni tekst bilo sposobno samostalno razvrstavati u pripadne kategorije. Kako je za ručno razvrstavanje teksta potrebno imati na raspolaganju stručnjake obrazovane u određenom području, što je razmjerno skupo, a sam postupak k tome i vremenski neefikasan, javila se potreba za delegiranjem takvog posla računalu.

Interes prema strojnoj klasifikaciji teksta pojavio se već 1960-ih godina, no značajno je porastao tijekom posljednjeg desetljeća prethodnog stoljeća [7], što se objašnjava popularizacijom interneta, čija je posljedica i vrlo značajan porast u količini široko dostupnih tekstnih dokumenata pohranjenih u digitalnim zapisima. Uz porast potrebe za rješavanjem problema strojne klasifikacije teksta javili su se razni pristupi spomenutom problemu. Neki od njih su pristupi temeljeni na bazama znanja i pravilima, modeliranim prema rasuđivanju stručnjaka. Nedostatak takvih sustava je relativno velik trud koji je čovjeku potrebno uložiti u ručno definiranje klasifikatora, koji vrlo brzo raste s povećanjem broja kategorija u koje se dani tekst razvrstava te samom veličinom (i semantičkom složenošću) teksta koji klasifikator treba razvrstavati. Područje takvih pristupa naziva se još i *inženjerstvo znanja* (engl. *knowledge engineering*). Drugi pristupi su usmjereni na pokušaje automatizacije koraka koji uključuje stvaranje klasifikatora, što je karakteristika podskupa umjetne inteligencije poznatog kao *strojno učenje*. Automatizacijom izrade klasifikatora postižu se značajne prednosti u odnosu na ekspertne sustave – izbacivanjem čovjeka iz samog postupka, uz ubrzanje, gubi se i potreba za izvršavanjem zahtjevnih ljudskih poslova, koje uključuje izrada klasifikatora, kao što je to npr. inženjerstvo baze znanja.

Klasični postupak koji se primjenjuje u strojnoj klasifikaciji teksta je predstavljanje tekstne jedinice kao vreće riječi (engl. *bag of words*), pri čemu se ona predstavlja vektorom, budući da postoji niz klasifikacijskih algoritama koji se temelje na vektorskom prostoru. Riječi se prethodno morfološki normaliziraju, kako bi se time ujednačili

morfološki različiti oblici iste riječi, što je naročito značajno za jezike visokog stupnja fleksije, kao što je to, na primjer, hrvatski jezik.

Model vreće riječi danas je raširen u klasifikaciji teksta zbog jednostavnosti i kvalitete rezultata koji se ostvaruju takvim pristupom. Unatoč tome, takvim modelom nije moguće postići razlučivanje sadržaja teksta na konceptualnoj razini, što se odražava u klasifikaciji teksta koji sadrži, na primjer, višeznačnice i srodne termine. To je potaknulo istraživanja i razvoj modela koji nadilaze vreću riječi uvođenjem semantičkog znanja u sam postupak klasifikacije. Jedan takav model, temeljen na Wikipediji i stroju potpornih vektora, razvili su 2008. P. Wang i C. Domeniconi te ga predstavili u radu [9], nadjenuvši mu naziv *semantička jezgrena funkcija* (engl. *semantic kernel*).

Termin semantičke jezgrene funkcije nadovezuje se na koncept jezgrene funkcije stroja potpornih vektora, što je funkcija kojom se predstavlja numerička mjera različitosti između dvije tekstne jedinice, kao što su to, na primjer, dokumenti, koja se koristi u postupku optimizacije koji rezultira klasifikatorom.

Cilj ovog rada je implementacija modela temeljenog na semantičkoj jezgrenoj funkciji i modela temeljenog na vreći riječi, primjena oba modela na klasifikaciju dokumenata na hrvatskom jeziku i usporedba dobivenih rezultata u smislu kakvoće klasifikacije ostvarene korištenjem tih modela.

U poglavlju 2 detaljnije je opisan problem klasifikacije teksta, nakon čega u poglavlju 3 slijedi opis modela temeljenog na vreći riječi, uz osnovnu definiciju stroja potpornih vektora. Potom je u poglavlju 4 opisan model semantičke jezgrene funkcije, po uzoru na izvorni opis iz [9], uz manje izmjene, nakon čega slijedi poglavlje 5, u kojemu su uspoređeni dobiveni rezultati nad korištenom ispitnom kolekcijom dokumenata, a zatim i zaključak (poglavlje 6).

2. Klasifikacija teksta

Problem klasifikacije teksta je zadatak koji zahtijeva raspoređivanje tekstnih jedinica (članaka, dokumenata, poruka) po pripadnim, prethodno određenim kategorijama. Primjerice, pravni dokumenti mogu se razvrstati po različitim kategorijama koje predstavljaju različite podvrste prava, kao što su to građansko pravo, trgovačko pravo i upravno pravo.

Pridjeljivanje kategorija tekstu može biti višestruko – jedna tekstna jedinica može biti razvrstana u više kategorija istovremeno. Na primjer, novinski članak, koji, uz Olimpijske igre, opisuje i političku klimu države na čijem se teritoriju igre održavaju, razumno je razvrstati u kategoriju *sport*, ali podjednako i u kategoriju *politika*, u obje spomenute kategorije, pa i niti u jednu od njih, ovisno o usvojenim konvencijama klasifikacije.

Sama klasifikacija teksta suštinski je u domeni područja *pretraživanja informacije* (engl. *information retrieval, IR*), koje obuhvaća aktivnosti usmjerene na izlučivanje čovjeku korisne informacije iz teksta. Informacija koja se dohvaća u slučaju klasifikacije teksta je pripadnost dane tekstne jedinice nekoj od određenih kategorija, što se postiže na temelju sadržaja samog teksta. Primjer može biti razvrstavanje dokumenata po geografskim entitetima na koje se odnose. Na primjer, može se dogoditi da je relativno visok broj pojavljivanja riječi „maslina" u tekstu indikator da je dani dokument potrebno razvrstati u kategoriju „Mediteran" – što je primjer pretraživanja informacije, koja nije nužno izravno izražena u dokumentu, na temelju samog teksta (nije nužno da se unutar dokumenta ni na kojem mjestu eksplicitno spominje Mediteran).

Formalno, klasifikacija teksta može se opisati kao pridjeljivanje Booleovih vrijednosti svakoj od uređenih dvojki $(d_i, c_j) \in \mathcal{D} \times \mathcal{C}$, gdje $\mathcal{D}$ predstavlja skup dokumenata koje je potrebno razvrstati po kategorijama, a $\mathcal{C}$ skup određenih kategorija, po kojima je dokumente potrebno razvrstati. Vrijednošću $\top$ pridijeljenom nekom uređenom paru (d_i, c_j) označava se razvrstavanje dokumenta d_i u kategoriju c_j, dok se pridjeljivanjem vrijednosti $\bot$ označava suprotna relacija (dokument d_i se tada ne smatra razvrstanim u

kategoriju c_j), gdje vrijedi:

$$1 \leq i \leq |\mathcal{D}|\,,\ i \in \mathbb{Z}$$
$$1 \leq j \leq |\mathcal{C}|\,,\ j \in \mathbb{Z}$$

Neka je Φ^* funkcija koja opisuje idealno pridjeljivanje kategorija dokumentima:

$$\Phi^* : \mathcal{D} \times \mathcal{C} \rightarrow \{\top, \bot\}$$

Tada se cilj klasifikacije teksta može izraziti kao težnja da se što preciznije (u odnosu na Φ^*) danim dokumentima pridijele određene kategorije, tj. da se idealnu funkciju Φ^* što točnije aproksimira nekom funkcijom:

$$\Phi : \mathcal{D} \times \mathcal{C} \rightarrow \{\top, \bot\}$$

Takva funkcija Φ^* poznata je još i kao *ciljna funkcija* (engl. *target function*), dok se Φ naziva *klasifikatorom* (engl. *classifier*) [7].

Skup dokumenata nad kojima se provodi klasifikacija naziva se još i *kolekcijom dokumenata* (engl. *document collection*). Uobičajen naziv, koji će se uglavnom koristiti u ovom radu je i *korpus*, po latinskom terminu *corpus* (*skup, tijelo*).

2.1. Pristupi

2.1.1. Pristupi temeljeni na pravilima

Pristupima temeljenima na pravilima smatraju se sustavi koji pohranjuju početno danu informaciju u bazu znanja, da bi ju potom manipulacijom raznim metodama (pravilima) povezali s odgovarajućom interpretacijom. Najtipičniji primjer sustava temeljenih na pravilima su sustavi kojima se, kroz primjenu raznih pravila na dane ulazne podatke, koji se nazivaju *činjenicama* (engl. *facts*), uz podatke predstavljene u *bazi znanja* (engl. *knowledge base*) emuliraju postupak donošenja odluke ljudskog eksperta – od čega potječe i naziv *ekspertni sustavi* (engl. *expert systems*). U takvim sustavima znanje je predstavljeno preko *ako-onda* (engl. *if-else*) konstrukata, koje je inženjer baze znanja prethodno preslikao iz stručnog znanja ili ulaznih podataka.

Ekspertni sustavi pojavili su se kasnih 1970-ih godina i još uvijek predstavljaju aktualno istraživačko područje unutar umjetne inteligencije.

2.1.2. Pristupi temeljeni na strojnom učenju

1980-ih godina 20. stoljeća sve aktualnijima su postajali pristupi koji su se, za razliku od pristupa temeljenih na pravilima, služili računalnim metodama za izgradnju klasifikatora. S obzirom na to da u tom slučaju računalo „uči", područje koje obuhvaća takve pristupe zove se *strojno učenje* (engl. *machine learning, ML*). Metode strojnog učenja u klasifikaciji teksta uglavnom se sastoje od primjene algoritma nad ulaznim podacima (engl. *training dataset*), nakon čega se izgradi model klasifikatora, koji se potom ispita na podacima za vrednovanje (engl. *evaluation dataset*), za koje je unaprijed poznata ciljna vrijednost, koju klasifikator treba s čim većom preciznošću odrediti. Na primjer, uz dane konačne skupove točaka koordinatnog sustava u kojemu os x predstavlja starost čovjeka (vrijeme), a os y tjelesnu visinu čovjeka u danom vremenskom trenutku, gdje pojedini skup točaka predstavlja visinu jednog čovjeka u ovisnosti o njegovoj starosti, moguće je matematičkim postupcima, kao što je to, na primjer, *polinomna regresija* (engl. *polynomial regression*) pokušati odrediti matematičku funkciju (klasifikator) koja bi se, na temelju svega nekoliko točaka (nekoliko podataka o tjelesnoj visini i starosti) mogla (s određenom preciznošću) koristiti za predviđanje trenda kretanja tjelesne visine neke osobe kroz vrijeme. Bitno je napomenuti da se ne mora raditi o osobi čiji su podaci (točke) korišteni za izgradnju takve funkcije – ideja strojnog učenja ne implicira ograničenost ulazne domene klasifikatora na podatke koji su korišteni na izgradnju istog. Naprotiv, cilj je stvoriti model koji na temelju skupa za učenje generalizira postupak klasifikacije.

U klasifikaciji teksta, osnovni pristup klasifikaciji dokumenata temelji se na leksičkoj analizi dokumenta, normalizaciji zasebnih riječi, filtriranje leksičkih jedinki (što može uključivati razne metode), nakon čega se dokumenti modeliraju kao vektori frekvencija riječi, s kojima se, nakon toga, provodi određeni algoritam izgradnje klasifikatora, koji također provodi razvrstavanje nad vektorima čiji je format jednak formatu vektora kojima je sam i izgrađen. Više riječi o takvom pristupu bit će u narednom poglavlju.

