Laboratorij za analizu teksta i inženjerstvo znanja – TakeLab

Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
Zavod za elektroniku, mikrolektroniku, računalne i inteligentne sustave
Unska 3, 10000 Zagreb, Hrvatska

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 1915

Ekstrakcija ključnih fraza metodama nadziranog strojnog učenja

Marko Horvatić

Zagreb, rujan 2012.

SADRŽAJ

1. Uvod **1**

2. Srodni radovi **3**

3. Skup podataka za učenje i evaluaciju **5**

4. Ekstrakcija ključnih fraza **6**

4.1. Predobrada 8

4.2. Generiranje kandidata 8

4.3. Značajke 10

4.3.1. Kandidat je u kategoriji 10

4.3.2. Kandidat je u naslovu 10

4.3.3. TFxIDF 10

4.3.4. Relativna prva pozicija lematizirane fraze 11

4.3.5. Broj riječi u frazi 11

4.3.6. Frekvencija najfrekventnije lematizirane riječi u frazi 11

4.3.7. Minimalan broj znakova 12

4.3.8. Prosječan broj znakova 12

4.3.9. Fraza je ključna fraza 12

4.3.10. Odbačene značajke 12

4.4. Izgradnja modela 12

4.4.1. Naivan Bayesov klasifikator 13

4.4.2. Višeslojni perceptron 13

5. Evaluacija i analiza rezultata **14**

5.1. Svrha, nedostaci i definicija evaluacije 14

5.2. Unija označivača 15

5.3. Prosjek označivača 17

5.4. Varijacija podskupova fraza prema označivačima 18

5.4.1. Evaluacija nad skupovima ključnih fraza oko kojih se slažu barem dva označivača . . . 18
5.4.2. Evaluacija nad skupovima ključnih fraza oko kojih se slažu barem tri označivača . . . 19
5.4.3. Evaluacija nad skupovima ključnih fraza oko kojih se slažu barem četiri označivača . . . 20
5.4.4. Analiza evaluacije nad podskupovima ključnih fraza . . . 21
5.5. Varijacija metoda diskretizacije . . . 22
5.6. Optimizacija . . . 24
5.6.1. Filtriranje na temelju vrste riječi . . . 24
5.7. Rezultati srodnih radova . . . 27
5.7.1. Ahel et