Pristupi temeljeni na pravilima dominirali su početkom 80-ih godina prošlog stoljeća, da bi ih u velikoj mjeri iz potisnuli pristupi temeljeni na metodama strojnog učenja, koji i danas obuhvaćaju najveći dio praktičnih i istraživačkih aspekata klasifikacije teksta. Moderno područje klasifikacije teksta može se stoga smatrati presjecištem strojnog učenja i pretraživanja informacija [7].

Dok se danas pristupi klasifikacije teksta metodama strojnog učenja uglavnom temelje na vektorima učestalosti termina (engl. *term frequency vectors*), isključivo na

riječima sadržanim u tekstu, postoje ideje usmjerene na nadogradnju takvih postupaka uvođenjem semantičke informacije o samom dokumentu (temeljem vanjskog izvora informacije) u postupak klasifikacije, na primjer, korištenjem vanjske baze znanja s kojom se termini prisutni u dokumentu povežu na temelju određenih značajki, kako bi se time, na temelju veće ulazne količine informacije, kojom se definira klasifikator, dobio precizniji model za klasifikaciju teksta. Neke od tih ideja su izvedene u okviru znanstvenih radova.

U radu [9] autori definiraju pojam *koncepta*, kao onog slijeda riječi (ili samo jedne riječi), koji je izravno predstavljen člankom na engleskoj Wikipediji, nakon čega se dokumenti predstavljaju vektorima koncepata te mjere sličnosti među takvim vektorima temelje na različitim mjerama sličnosti pripadnih članaka Wikipedije, kao što je to udaljenost na grafu taksonomije. Takav se pristup koristi kao nadogradnja klasičnog pristupa, tzv. metode *vreće riječi* (engl. *bag-of-words*), koji se temelji na predstavljanju dokumenta nizom indeksa i težinskih vrijednosti, koje predstavljaju učestalosti pojavljivanja riječi iz nekog rječnika unutar tog dokumenta.

U ovom radu izveden je jedan takav model, koji se temelji na uvođenju znanja izlučenog iz hrvatske Wikipedije u postupak klasifikacije dokumenata na hrvatskom jeziku te je provedena usporedba s klasičnim tf-idf (*term frequency – inverse document frequency*) modelom. Oba modela temelje se na korištenju stroja potpornih vektora za izgradnju klasifikatora.

3. Pristup temeljen na vreći riječi

Klasični pristup rješavanju problema klasifikacije teksta korištenjem metoda strojnog učenja usmjeren je na pretvorbu danih tekstnih jedinica u vektore značajki (engl. *feature vectors*), koji se potom koriste za izradu klasifikatora te za svođenje ulaznog teksta na oblik koji klasifikator potom može razvrstati. U sklopu ovog rada, implementiran je model koji se temelji na leksičkoj analizi i morfološkoj normalizaciji tekstnih dokumenata, nakon kojih slijedi izgradnja vektora značajki, koji predstavljaju dane dokumente. Nakon cijelog postupka provode se izrada i vrednovanje klasifikatora, korištenjem prethodno stvorenih vektora značajki.

3.1. Obrada ulaznih dokumenata

3.1.1. Leksička analiza

Kao što je to slučaj kod obrade programskog jezika, tako je i kod obrade prirodnog jezika prvi korak najčešće leksička analiza. Leksička analiza prirodnog jezika uključuje razgraničavanje rečenica (engl. *sentence splitting*) te pretvorbu ulaznog teksta u niz leksičkih jedinki, što se još naziva i tokenizacijom (engl. *tokenization*).

Tokenizacija je postupak koji razdvaja dani tekst na pojavnice (*tokene*), uz, ovisno o definiciji, moguće odbacivanje određenih znakova (na primjer, znakova interpunkcije) [4].

Primjerice, za ulaznu rečenicu *„Dan je dug, ali noć je još mlada."*, provedba postupka tokenizacije koji ne uključuje izbacivanje interpunkcijskih znakova rezultirat će sljedećim nizom leksičkih jedinki:

Dan	je	dug	,	ali	noć	je	još	mlada	.

Razgraničavanje rečenica postupak je koji ulazni tekst pretvara u slijed rečenica, razdvojenih na temelju graničnika definiranih jezikom. Pristupi razgraničavanju rečenica nerijetko se temelje na pravilima ovisnim o jeziku koji se obrađuje – primjerice, za

hrvatski se jezik točka najčešće može koristiti kao graničnik između rečenica, uz ograničen broj iznimki (redni brojevi, opisivanje točke kao simbola, kratice i dr.), na čemu se i temelji pristup koji izvodi alat koji je korišten u sklopu ovog rada.

Za ulazni tekst *„Godina ima četir' puta po mjeseca tri. Ljepšeg od svibnja međ' njima ni'."*, postupkom razgraničavanja dobiju se dvije rečenice:

Godina ima četir' puta po mjeseca tri.	Ljepšeg od svibnja međ' njima ni'.

Sam postupak razgraničavanja rečenica nije izravno važan za izvođenje značajki iz teksta, ali je bitan korak za tokenizaciju, kojoj prethodi.

3.1.2. Lematizacija

Uzevši u obzir visok stupanj fleksije hrvatskog jezika, kao i činjenicu da hrvatski jezik obuhvaća velik broj riječi (primjerice, korišteni alat raspoznaje broj riječi reda veličine 10^5 u odabranoj ispitnoj zbirci, sastavljenoj od 13205 dokumenata), može se zaključiti da implementacija koja bi koristila sve oblike svih vrsta riječi trenutno nije ostvariva za praktičnu uporabu. Zbog toga je potrebno provesti postupak morfološke normalizacije – svođenja više morfoloških varijanti iste riječi na normalni oblik, odnosno ujednačavanje istih za potrebe klasifikacije. Stoga je, u okviru implementacije ostvarene u sklopu ovog rada, prije uvođenja samih riječi u rječnik proveden postupak lematizacije. Lematizacija je postupak kojim se dana riječ svodi na vlastitu lemu (engl. *lemma*). Kod imenskih riječi, radi se o nominativu jednine, dok je to kod glagola infinitiv. Izjednačavanjem riječi lematizacijom postiže se precizniji opis sastava dokumenta, što je ključno za izvedbu klasifikatora.

Lematizacija, kao morfološka pretvorba riječi iz jednog oblika u drugi, može se opisati kao preslikavanje među nizovima znakova:

$$lema : \mathcal{R} \rightarrow \mathcal{L},$$

gdje $\mathcal{R}$ predstavlja riječ u izvornom obliku, a $\mathcal{L}$ morfološki normalni oblik te riječi.

Primjer: *„Donio je kući veliku ribu." „Netko ribi grize rep."*

Danim rečenicama zajedničko je to što se odnose na ribu, što bi se moglo prevesti u relevantnu značajku za neki klasifikator. Bez morfološke normalizacije, prilikom usporedbe riječi *ribe* (imenica u akuzativu) i *ribi* (imenica u dativu) dalo bi se zaključiti da se radi o dva različita termina, dok je zapravo riječ o morfološkim varijantama

jednog termina (riječi), što se najjednostavnije primjećuje po provedbi morfološke normalizacije:

$$ribu \rightarrow riba$$
$$ribi \rightarrow riba$$

Pristupi lematizaciji mogu se podijeliti na pristupe temeljene na pravilima (engl. *rule-based*) i pristupe temeljene na leksikonu (engl. *lexicon-based*). Pristupi temeljeni na pravilima najčešće se oslanjaju na sufiksne pretvorbe nad nizovima znakova, dok pristupi temeljeni na leksikonu koriste upite nad flektivnim morfološkim leksikonom.

U izvedbi ovog rada korišten je sustav za lematizaciju koji svoj postupak temelji na leksikonu, opisan u [15].

3.1.3. Zaustavne riječi

Zaustavne riječi (engl. *stop words*) su riječi koje ne pridonose klasifikaciji teksta te se stoga uklanjaju prije nego što započinje klasifikacija i sam postupak izgradnje klasifikatora. To su, u hrvatskom jeziku, veznici, usklici, čestice, brojevi i ostale riječi, čija uporaba uglavnom ne ovisi o kontekstu u kojemu su korištene u mjeri značajnoj za postupak klasifikacije. Neke od mogućih zaustavne riječi su:

a
ajme
bijahu
četrnaestero
pa
ponegdje
tad
...

Unutar obrade ulaznog teksta izvedene u sklopu ovog rada, nakon leksičke analize izbacuju se zaustavne riječi, temeljem ručno sastavljene liste, koja sadrži 2024 različitih leksema.

3.1.4. Frekvencije termina

Nakon provođenja postupka morfološke analize slijedi izračun frekvencija termina, kao karakteristike koja pridjeljuje određenu „težinu" pojedinom terminu, s ciljem kodiranja informacije o učestalosti pojavljivanja tog termina unutar dokumenta, ali i unutar

cijelog korpusa. Termin se ovdje definira kao klasa svih morfološki normaliziranih leksičkih jedinki (tokena) istog zapisa.

Osnovni pristup pridjeljivanja frekvencije pojedinom terminu koji se nalazi u tekstu zasniva se na definiciji frekvencije kao broja pojavljivanja primjeraka tog termina u tekstu. Nedostatak takvog pristupa je u tome što se svim terminima pridjeljuje jednaka važnost (težina) na razini čitavog korpusa. Stoga su predložene različite težinske sheme. Među najpopularnijima je korištenje statističke mjere poznate pod engleskim nazivom *term frequency – inverse document frequency*, skraćeno *tf-idf*.

Uvode se dva pojma: frekvencija termina (engl. *term frequency*) i frekvencija dokumenata (engl. *document frequency*).
Frekvencija termina mjera je učestalosti pojavljivanja nekog termina unutar jednog dokumenta. Oznaka za frekvenciju termina je $t_{t,d}$, gdje indeks t označava jedan termin, a indeks d jedan od dokumenata obuhvaćenih postupkom obrade.
Frekvencija dokumenata za dani termin mjera je učestalosti pojavljivanja nekog termina na razini korpusa. Oznaka za frekvenciju dokumenata je df_t, gdje indeks t predstavlja neki termin. U tablici 3.1, načelno preuzetoj iz knjige [3], opisane su neke od shema za izračun frekvencije termina, odnosno frekvencije dokumenata.

Tablica 3.1: Neke od shema za izračun frekvencije termina, odnosno frekvencije dokumenata

naziv	$\boldsymbol{tf_{t,d}}$	**naziv**	$\boldsymbol{df_t}$
n (prirodna)	$n_{t,d}$	n (prirodna)	d_t
l (logaritamska)	$1 + log(n_{t,d})$	t (inverzna)	$log\frac{N}{d_t}$
a (pojačana)	$0.5 + \frac{0.5 \times n_{t,d}}{max_t(n_{t,d})}$		

$n_{t,d}$ – broj pojavljivanja termina t u dokumentu d
d_t – broj dokumenata unutar korpusa koji sadrže termin t
N – ukupan broj dokumenata u korpusu

U klasifikaciji teksta, za izračun frekvencije dokumenata često je se koristi izraz iz retka *t* tablice 3.1, koji je nazvan *inverzna frekvencija dokumenata* (engl. *inverse document frequency*). Ideja inverzne frekvencije dokumenata izvorno potječe iz rada [8], gdje je ta mjera predložena kao jedna od heuristika za pridjeljivanje težine terminima unutar dokumenata. U domeni klasifikacije teksta, inverzna frekvencija dokumenata pokazala se kao dobra mjera za dodjelu težine svakom od pojedinih termina.

Pokušaji pronalaska formalizma koji opravdava korisnost inverzne frekvencije dokumenata često sežu u područje teorije informacije, kao što je to, na primjer, riječ u radu [6].

Već spomenuta shema težine koja se pridjeljuje nekom terminu, tf-idf, dana je izrazom:

$$tf\text{-}idf = tf_{t,d} \times idf_t \tag{3.1}$$

U izrazu 3.1, inverzna frekvencija dokumenata temeljem termina t (idf_t) određena je izrazom za izračun frekvencije dokumenata danim u retku *t* tablice 3.1. Izraz $idf_t = log\frac{N}{d_t}$ poprima vrijednost 0 kad je termin t prisutan unutar svih dokumenata u korpusu – tada je taj termin za potrebe klasifikacije efektivno zaustavna riječ te se, pridjeljivanjem težine, koja prema izrazu 3.1, u tom slučaju također iznosi 0, njegova relevantnost u kontekstu klasifikacije teksta poništava.

U mjeri tf-idf postoji čitav niz mogućih odabira sheme frekvencije dokumenata ($tf_{t,d}$), od kojih se svaka u literaturi opravdava numeričkom izražajnošću u smislu iskazivanja mjere relevantnosti termina na razini dokumenta.

U ovom radu odabrana je shema opisana retkom n tablice 3.1, koja predstavlja najjednostavniji oblik iskazivanja značajnosti termina t unutar dokumenta d – brojem pojavljivanja tog termina unutar samog dokumenta.

3.2. Vektor značajki

Modeliranje semantičke sličnosti jedinica teksta, kao što su to dokumenti, mjerama vektorske sličnosti pokazalo se kao dobra tehnika u domeni klasifikacije teksta [3]. Sličnost dvaju vektora najčešće se opisuje kosinusom kuta određenog tim vektorima:

$$cos(\theta) = \langle \hat{\boldsymbol{v}}, \hat{\boldsymbol{w}} \rangle \tag{3.2}$$

U izrazu 3.2 opisana je metoda izračuna kosinusa kuta (θ), tj. kuta između vektora $\boldsymbol{v}$ i $\boldsymbol{w}$, korištenjem skalarnog umnoška jediničnih vektora – vektora smjera $\boldsymbol{v}$ odnosno $\boldsymbol{w}$. Trigonometrijska funkcija *kosinus* pogodna je za izražavanje različitosti vektora iz razloga što vrijedi nejednakost $0 \leq |cos(\theta)| \leq 1$, gdje vrijednost 0 kosinusa kuta θ upućuje na ortogonalnost dvaju vektora (najveći stupanj različitosti), a vrijednost 1 na vektore koji leže na istom pravcu i istoga su smjera, odnosno vrijednost -1 na vektore koji leže na istom pravcu, ali su međusobno suprotnog usmjerenja.

Modeliranjem dokumenata vektorima značajki, gdje svaka značajka predstavlja težinu dodijeljenu nekom terminu t, a vrijednost te značajke za dokument d u pripadnom vektoru mjeru izraženosti samog termina predstavljenog značajkom unutar

dokumenta, dobiva se jednostavna mjera sličnosti između neka dva dokumenta, koja proizlazi iz skalarnog umnoška vektorskog prikaza tih dokumenata.

Mjera sličnosti dokumenata omogućava korištenje mnogih metoda klasificiranja temeljem strojnog učenja. Nekih od tih metoda su algoritmi grupiranja (na primjer, algoritam k najbližih susjeda i algoritam k srednjih vrijednosti), koji spadaju u domenu nenadziranog učenja (engl. *unsupervised learning*). Nenadziranim učenjem mogu se, ugrubo, smatrati one metode strojnog učenja koje ne zahtijevaju prethodno (najčešće od strane čovjeka) označene podatke.

S druge strane, metode nadziranog učenja (engl. *supervised learning*) temelje se na označenim skupovima podataka (engl. *labeled data sets*), na temelju kojih se potom izgrade klasifikatori. Neki od modela temeljenih na nadziranom učenju su naivni Bayesov klasifikator, koji se temelji na probabilističkoj klasifikaciji korištenjem vektora značajki te model stroja potpornih vektora.

3.3. Stroj potpornih vektora

3.3.1. Model

Stroj potpornih vektora (engl. *support vector machine, SVM*), poznat i kao *mreža potpornih vektora* (engl. *support-vector network*) linearni je diskriminativni model koji omogućava binarnu klasifikaciju podataka predstavljenih vektorima, predstavljen 1995. godine u radu [1].
Klasifikacija strojem potpornih vektora sastoji se u podjeli danih vektora na pozitivne i negativne, što znači da SVM rješava problem binarne klasifikacije.

Model se temelji na kriteriju *maksimalne margine* (engl. *maximum margin*). Postavljanjem granice takve da je prostor između pozitivnih i negativnih primjera što veći postiže se podjela vektorskog prostora na potprostor koji pripada pozitivnim primjerima i potprostor koji pripada negativnim primjerima, što omogućuje klasifikaciju ulaznog vektora temeljem njegove pripadnosti jednom od dva potprostora.

Linearan model SVM-a opisan je sljedećom jednadžbom:

$$h(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) + w_0 \tag{3.3}$$

U izrazu 3.3 je $\boldsymbol{x} \in R^n$ ulazni primjer, dok je $\boldsymbol{\phi} : R^n \to R^m$ funkcija koja preslikava ulazni primjer u m-dimenzionalni prostor značajki. Jednakošću $h(\boldsymbol{x}) = 0$, odnosno $\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) + w_0 = 0$ određena je m-dimenzionalna hiperravnina odluke (engl. *decision hyperplane*), skalarni umnožak čije normale, $\boldsymbol{w}$, i vektora značajki danog primjera,

x, određuje njihov odnos – vektori značajki za koje izraz daje vrijednost pozitivnog predznaka nalaze se *„iznad"* hiperravnine odluke, dok su primjeri za koje izraz daje negativno predznačenu vrijednost, koja se nalazi *„ispod"* hiperravnine.
Na taj način definirana je za neki primjer $\boldsymbol{x}$ i predikcija, temeljem predznaka vrijednosti dobivene uvrštavanjem vektora značajki primjera u jednadžbu hiperravnine odluke (3.3): $y = sgn(h(\boldsymbol{x}))$, gdje vrijedi $y \in \{-1, 1\}$.

3.3.2. Problem maksimalne margine

Za izgradnju klasifikatora modelom stroja potpornih vektora potrebno je, na temelju skupa primjera za učenje ($\mathcal{D}$), izračunati koeficijente hiperravnine odluke: njezinu normalu, $\boldsymbol{w}$ i konstantu w_0, prema izrazu 3.3.
Pretpostavka linearne razdvojivosti primjera iz skupa $\mathcal{D}$ (ili linearne razdvojivosti njihovih pripadnih vektora značajki, dobivenih preslikavanjem $\boldsymbol{\phi}$) povlači postojanje $\boldsymbol{w}$ i w_0 za koje vrijedi $h(\boldsymbol{x_i}) \geq 0$ za $y_i = 1$, odnosno $h(\boldsymbol{x_i}) \leq 0$ za $y_i = -1$, $\forall x_i \in \mathcal{D}$.

Udaljenost nekog primjera $\boldsymbol{x_i}$ od hiperravnine $h(\boldsymbol{x}) = 0$ dana je izrazom $\|h(\boldsymbol{x_i})\|/\|\boldsymbol{w}\|$, koji se, zbog toga što vrijedi nejednakost $y_i h(\boldsymbol{x_i}) \geq 0$, može zapisati i kao:

$$\frac{y_i h(\boldsymbol{x_i})}{\|\boldsymbol{w}\|} = \frac{y_i(\boldsymbol{w^T}\boldsymbol{\phi}(\boldsymbol{x_i}) + w_0)}{\|\boldsymbol{w}\|}$$

Kako je margina jednaka udaljenosti najbližeg primjera iz skupa $\mathcal{D}$, ona se može opisati sljedećim izrazom:

$$\frac{1}{\|\boldsymbol{w}\|} \min_i \left\{ y_i(\boldsymbol{w^T}\boldsymbol{\phi}(\boldsymbol{x_i}) + w_0) \right\} \tag{3.4}$$

Temeljem definicije margine (3.4) moguće je formulirati optimizacijski problem pronalaska maksimalne margine:

$$\arg\max_{\mathbf{w}, w_0} \left\{ \min_i \left\{ y_i(\boldsymbol{w^T}\boldsymbol{\phi}(\boldsymbol{x_i}) + w_0) \right\} \right\} \tag{3.5}$$

Vektore koji leže na samim rubovima margine (s pozitivne i negativne strane) nazivamo *potpornim vektorima* (engl. *support vectors*), otkud potječe i sam naziv modela. Potporni vektori jedini su vektori koji određuju klasifikator te time i jedini koji su bitni za klasifikaciju – nakon učenja modela, ostale vektore možemo zanemariti.

Uzimajući u obzir da skaliranje normale $\boldsymbol{w}$ i koeficijenta w_0 hiperravnine odluke proizvoljnom realnom konstantom (izuzev 0) ne utječe na iznos udaljenosti između hiperravnine i primjera može se, bez gubitka općenitosti, skalirati ravninu nekom kon-

stantom $\alpha \in \mathbb{R}$ (uz uvjet $\alpha \neq 0$), takvom da vrijedi:

$$y_i \left(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x_i}) + w_0 \right) \geq 1, \qquad 1 \leq i \leq N, i \in \mathbb{Z} \tag{3.6}$$

U izrazu 3.6 broj primjera za učenje označava se simbolom N ($N = |\mathcal{D}|$). Jednakost $y_i \left(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x_i}) + w_0 \right) \geq 1$ postiže se za one $\boldsymbol{x_i}$ koji su najbliži ravnini $h(\boldsymbol{x}) = 0$. Širina maksimalne margine tada iznosi $2/\|\boldsymbol{w}\|$, budući da je udaljenost od margine do najbližeg pozitivnog primjera jednaka $1/\|\boldsymbol{w}\|$, kao i udaljenost margine do najbližeg negativnog primjera. Stoga se optimizacijski problem pronalaska maksimalne margine (3.5) može izraziti i na sljedeći način:

$$\arg\min_{\boldsymbol{w}, \boldsymbol{w_0}} \frac{1}{2} \|\boldsymbol{w}\|^2 \tag{3.7}$$

S obzirom na to da se suštinski radi o problemu konveksne optimizacije (pronalasku jedinog minimuma optimizirane funkcije), mogu se primijeniti specijalizirani algoritmi, kao što je to, na primjer, algoritam *stohastičkog gradijentnog spusta* (engl. *stochastic gradient descent, SGD*).

Kombiniranjem izraza 3.7 i ograničenja 3.6 moguće je formulirati problem kvadratnog programiranja (engl. *quadratic programming*), vremenske složenosti rješavanja $\mathcal{O}(n^3)$. Problem kvadratnog programiranja moguće je, metodom Lagrangeovih multiplikatora, svesti na dualni oblik.

Kodiranje svakog od ograničenja iz 3.6 jednim Lagrangeovim multiplikatorom α_i rezultira sljedećom Lagrangeovom funkcijom:

$$L(\boldsymbol{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2} \|\boldsymbol{w}\|^2 - \sum_{i=1}^{N} \alpha_i \left[y_i (\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x_i}) + w_0) - 1 \right] \tag{3.8}$$

U izrazu 3.8 $\boldsymbol{\alpha}$ je vektor sastavljen od svih N Lagrangeovih multiplikatora. Lagrangeova funkcija (3.8) može se svesti i na dualni oblik:

$$\tilde{L}(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{\phi}(\boldsymbol{x_i})^T \boldsymbol{\phi}(\boldsymbol{x_j}) \tag{3.9}$$

Korištenjem dualnog oblika (3.9) može se formulirati optimizacijski problem koji je istovjetan problemu danom izrazom 3.7:

$$\begin{aligned} &\underset{\boldsymbol{\alpha}}{\textit{maksimiziraj}} && \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{\phi}(\boldsymbol{x_i})^T \boldsymbol{\phi}(\boldsymbol{x_j}) \\ &\textit{uz ograničenja} && \alpha_i \geq 0, \qquad 1 \leq i \leq N, \ i \in \mathbb{Z} \\ & && \sum_{i=1}^{N} y_i \alpha_i = 0 \end{aligned} \tag{3.10}$$

Kao i 3.7, dualni problem 3.10 također je problem konveksne optimizacije, ali je sada broj varijabli po kojima se funkcija optimizira N (broj primjera za učenje), za razliku od n (dimenzionalnost prostora značajki) u izvornom problemu. Također, bitno je napomenuti da postoje algoritmi koji problem 3.10 mogu riješiti uz vremensku složenost $\mathcal{O}(N^2)$, što može biti bitno vremenski efikasnije u odnosu na složenost rješavanja problema u primalnom obliku, $\mathcal{O}(n^3)$, ovisno o odnosu broja primjera za učenje i broja značajki.

3.3.3. Jezgrena funkcija

S obzirom na to da se u izrazu 3.10 funkcija $\boldsymbol{\phi}$ pojavljuje isključivo u skalarnom umnošku oblika $\boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\phi}(\boldsymbol{x'})$, kojim se izražava mjera različitosti pojedina dva primjera $\boldsymbol{x}$ i $\boldsymbol{x'}$, moguće definirati funkciju koja će zamijeniti taj skalarni umnožak:

$$\kappa(\boldsymbol{x}, \boldsymbol{x'}) = \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\phi}(\boldsymbol{x'}) \tag{3.11}$$

Funkcija κ (3.11) naziva se *jezgrenom funkcijom* (engl. *kernel function*), a zamjena spomenutog skalarnog umnoška tom funkcijom *jezgrenim trikom* (engl. *kernel trick*).

Jezgreni trik omogućava jednostavniju definiciju mjere različitosti dva ulazna primjera te klasifikaciju vektora koji početno nisu linearno razdvojivi, primjenom specifičnih jezgrenih funkcija. Neke od jezgrenih funkcija su:

$$\begin{aligned}
&\kappa(\boldsymbol{x}, \boldsymbol{x'}) = \boldsymbol{x}^T\boldsymbol{x'} && \text{(linearna)}\\
&\kappa(\boldsymbol{x}, \boldsymbol{x'}) = (\boldsymbol{x}^T\boldsymbol{x'})^p && \text{(polinomna)}\\
&\kappa(\boldsymbol{x}, \boldsymbol{x'}) = \|\boldsymbol{x} - \boldsymbol{x'}\| && \text{(homogena)}
\end{aligned}$$

Vrijednosti jezgrene funkcije mogu se unaprijed izračunati za sve parove primjera za učenje:

$$\boldsymbol{K} = \begin{pmatrix}
\kappa(\boldsymbol{x_1}, \boldsymbol{x_1}) & \kappa(\boldsymbol{x_1}, \boldsymbol{x_2}) & \dots & \kappa(\boldsymbol{x_1}, \boldsymbol{x_N}) \\
\kappa(\boldsymbol{x_2}, \boldsymbol{x_1}) & \kappa(\boldsymbol{x_2}, \boldsymbol{x_2}) & \dots & \kappa(\boldsymbol{x_2}, \boldsymbol{x_N}) \\
\vdots & \vdots & \ddots & \vdots \\
\kappa(\boldsymbol{x_N}, \boldsymbol{x_1}) & \kappa(\boldsymbol{x_N}, \boldsymbol{x_2}) & \dots & \kappa(\boldsymbol{x_N}, \boldsymbol{x_N})
\end{pmatrix} \tag{3.12}$$

Matrica $\boldsymbol{K}$ (3.12) naziva se *Gram matricom* (engl. *Gram matrix*). Izračun Gram matrice praktičan je za jezgrene funkcije određene algoritmima koje je teže opisati matematičkim izrazima, pri čemu je prilikom predikcije (klasifikacije) potrebno računati vrijednost primjene jezgrene funkcije na vektor za koji se računa predikcija i svaki od primjera za učenje.

3.4. Primjena

U okviru ovog rada, za osnovni pristup klasifikaciji teksta korišten je stroj potpornih vektora linearne jezgrene funkcije.
S obzirom na broj značajki reda veličine 10^5, za klasifikaciju dokumenata odabranih korpusa korištena je biblioteka LIBLINEAR [2], koja pruža implementaciju algoritma za treniranje stroja potpornih vektora linearne jezgrene funkcije u vremenskoj složenosti $\mathcal{O}(n)$, gdje je n broj značajki prisutnih u skupu primjera za učenje.

Uz zaustavne riječi i interpunkcijske znakove, iz ulaznog teksta uklonjene su sve pojavnice koje započinju ili završavaju znamenkom. Motivacija iza tog postupka je eksperimentalno utvrđena činjenica da su, u korištenim korpusima, takve pojavnice uglavnom jedinstvene na razini cijelog korpusa.

3.5. Klasifikacija s višestrukim oznakama

Stroj potpornih vektora može se izravno primijeniti na klasifikaciju s jednom oznakom, odnosno dvije, gdje druga oznaka predstavlja negaciju prve. To se naziva klasifikacijom s jednom oznakom (engl. *single-label classification*). Kod klasifikacije teksta često se radi o više od dvije oznake, što se naziva klasifikacijom s višestrukim oznakama (engl. *multi-label classification*), na što se model stroja potpornih vektora ne može izravno primijeniti.

Jedan od pristupa rješavanju problema klasifikacije s višestrukim oznakama korištenjem SVM-a sastoji se od podjele izvornog problema na manje potprobleme, od kojih je svaki zasebni problem klasifikacije s jednom oznakom, što se može izvesti na sljedeći način:
Neka je $\mathcal{L}$ skup svih oznaka, gdje vrijedi $|\mathcal{L}| > 2$, i neka je potrebno svakom od danih dokumenata $d \in \mathcal{D}$ pridijeliti jedan od podskupova skupa svih oznaka, $l_d \in 2^{\mathcal{L}}$. Za određivanje skupa oznaka koji se pridjeljuje nekom dokumentu potrebno je izgraditi $|\mathcal{L}|$ modela, od kojih se svaki koristi za binarnu klasifikaciju pripadne oznake. Predikcijom svakog od tako dobivenih modela utvrđuje se treba li pripadnu oznaku pridijeliti danom dokumentu ili ne. Takav je pristup korišten i u ovom radu.

Važno je napomenuti da se prethodni pristup temelji na pretpostavci da su potproblemi izvornog problema međusobno nezavisni, što je slučaj u ispitnoj zbirci korištenoj u ovom radu, budući da prisutnost neke oznake u skupu oznaka bilo kojeg od dokumenata nije ni na koji način označivaču upućivala na prisutnost neke druge.

4. Semantička jezgrena funkcija

4.1. Motivacija

U pristupu temeljenom na terminima kao značajkama, opisanom u poglavlju 3, mjera sličnosti između dva dokumenta temelji se na pojavljivanju istih termina unutar tih dokumenata. U jeziku postoje načini da se istim izrazima, ovisno o kontekstu pridijele različita značenja (višeznačnice), kao i načini da se različitim izrazima pridijele identična značenja (sinonimi), što će, primjenom modela temeljenog na terminima i znanju dobivenom isključivo na izrazima koji se pojavljuju u dokumentu rezultirati klasifikacijom nešto lošije kvalitete od očekivane.
Jedan takav primjer je i par rečenica:

„Oštrom se kosom trava lako kosi."
„Kosa štiti glavu od sunca."

U obje rečenice će se, nakon leksičke analize i lematizacije pojaviti termin *kosa*. Uzimajući to u obzir, klasifikacija modelom koji se temelji isključivo na vektorima termina će, uz pretpostavku da taj termin ne biva poništen nekom pridijeljenom frekvencijom (3.1), temeljem prisutnosti tog termina u obje rečenice rezultirati većom vrijednošću mjere sličnosti između te dvije rečenice, nego što bi to bio slučaj da ti termini nisu prisutni u obje rečenice – iako je njihov izraz isti, radi se o dva leksema različitih značenja. U prvoj rečenici termin *kosa* označava alat, ukrivljen dug nož na držalu, dok u drugoj rečenici *kosa* predstavlja vlasi na glavi čovjeka (po definiciji preuzetoj iz [5]).

U tekstu se nerijetko pojavljuju i *sinonimi*, posebice u kolekcijama temeljenim na novinskim člancima različitih autora, prikupljenim kroz dulje vrijeme. Sinonimi su leksemi različitih izraza i identičnog značenja. Problem prethodno opisanog modela se kod sinonima prisutnih u različitim dokumentima odražava na nemogućnost uparivanja sinonima temeljem njihovog značenja. Prilikom izračuna mjere sličnosti između dva dokumenta, sinonimi prisutni u oba dokumenta neće pozitivno utjecati na vrijednost koja opisuje njihovu sličnost – štoviše, u slučaju u kojemu je leksem prisutan isključivo

u jednom od dokumenata, a njegov sinonim isključivo u drugom, oni će negativno utjecati tu vrijednost. To je slučaj u sljedećem paru rečenica:

„Želim otići u knjižnicu."
„Htio bih posjetiti biblioteku.."

Termin *knjižnica* u prvoj rečenici i termin *biblioteka* u drugoj odnose se na isti pojam – radi se o ustanovi u kojoj se čuvaju i posuđuju knjige. Zbog pristupa koji se temelji isključivo na samim terminima pronađenim unutar dokumenata ta dva termina ni na koji način ne mogu biti promatrana na razini značenja, bez korištenja nekog oblika vanjskog znanja.

Wang i Domeniconi u svom radu [9] predlažu pristup koji se temelji na korištenju Wikipedije kao baze semantičkog znanja, koje se potom unosi u klasični tf-idf vektorski zapis dokumenata, koji je pogodan za klasifikaciju. Takav pristup nazivaju *semantičkom jezgrenom funkcijom* (engl. *semantic kernel*), čiji naziv proizlazi iz konteksta korištenja semantički obogaćenog teksta, uz neku od tradicionalnih jezgrenih funkcija, za izvedbu klasifikacije modelom stroja potpornih vektora.

4.2. Wikipedija

4.2.1. Struktura

Wikipedija (engl. *Wikipedia*) je višejezična internetska enciklopedija slobodnog sadržaja, koju podupire neprofitna udruga Wikimedia Foundation. Sam projekt pokrenuli su L. Sanger i J. Wales 2001. godine, a počiva na principima slobodnog uređivanja, provjerljivosti i neutralne perspektive. Engleska inačica Wikipedije, u vrijeme pisanja ovog rada, brojila je preko 4.5 milijuna valjanih članaka, dok je ukupan broj svih članaka svih jezičnih inačica Wikipedije iznosio više od 30 milijuna.
Inačica Wikipedije na hrvatskom jeziku pokrenuta je 2003. godine. U vrijeme pisanja ovog rada, broj valjanih članaka bio je veći od $140\ 000$, čime je u poretku po broju članaka 39. po redu. Procjenjuje se[1] da ju održava aktivnih 547 suradnika, od $125\ 965$ regstriranih [11].

Članci Wikipedije organizirani su u kategorije, gdje kategorije predstavljaju skupine članaka ili kategorija slične tematike. Na primjer, članci *Brdski biciklizam* i *Cestovni biciklizam* nalaze se u kategoriji *Bicklizam*, koja se nalazi u kategoriji *Šport*. U daljnjem tekstu će se za spomenutu strukturu kategorija upotrebljavati termin *taksonomija*.

[1] Izračunato na temelju novih unosa ili izmjena na Wikipediji unutar prethodnih 30 dana.

Iako intuitivno opis taksonomije upućuje na to da se radi o strukturi koja se može formalizirati stablastim grafom, radi se o usmjerenom grafu s ciklusima. Uz usmjerenost grafa, koja se odražava u tome da kategorija sadrži dvosmjerne veze (u obliku poveznica) članke i potkategorije koje obuhvaća, ciklusi proizlaze iz činjenice da pojedini članak (ili potkategorija) može biti svrstana u više kategorija.

Podjelom članaka po tematskim kategorijama ostvaruje se mjera semantičke povezanosti – članci koji se nalaze unutar iste kategorije (na minimalnoj udaljenosti u grafu taksonomije) semantički su bliskiji u odnosu na članke između kojih je najkraći put u grafu taksonomije dulji (obuhvaća više kategorija).

U strukturi Wikipedije razrješavanje sinonimije modelirano je stranicama preusmjeravanja. Na primjer, podstranica Wikipedije *Auto* preusmjerava na članak *Automobil*, što istovremeno olakšava korisničko pretraživanje Wikipedije i tvori bazu termina, povezanih na osnovi sinonimije.

Višeznačnost termina je utjecala na strukturu Wikipedije potrebom za izdvajanjem stranica razrješavanja (engl. *disambiguation pages*), koje sadrže poveznice na članke čiji je naslov isti izraz, uz opis značenja svakog od ponuđenih izraza. Stranice razrješavanja također, kao što to čine stranice preusmjeravanja, ubrzavaju pretraživanje Wikipedije od strane korisnika, ali i, u kontekstu obrade prirodnog jezika, tvore bazu termina povezanih na temelju izraza.

4.2.2. Wikipedija kao izvor znanja

Zbog tematske i hijerarhijske organizacije kategorija i članaka Wikipedija se može koristiti kao izvor semantičkog znanja. U radu [10] poveznicama između članaka Wikipedije modelira se semantička povezanost među terminima, što uključuje relacije višeznačnosti (homonimije i polisemije) i sinonimije. Članci Wikipedije mogu se predstaviti kao koncepti – gdje je svaki koncept predstavljen pripadnim člankom Wikipedije.

Strukturirano znanje, kakvo pruža Wikipedija u obliku grafa članaka i kategorija, korišteno je u radu [12] za određivanje semantičke povezanosti dokumenata postupkom nasumičnog hoda (engl. *random walk*), uz različite strategije odabira posjećenih bridova na grafu. Predstavljanje Wikipedije grafom, s obzirom na tematsku podjelu članaka i kategorija, intuitivno modelira semantičku povezanost različitih koncepata (predstavljenih člancima), pod uvjetom da se ne narušava konzistentnost pravila kategorizacije određenih internim konvencijama Wikipedije od strane autora koji unose i izmjenjuju sadržaj.

4.3. Koncepti

Pristup autora u radu [9] temelji se na proširivanju vektora značajki semantičkim entitetima dobivenih na temelju povezivanja izvornih termina dokumenata s Wikipedijom. Ideja se temelji na intuiciji da tf-idf vektor proširen tako izlučenim *konceptima* bolje modelira semantičke značajke pripadnog dokumenta, čime se teži prema izgradnji kvalitetnijeg klasifikatora. Otud i naziv *semantička jezgrena funkcija* (engl. *semantic kernel*) – takva jezgrena funkcija, uz ulazni tf-idf vektor termina računa mjeru sličnosti dokumenata uz unošenje vanjskog semantičkog znanja, koje u ovom radu, po uzoru na [9], proizlazi iz Wikipedije.

Izlučivanje koncepata provedeno je u nekoliko koraka. Opis svakog od tih koraka slijedi kroz idućih nekoliko odjeljaka.

Filtriranje značajnih članaka

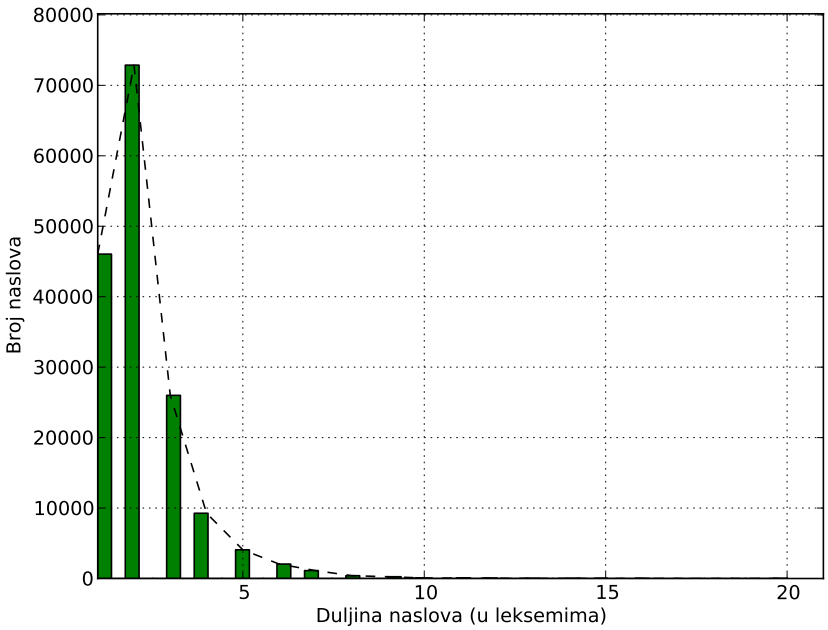
Filtriranje značajnih članaka temelji se na odbacivanju onih članaka Wikipedije za koje se procjenjuje da beznačajno malo, ako uopće, mogu utjecati na klasifikaciju dokumenata. U ovom slučaju, radi se o člancima koji zadovoljavaju barem jedan od sljedećih uvjeta:

1. Naslov članka sastoji se od više od 6 leksema
2. Naslov članka sastoji se isključivo od brojeva
3. Članak ne sadrži tekst

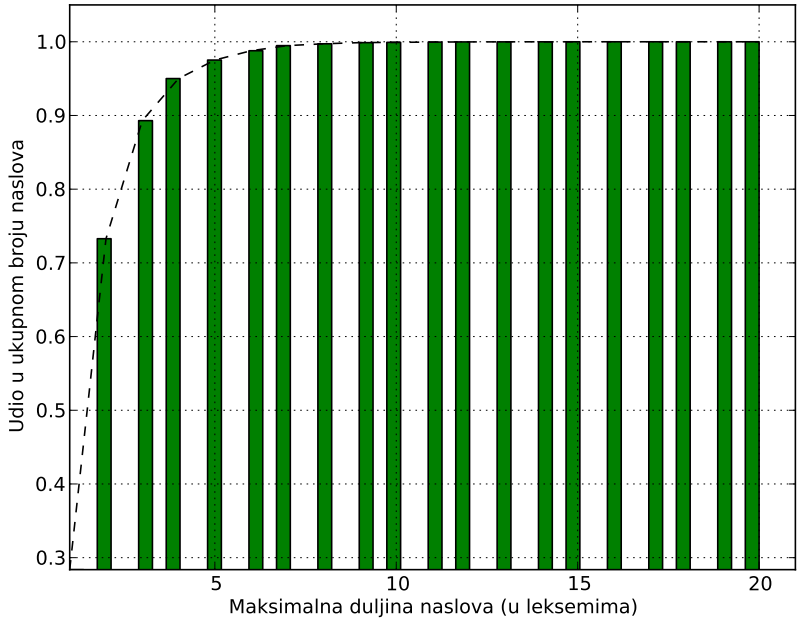
Motivacija za ograničavanje duljine naslova (izražene brojem leksema) članaka koji se koriste kao koncepti proizlazi iz analize raspodjele broja članaka po dokumentima.

Prema slici 4.1 vidljivo je da se odbacivanjem članaka čiji naslovi sadrže više od 6 leksema eliminiraju svega 1972 primjerka od njih ukupno $162\ 278$, čime se uvelike umanjuje složenost problema pronalaska koncepata u tekstu, uz manje značajnu cijenu izostavljanja dijela članaka. Drugim riječima, ograničavanjem duljine naslova na 6 leksema obuhvaća se 98.78 % članaka koji mogu predstavljati koncepte (slika 4.2). Više riječi o samom algoritmu korištenom za pronalazak koncepata bit će u jednome od odjeljaka koji slijede.

Izbacivanje članaka čiji se naslovi sastoje isključivo od brojeva provedeno je zbog toga što su pojavljivanja takvih članaka razmjerno rijetka. Najčešće se radi o specifičnim datumima i brojevima, čiji je sadržaj tipično stranica koja sadrži poveznice na



Slika 4.1: Raspodjela duljina naslova po člancima



Slika 4.2: Udio članaka obuhvaćenih ograničenjem duljine naslova

više tematski nesrodnih članaka, kao i kratki tekst koji opisuje svaki od njih. Na primjer, takvi su članci kategorije *Godine 20. stoljeća*. Članak posvećen jednoj od godina ukratko opisuje svaki od događaja koji su se zbili te godine, uz poveznicu na članak sadržaja koji rječitije i detaljnije opisuje taj događaj.

Leksička i morfosintaktička analiza

Kako bi pronalazak koncepata u tekstu bio što jednostavniji, naslovi članaka koji su prošli filtriranje po broju leksema podvrgnuti su istom postupku leksičke analize i morfosintaktičke normalizacije kao i dokumenti koji se klasificiraju, što je opisano u poglavlju 3.

Izlučivanje koncepata iz dokumenta

Nakon filtriranja, leksičke i morfosintaktičke obrade naslova slijedi izlučivanje koncepata iz teksta dokumenti se podvrgavaju postupku koji pretražuje njihov sadržaj s ciljem pronalaska koncepata. Kako je sam problem pronalaska koncepata zaseban problem pretraživanja tekstne informacije, radi jednostavnosti je odabran jednostavni algoritam, koji se temelji na usporedbi nizova normaliziranih leksema sa skupom naslova te uparivanjem koncepta s dokumentom u slučaju da se je pripadni naslov izražen nekim od nizova leksema koji se nalaze unutar tog dokumenta. Uzevši u obzir da je veličina naslova članaka koji predstavljaju valjane koncepte nakon filtriranja (4.3) manja ili jednaka 6 leksema, moguće je pretraživanje koncepata izvršiti algoritmom traženja podniza unutar niza koji predstavlja dani dokument, nakon kodiranja leksema cjelobrojnim identifikatorima njihovih lema, što je i izvedeno u programskom ostvarenju postupka opisanog u ovom radu. Bitno je napomenuti da se, u slučaju da je neki n-gram (niz n leksema), koji predstavlja koncept, sadržan u m-gramu, koji također predstavlja koncept, gdje vrijedi $m > n$, tada se kao koncept pronađen u tekstu uzima samo onaj koji je određen m-gramom.

Razlika između takvog postupka izlučivanja koncepata iz dokumenata u odnosu na rad [9] je u morfosintaktičkoj analizi – u radu [9] primijenjena je tehnika potpunog poklapanja (engl. *exact match*), pri čemu se nizovi pojavnica u izvornom obliku uparuju s naslovima članaka. Postupak morfosintaktičke normalizacije nad naslovima članaka uveden je zbog visoko flektivne prirode hrvatskog jezika.

4.4. Konceptualne značajke

Vektor značajki korišten u pristupu koji se temelji na semantičkoj jezgrenoj funkciji sastoji se od tf-idf vektora termina, opisanog u poglavlju 3. Takav vektor proširuje se značajkama koje predstavljaju koncepte, odnosno njihovu tf-idf mjeru, temeljem semantičke povezanosti dokumenta i tih koncepata te koncepata koji su srodni onima koji su pronađeni unutar dokumenta. Razlika u odnosu na izvorni rad koji opisuje semantičku jezgrenu funkciju je pritom izostavljanje modeliranja hiponomije, kao hijerarhijskog odnosa među konceptima, zbog toga što je, u odnosu na inačicu Wikipedije na engleskom jeziku, za hrvatski jezik teže izlučiti takav odnos temeljem Wikipedije, budući da relativno manji broj kategorija sadrži pripadne članke, kojima bi se iste mogle modelirati kao cjeloviti koncepti.

4.4.1. Višeznačnice

Budući da su višeznačnice na Wikipediji obuhvaćene stranicama razrješavanja slijedi da je izraz unutar danog dokumenta, koji se poklapa s nazivom stranice razrješavanja (odnosno izrazom višeznačnice koje ona obuhvaća) potrebno povezati s jednim od koncepata na koje usmjerava takva stranica. Za razlučivanje višeznačnica korišten je postupak uspoređivanja vektora termina dobivenog temeljem dokumenta sa svakim od vektora termina dobivenih na temelju teksta članaka na koji usmjerava stranica razrješavanja te odabirom onog koncepta (članka) čija je sličnost s dokumentom, u smislu kosinusa kuta između vektora termina, najveća. Takav način razrješavanja razdvojbe ovisi o opsežnosti različitih članaka na koje upućuje stranica razrješavanja. U slučaju da neki od članaka kandidata sadrže značajno manju količinu teksta od drugih tada će njihova izražajnost biti manja te će se oni u pravilu rjeđe pojavljivati kao odabrani koncepti koji slijede iz razdvojbi, što djelomično slijed i iz problema nedovršenosti dijela članaka Wikipedije.

4.4.2. Sinonimi

Sličnost temeljena na sinonimiji također se postiže proširivanjem vektora značajki konceptima izlučenim iz dokumenata. Svaki vektor značajki proširuje se, uz koncept pronađen pretraživanjem pripadnog dokumenta, konceptima koji na Wikipediji preusmjeravaju na taj koncept (engl. *redirects*), koji poprimaju istu vrijednost kao i izvorni koncept, na temelju kojeg su oni uvedeni u vektor značajki.

4.4.3. Srodni koncepti

Koncepti se smatraju srodnima u slučaju da postoji (jednosmjerna ili dvosmjerna) poveznica među člancima Wikipedije koji ih predstavljaju. Dvosmjerna poveznica između članaka intuitivno upućuje na srodnost, dok jednosmjerna, pogotovo u slučaju hiponimije (značenje jednog koncepta je specijalizacija značenja drugog) može modelirati podjednako izraženu srodnost. Obrazloženje korištenja i jednosmjernih veza za modeliranje srodnosti detaljnije je opisano u radu [10] uz prikladne primjere, otkud je sam postupak koji je izveden u ovom radu i preuzet.

Značajni koncepti uvode dodatne, *srodne koncepte* (engl. *related concepts*) u skup koncepata tog dokumenta. Uvođenje takvih koncepata motivirano je potrebom izražavanja stupnjevite sličnosti među konceptima koji nisu identični (ne preslikavaju se u značajke istog indeksa), koja se na taj način ostvaruje uz korištenje neke od konvencionalnih jezgrenih funkcija uz semantički prošireni vektor značajki.

4.4.4. Mjera sličnosti koncepata

Za potrebu određivanja srodnih koncepata uvodi se numerička mjera sličnosti (srodnosti) koncepata. Mjera sličnosti koncepata izražena je korištenjem tri različita kriterija usporedbe: kriterija određenog udaljenošću koncepata u grafu taksonomije, kriterija određenog tf-idf sličnošću tekstnog sadržaja pripadnih članaka Wikipedije te kriterija određenog brojem zajedničkih kategorija članaka Wikipedije na koje upućuju poveznice unutar oba dokumenta.

Kriterij udaljenosti u grafu taksonomije

Udaljenost unutar grafa taksonomije korištena je kao komponenta mjere sličnosti koncepata zbog tematske raspodjele članaka na Wikipediji. Ovaj kriterij će slične koncepte upariti na temelju toga što se njihovi pripadni članci nalaze u istim ili bliskim kategorijama, oslanjajući se pritom na usklađenost unosa s konvencijom organizacije sadržaja na Wikipediji. Kako mjera treba upariti članke kraćih udaljenosti na grafu, ovaj kriterij definira se kao vrijednost udaljenosti između dva dana koncepta na grafu taksonomije, normalizirana promjerom grafa (najvećom mogućom vrijednošću udaljenosti između dva koncepta u grafu taksonomije), čime se raspon mogućih vrijednosti kriterija svodi na interval $[0, 1]$. U matematičkom zapisu ovog kriterija u nastavku rada koristit će se vrijednost dana tim kriterijem oduzeta od 1, kako bi se izrazila mjera sličnosti koncepata, a ne različitosti, što izvorno predstavlja udaljenost na grafu.

U programskoj izvedbi sustava za izračun vrijednosti ovog kriterija korišten je algoritam pretraživanja u širinu (engl. *breadth-first search, BFS*) zbog činjenice da graf taksonomije nije težinski graf te nije stablo – svaka kategorija dijete može biti potkategorija više roditeljskih kategorija, koje mogu biti djeca iste kategorije – čime u grafu nije isključena mogućnost pojavljivanja ciklusa.

U formulama koje definiraju izračun sličnosti koncepata, za ovu mjeru koristit će se indeks tax.

Kriterij tf-idf sličnosti

Sličnost koncepata temeljena na skalarnom umnošku vektora frekvencija termina pripadnih članaka (tf-idf) koristi se kao jedan od kriterija koji čine mjeru sličnosti koncepata zbog činjenice da sami članci Wikipedije sadrže tekst, što ih čini dokumentima, tj. jedinicama teksta na koje se mogu primijeniti klasične mjere sličnosti dokumenata, od kojih je jedna i tf-idf. Vektorski umnožak tf-idf vektora normalizira se na jedinični interval dijeljenjem umnoškom normi tih dvaju vektora.

Matematička notacija za članove koji se odnose na ovaj kriterij izražena je indeksom $tfidf$.

Kriterij izlaznih poveznica

Posljednji i najjednostavniji kriterij je kriterij koji kao mjeru sličnosti dvaju koncepata uzima broj zajedničkih kategorija članaka na koje članci koncepata koji se uspoređuju sadrže poveznice. Ideja potječe iz pretpostavke da kategorije članaka koji su srodni odabranom članku sadrže informaciju o semantici tog članka. Vrijednost koja može poprimiti taj kriterij također je na intervalu $[0, 1]$, što se postiže dijeljenjem broja zajedničkih kategorija srodnih članaka manjim od dva broja kategorija srodnih članaka.

U matematičkim izrazima, za opis kriterija izlaznih poveznica koristit će se indeks out.

Sama mjera sličnosti među konceptima definira se kao linearna kombinacija prethodno navedenih kriterija:

$$\theta(c_i, c_j) = \lambda_{tax}\left(1 - \theta_{tax}(c_i, c_j)\right) + \lambda_{tfidf}\theta_{tfidf}(c_i, c_j) + \lambda_{out}\theta_{out}(c_i, c_j) \qquad (4.1)$$

U navedenom izrazu, konstante λ_{tax}, λ_{tfidf} i λ_{out} su eksperimentalno utvrđeni pozitivni cijeli brojevi, pri čemu su im pridijeljene vrijednosti 0.4, 0.4, odnosno 0.2, po uzoru na rad [9]. Funkcija $\theta : \mathbb{Z} \times \mathbb{Z} \to \mathbb{R}$ funkcija je koja određuje sličnost dva dana koncepta, c_1 i c_2, gdje je sličnost realna vrijednost, između 0 i 1, sukladno definiciji konstanti

i normalizaciji kriterija, dok funkcije θ_{tax}, θ_{tfidf} i θ_{out} preslikavaju par ulaznih koncepata, odnosno njihovih cjelobrojnih oznaka, u realne vrijednosti dane kriterijima na koje se odnose.

S obzirom na to da se sličnost dvaju koncepata opisuje realnom vrijednošću iz intervala $[0, 1]$, konstante θ_{tfidf} i θ_{out} jednoznačno određuju θ_{tax} te se izraz 4.1 može zapisati i na sljedeći način:

$$\theta(c_i, c_j) = (1 - \lambda_{tfidf} - \lambda_{out})\,(1 - \theta_{tax}(c_i, c_j)) \;+\; \lambda_{tfidf}\theta_{tfidf}(c_i, c_j) \;+\; \lambda_{out}\theta_{out}(c_i, c_j) \tag{4.2}$$

4.5. Proširenje vektorskog prostora modela

Model dokumenta kao vektora termina opisan u poglavlju 3 proširuje se značajkama koncepata. Za potrebe modeliranja konceptualne sličnosti dokumenata potrebno je uključiti mjere sličnosti među konceptima sadržanim u dokumentima. Svaki koncept predstavljen je zasebnom značajkom, s tim da se prisutnost koncepta u dokumentu također opisuje značajkama, koje u vektoru slijede nakon značajka dobivenih temeljem prisutnih termina.

Kako je temeljna ideja modeliranja dokumenta na konceptualnoj razini uparivanje koncepata koji se nužno ne pojavljuju u istom obliku ili intenzitetu, prisutnost nekog koncepta kao značajke povlači prisutnost srodnih koncepata u ovisnosti o intenzitetu srodnosti, kao i prisutnost sinonima, kojima se naznačuje da je dani koncept prisutan u dokumentu u nekom od mogućih oblika.

4.5.1. Matrica semantičkog modela

Pretvorba vektorskog zapisa dokumenta modeliranog terminima (tf-idf) može se opisati u dva koraka. Prvi korak je proširivanje $n-$dimenzionalnog vektora termina dodatnim značajkama koje predstavljaju koncepte prisutne u dokumentu, njih m :

$$\boldsymbol{\phi}(\boldsymbol{x}) = [t_0\ t_1\ \cdots\ t_{n-2}\ t_{n-1}] \rightarrow \boldsymbol{\phi_c}(\boldsymbol{x}) = [t_0\ t_1\ \cdots\ t_{n-2}\ t_{n-1}\ c_0\ c_1\ c_2\ \cdots\ c_{m+n-2}\ c_{m+n-1}] \tag{4.3}$$

Uvođenjem neke značajke c_i opisuje se prisutnost koncepta koji pripada toj značajki u dokumentu, tako da se toj značajki pridijeli vrijednost 0 ako koncept nije prisutan u dokumentu, 1 ako je prisutan samo na jednom mjestu, odnosno $n \in \mathbb{N}$ ako je koncept u dokumentu prisutan n puta. Ovakav način predstavljanja dokumenta na konceptualnoj

razini može se smatrati osnovnim modelom – u vektoru značajki uključeni su koncepti koji se nalaze u dokumentu, uključujući one dobivene prethodno opisanim postupkom razlučivanja višeznačnica. Nedostatak takvog pristupa je nemogućnost izražavanja semantički „slabijih" povezanosti među konceptima, kao što je to srodnost. Zbog toga se uvodi dodatna pretvorba vektora značajki, koja se može opisati sljedećom matricom:
Elementi podmatrice matrice $\boldsymbol{S}$ (a, b, c, ...) koja određuje proširenje konceptualnog

Tablica 4.1: Matrica $\boldsymbol{S}$

	Termini				Koncepti			
Termini	1	0	...	0	0	0	...	0
	0	1	...	0	0	0	...	0
	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮
	0	0	...	1	0	0	...	0
Koncepti	0	0	...	0	1	a	...	b
	0	0	...	0	a	1	...	c
	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮
	0	0	...	0	b	c	...	1

dijela vektora definiraju se na sljedeći način:

$$d_{i,j} = d_{j,i} = \begin{cases} i = j & 1 \\ c_i \text{ i } c_j \text{ su sinonimi} & 1 \\ c_i \text{ i } c_j \text{ su srodni} & \theta(c_i, c_j) \text{ (prema izrazu 4.1)} \end{cases}$$

Matricom $\boldsymbol{S}$ modelira se, uz razlikovanje višeznačnica i ujednačavanje sinonima korištenjem koncepata temeljenih na Wikipediji, i srodnost među konceptima. Nakon izgradnje takve matrice, vektor $\boldsymbol{\phi_c}(\boldsymbol{x})$ množenjem matricom S pretvara se u vektor $\boldsymbol{\phi_c}'(\boldsymbol{x})$, iz čega slijedi izraz za semantičku jezgrenu funkciju:

$$\kappa_s(\boldsymbol{x}, \boldsymbol{x'}) = \boldsymbol{\phi_c}'(\boldsymbol{x})\boldsymbol{\phi_c}'(\boldsymbol{x'})^T = \boldsymbol{\phi_c}(\boldsymbol{x})\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{\phi_c}(\boldsymbol{x'})^T \tag{4.4}$$

Klasifikacija dokumenata korištenjem semantičke jezgrene funkcije može se izvesti prethodnom pretvorbom vektora termina proširivanjem konceptima, uz modeliranje sličnosti koncepata množenjem vektora značajki matricom $\boldsymbol{S}$. Uz odabir linearne jezgrene funkcije, odnosno preslikavanja značajki $\boldsymbol{\phi} : \boldsymbol{x} \rightarrow \boldsymbol{x}$, za klasifikaciju je moguće koristiti postojeću linearnu jezgrenu funkciju.

U okviru ovog rada, klasifikator je izveden korištenjem stroja potpornih vektora linearne jezgrene funkcije, čemu prethodi provedba opisanog postupka proširivanja vektora značajki, čime je izbjegnut računski zahtjevan posao izgradnje Gram matrice za semantičku jezgrenu funkciju i skup dokumenata za učenje. Vremenska i prostorna složenost izračuna Gram matrice proizlazi i iz velikog broja jedinstvenih leksema dobivenih temeljem Wikipedije reda veličine 10^5. Na složenost izračuna također značajno utječe potreba za grafom udaljenosti najkraćih puteva među kategorijama Wikipedije. U vrijeme pisanja ovog rada, hrvatska Wikipedija obuhvaćala je 14390 različitih kategorija, što povlači $\binom{14390}{2}$ udaljenosti najkraćih puteva – u najboljem slučaju, memorijski zahtjev pohranjivanja takve strukture podataka iznosi otprilike 1 GB, pod uvjetom da su za same udaljenosti najkraćih puteva korišteni 8-bitni primitivi. Konačno, s obzirom na dimenzionalnost proširenog vektora semantičkog modela, koja iznosi 283534 za korištenu ispitnu zbirku, i sam postupak određivanja različitosti dvaju vektora doprinosi vremenskoj složenosti izračuna Gram matrice.

5. Eksperimentalno vrednovanje

5.1. Ispitna zbirka dokumenata

Model je ispitan na zbirci dokumenata NN13205. NN13205 naziv je kolekcije koja se sastoji od 13205 indeksiranih pravnih dokumenata, izdanih u časopisu *Narodne Novine*, službenom glasilu Republike Hrvatske. Kolekcija je nastala s ciljem sastavljanja instance višejezičnog, višedisciplinarnog pojmovnika Europske unije, poznatog pod nazivom „EuroVoc", u okviru projekta CADIAL[1] (Computer Aided Document Indexing for Accessing Legislation), u okviru kojega je razvijena javno dostupna semantička tražilica za pretraživanje pravnih dokumenata Republike Hrvatske.

Dokumentima je pridijeljeno 3951 od ukupno 6797 opisnika (engl. *descriptors*), određenih specifikacijom višejezičnog korpusa Eurovoc. Opisnici su organizirani u hijerarhijsku strukturu, koja se sastoji od 8 razina. Na prvoj razini nalazi se 21 vršnih opisnika. Prosječan dokument kolekcije sastoji se od oko 3000 leksema, Opisnici prve razine uzeti su kao ciljne vrijednosti za klasifikaciju dokumenata u ovom radu, što je jedna od mogućnosti predloženih u izvornom opisu korpusa ([14]).

S obzirom na to da je stroj potpornih vektora suštinski binaran klasifikacijski model, temeljem dokumenata ovog korpusa izgrađen je 21 klasifikator, od kojih je svaki stvoren s ciljem binarne klasifikacije pojedine oznake (vršnog opisnika).

Parametar C je za svaki od klasifikatora utvrđen usporedbom rezultata za različite vrijednosti iz skupa $\{0.01, 0.1, 1, 10, 100\}$.

5.2. Mjere vrednovanja

Iscrpno vrednovanje kvalitete izvedenih klasifikatora provedeno je nad zbirkom dokumenata NN13205. Kao mjere kakvoće modela korištene su preciznost (engl. *precision*), odziv (engl. *recall*) i F-mjera (engl. *F-score*), uz koje je priložena i točnost.

[1] http://cadial.hidra.hr

Preciznost klasifikacijskog modela za neku oznaku definira se sljedećim izrazom:

$$P = \frac{tp}{tp + fp} \tag{5.1}$$

U izrazu 5.1 oznakom *tp* (engl. *t*rue positive) predstavljen je broj primjeraka za koje se rezultati predikcije u smislu (pozitivne) prisutnosti neke oznake podudaraju s ručno pridijeljenom oznakom, dok *fp* (engl. *false positive*) predstavlja broj primjeraka za koje rezultati predikcije iste oznake upućuju na prisutnost te oznake, dok to nije slučaj kod ručno pridijeljene oznake.

Odziv modela definiran je kao omjer broja primjeraka za koje je prisutnost oznake pozitivno utvrđena predikcijom uz podudaranje predikcije (*tp*) i ručno pridijeljene oznake i sume tog broja i broja primjeraka za koje je prisutnost oznake u stvarnosti pozitivna, dok je predikcijom utvrđen suprotan rezultat (neprisutnost), što predstavlja oznaka *fn* (engl. *false negative*):

$$R = \frac{tp}{tp + fn} \tag{5.2}$$

F-mjera (engl. *F-measure, F-score*) harmonijska je sredina preciznosti i odziva. Smisao F-mjere je numerička vrijednost koja predstavlja kvalitetu klasifikatora, a podjednako ovisi o odzivu i preciznosti. F-mjera definirana je izrazom:

$$F = 2\frac{P \cdot R}{P + R} \tag{5.3}$$

Konačno, uz prethodno navedene metrike, izračunata je i točnost, kao omjer broja ispravnih oznaka dobivenih predikcijom i ukupnog broja mogućih oznaka u cijeloj kolekciji, pri čemu se uračunava i podudaranje neprisutnosti pojedine oznake s ručno pridijeljenim oznakama. Premda točnost nije naročito indikativna mjera kvalitete klasifikacijskog modela, ona je priložena radi detaljnijeg opisa ostvarene funkcionalnosti.

Provedeno je vrednovanje kvalitete tri različita klasifikatora na zbirci dokumenata NN13205. Za svaki od klasifikatora provedeno je peterostruko križno vrednovanje (engl. *5-fold cross-validation*), što uključuje podjelu dokumenata iz zbirke na skup za učenje, koji sadrži $4/5$ svih dokumenata, dok je ostatak korišten kao ispitni skup te ponavljanje postupka ispitivanja 5 puta, uz uzimanje prosječne vrijednosti mjera kakvoće klasifikacije. U svakom od pojedinih ispitivanja, dobivene mjere za svaku od oznaka agregiraju se u prosjek, koji se smatra konačnom mjerom kakvoće klasifikacije za taj slučaj raspodjele dokumenata na skup za učenje i skup za treniranje. Iznimka je F mjera, koja se računa temeljem odziva i preciznosti dobivenih prethodno opisanim postupkom. Prosječne mjere kakvoće klasifikatora izračunate na opisani način nazivaju se u literaturi *macro-average* mjerama [3].

Podjela dokumenata u ta dva skupa vrši se nasumično, po jednolikoj razdiobi.

5.3. Rezultati

Prvi klasifikator opisan je u poglavlju 3. Radi se o linearnom klasifikatoru, temeljenom na stroju potpornih vektora, koji ne koristi konceptualne (semantičke) značajke dobivene korištenjem Wikipedije. U tablici 5.1 taj je klasifikator određen nazivom *Osnovni*.

Drugi klasifikator opisan je u poglavlju 4. U tablici 5.1 za takav je klasifikator korišten naziv *SK*. Treći izvedeni klasifikator presjek je prvog i drugog klasifikatora – za klasifikaciju se koriste linearna jezgrena funkcija i isključivo koncepti, dok se sami termini ne koriste kao značajke. U tablici je taj klasifikator oslovljen nazivom *Čisti SK*. Za klasifikaciju teksta modelom semantičke jezgrene funkcije korišteno je ukupno 24904 različitih koncepata, dobivenih temeljem članaka Wikipedije.

Dobiveni rezultati prikazani su u tablici 5.1.

Tablica 5.1: Rezultati vrednovanja izvedenih klasifikatora

Model	F mjera	Odziv	Preciznost	Točnost
Osnovni	69.22 %	80.04 %	60.98 %	96.05 %
SK	70.83 %	82.42 %	62.01 %	96.07 %
Čisti SK	58.83%	72.21 %	49.64 %	94.72 %

5.4. Diskusija

Iz tablice rezultata vidljivo je da je razlika između klasifikatora semantičke jezgrene funkcije i klasifikatora temeljenog isključivo na vreći riječi tek oko 1.6% u smislu F mjere, temeljem sukladnih razlika u odzivu, odnosno preciznosti . Višestrukom provedbom postupka vrednovanja je utvrđeno da je razlika u F mjeri uglavnom posljedica stohastičke raspodjele dokumenata na skup za učenje i ispitni skup, s obzirom na varijacije u dobivenim iznosima mjera kakvoće klasifikatora. Značajni porast kakvoće klasifikacije modelom semantičke jezgrene funkcije, u odnosu na linearni model vreće riječi, nije ostvaren, u usporedbi s rezultatima autora izvornog modela semantičke jezgrene funkcije ([9]), gdje je primijećen porast u F mjeri iznosa 5 %.

Treći klasifikator (*Čisti SK*) izveden je kao svojevrsna provjera valjanosti semantičkih značajki u kontekstu klasifikacije.

Iako se modeliranjem dokumenata ostvaruje razlučivanje leksema temeljem značenja, a time i semantički precizniji opis dokumenta, radi se o postupku ovisnom o bazi znanja i postupku uparivanja svakog od dokumenata s pripadnim semantičkim značajkama, kao i o sadržajnoj domeni dokumenata nad kojima se vrši klasifikacija. Priprema podataka za učenje i predikciju je u tom slučaju vrlo bitna u smislu utjecaja filtriranja i obrade ulaznog skupa dokumenata na ishod vrednovanja klasifikatora.

Budući su autori rada [14] utvrdili nedostatke ispitne zbirke koja je korištena u ovom radu, jedna od od uzroka sličnosti rezultata klasifikacije modelom temeljenim na semantičkoj jezgrenoj funkciji modelom vreće riječi može se pripisati šumu, nastalom uslijed lošeg označavanja dokumenata zbirke.

Pretpostavka je da semantičke jezgrene funkcije bolje djeluju u klasifikaciji kraćih jedinica teksta, u smislu veće izražajnosti koncepata u tekstu čiji sadržaj svojim opsegom ne doprinosi klasifikaciji modelom vreće riječi, zbog relativno malog broja značajki. Osnovna ideja je oslanjati se na uvođenje koncepata srodnih konceptima koji su raspoznati u tekstu za rješavanje problema premalog broja značajki, kako bi se značajno pospješila kakvoća klasifikacije.

6. Zaključak

Zadatak ovog rada bio je izvesti model semantičke jezgrene funkcije prema radu [9] i linearni model, temeljen na vreći riječi, korištenjem stroja potpornih vektora te provesti usporedbu tih modela nad ispitnom zbirkom dokumenata na hrvatskom jeziku, s ciljem utvrđivanja primjenjivosti semantičke jezgrene funkcije u klasifikaciji teksta na hrvatskom jeziku.

Temeljem usporedbe rezultata klasifikacije dobivenih u ovom radu primjećuje se minimalan doprinos semantičkog proširenja klasičnog modela vreće riječi kod klasifikacije dokumenata pravne domene, iako je čisti semantički model (bez vreće riječi), u kontekstu F-mjere, dovoljan za izgradnju tek nešto manje točnog klasifikatora od jednog koji se temelji isključivo na vreći riječi, odnosno drugog, koji se temelji na modelu sematičke jezgrene funkcije, koja uključuje i vreću riječi i semantički model.

Nastavak istraživanja primjene semantičke jezgrene funkcije u klasifikaciji teksta na hrvatskom jeziku mogao bi se ostvariti u smjeru razvijanja naprednije mjere semantičke povezanosti koncepata, budući da su autori izvornog rada u kojem je opisan postupak semantičke jezgrene funkcije [9] koristili bazu znanja dobivenu obradom engleske Wikipedije, u čiji razvoj je uložen značajan trud, što je u konačnici rezultiralo radom [10]. Također, predlaže se i vrednovanje na zbirkama teksta pisanog u manje formalnom obliku, kao što su to, primjerice, *tweet*-ovi ili komentari, s obzirom na očekivano veću raznovrsnost koncepata koji se pojavljuju u takvom tekstu, u odnosu na ispitnu zbirku dokumenata koja je korištena u ovom radu.

LITERATURA

[1] C. Cortes i V. Vapnik. Support-vector networks. *Machine learning*, 1995.

[2] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, i C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 2008.

[3] C. D. Manning i H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[4] C. D. Manning, P. Raghavan, i H. Schütze. *An Introduction to Information Retrieval*. Cambridge University Press, 2007.

[5] Novi Liber and Srce. Hrvatski jezični portal, 2004. URL `http://hjp.novi-liber.hr/`.

[6] S. Robertson. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 2004.

[7] F. Sebastiani. Machine Learning in Automated Text Categorization. *ACM computing surveys (CSUR)*, 2001.

[8] K. Spärck Jones. Index term weighting. *Information Storage and Retrieval*, 1973.

[9] P. Wang i C. Domeniconi. Building Semantic Kernels for Text Classification using Wikipedia. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008.

[10] P. Wang, J. Hu, H.-J. Zeng, L. Chen, i Z. Chen. Improving Text Classification by Using Encyclopedia Knowledge. *Seventh IEEE International Conference on Data Mining*, 2007.

[11] Wikimedia Foundation. Wikipedia, 2001. URL `http://en.wikipedia.org/`.

[12] E. Yeh, D. Ramage, C. D. Manning, E. Agirre, i A. Soroa. WikiWalk: Random walks on Wikipedia for Semantic Relatedness. *TextGraphs-4 Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, 2009.

[13] Torsten Zesch, Christof Müller, i Iryna Gurevych. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. U *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco, Svibanj 2008. electronic proceedings.

[14] F. Šarić, B. Dalbelo Bašić, F.-M. Moens, i J. Šnajder. Multi-label Classification of Croatian Legal Documents Using EuroVoc Thesaurus. *Proceedings of SPLeT - Semantic processing of legal texts: Legal resources and access to law workshop*, 2014.

[15] J. Šnajder i B. Dalbelo Bašić. Automatic acquisition of inflectional lexica for morphological normalisation. *Information Processing and Management*, 2008.

Dodatak A
Tehnologija

A.1. Java

Java je statički tipiziran, objektno orijentiran i imperativan programski jezik, koji upravljanje memorijom temelji na postupku „sakupljanja smeća" (engl. *garbage collection*). Izdaje ga je 1995. godine tvrtka Sun Microsystems, a trenutno održava i nadograđuje tvrtka Oracle.

Specifičnost programskog jezika Java njegovo je izvođenje na virtualnom stroju (engl. *Java Virtual Machine, JVM*), koje je omogućeno prethodnim prevođenjem izvornog koda u međukod, *bytecode*. Izvršavanje programa na virtualnom stroju, uz odgovorno korištenje platformski neovisnih biblioteka, otvara mogućnost izvođenja Java programa na bilo kojem uređaju koji sadrži ispravno podešen Java virtualni stroj, što se opisuje krilaticom „*Write once, run anywhere*.".

Programski jezik Java odabran je za izvedbu opisanih postupaka zbog velike količine slobodno dostupnih biblioteka napisanih u istom te performansi koje su u praksi do najviše 2 puta sporije od programskog jezika C.

U ovom radu korišteno je standardno izdanje programske platforme Java 8 (Java SE 8), dostupno od 18. ožujka 2014., kojim je po prvi put u programski jezik Java uvedeno nekoliko jezičnih konstrukata tipičnih za funkcijsku paradigmu, kao što su to anonimne funkcije te funkcionalnosti mapiranja, redukcije i filtriranja nad temeljnim podatkovnim kolekcijama.

A.2. MySQL

Sustav za upravljanje relacijskom bazom podataka MySQL izdala je 1995. godine tvrtka Oracle, koja ga i održava. Radi se o popularnom, slobodno dostupnom sustavu

za upravljanje bazom podataka otvorenog koda (engl. *open-source*). U ovom radu sustav MySQL korišten je za pohranu strukturiranog sadržaja hrvatske Wikipedije, kojemu se potom pristupa preko JDBC (*Java Database Connectivity*) sučelja iz Java programa, za potrebe izračuna semantičkih značajki dokumenata.

Također, s obzirom na rad s razmjerno velikom količinom podataka, koji zahtijeva izvedba semantičke jezgrene funkcije, sve potrebne podatke nije bilo moguće u svakom trenutku imati pohranjene u radnoj memoriji. Za potrebe dohvaćanja podataka iz skupa podataka koji veličinom nadilazi ograničenje veličine gomile (engl. *heap*) od 4GB na računalu na kojem je razvijeno programsko rješenje, u programskom jeziku Java je iz sučelja `java.util.Map` izveden razred koji podatke pohranjuje u bazu podataka, kojom se upravlja putem sustava MySQL.

A.3. Priprema podataka

Sadržaj hrvatske Wikipedije preuzet je u XML formatu iz repozitorija dostupnog putem sljedeće poveznice: `dumps.wikimedia.org/hrwiki/`. S ciljem omogućavanja jednostavnijeg i efikasnijeg pristupanja podacima, iz XML zapisa izgrađen je niz SQL relacija i pripadnih tablica korištenjem alata dostupnih unutar biblioteke JWPL, kao i pomoćnih skripti napisanih u programskom jeziku Python, koje su korištene za ručne ispravke neskladno uređenih članaka Wikipedije.

Biblioteka JWPL (punim nazivom *Java Wikipedia Library*) predstavlja skup programskih alata razvijenih u programskom jeziku Java, čija je svrha olakšavanje korištenja podataka Wikipedije. Sama biblioteka temelji se na izlučivanju sadržaja Wikipedije iz XML zapisa, koji se potom sprema u bazu podataka, sukladno specifikaciji biblioteke. Potom se, kroz razrede programskog jezika Java u korisničkom kôdu, preko objektno-relacijskog preslikavanja (izvedenog alatom Java Hibernate) pristupa podacima Wikipedije korištenjem prikladnih Java razreda. Primjerice, zasebnim primjerkom razreda `Page` modelira se svaka od stranica Wikipedije, čiji je sadržaj izražen nizom znakova, naslov primjerkom razreda `Title`, kategorije primjercima razreda `Category`, uz dodatne atribute, koji pobliže opisuju sadržaj stranice [13].

Primjena semantičkih jezgrenih funkcija u klasifikaciji teksta

Sažetak

Klasifikacija teksta temeljem sadržaja jedan je od osnovnih zadataka koji se javljaju u domeni dubinske analize teksta. Često korišteni postupci uključuju predstavljanje dokumenata u vektorskom obliku, korištenjem vreća riječi. Iako su takve metode jednostavne i učinkovite, njima nije moguće modelirati dokument na konceptualnoj razini, što negativno utječe na kvalitetu klasifikatora koji se temelje na istima. Stoga je nedavno predloženo nekoliko različitih pristupa koji se temelje na povezivanju ontološkog znanja s tekstom. P. Wang i C. Domeniconi u svom radu iz 2008. opisuju jedan takav pristup, koji se temelji na korištenju Wikipedije za modeliranje dokumenata na semantičkoj razini, nazivajući ga „semantičkom jezgrenom funkcijom". U okviru ovog rada izvedena su dva klasifikatora temeljena na strojevima potpornih vektora – jedan koji koristi vreće riječi i drugi, koji se temelji na semantičkoj jezgrenoj funkciji. Točnosti klasifikacije oba modela uspoređene su temeljem iscrpnog vrednovanja provedenog na zbirci dokumenata na hrvatskom jeziku.

Ključne riječi: Jezgrena funkcija, klasifikacija teksta, stroj potpornih vektora, Wikipedija

Applying Semantic Kernel Functions in Text Classification

Abstract

Content-based text classification is one of the basic tasks in the domain of text analysis. Popular methods involve mapping text documents to bags of words, represented by vectors. Although quite effective in practice, such methods fail to describe text documents on a conceptual level, which negatively impacts the quality of the implementing classifiers. Several approaches which bind ontological knowledge to text documents have been proposed recently. P. Wang. and C. Domeniconi describe one such approach which relies on Wikipedia to represent documents on a semantic level in their 2008 paper, denoting the resulting model with the term „semantic kernel". Two support vector machine classifiers have been implemented as part of this thesis – one based on the traditional bag of words approach, the other being based on the semantic kernel. The performance of the models is then compared and evaluated by applying them on a collection of Croatian language documents.

Keywords: Kernel function, text classification, SVM, Wikipedia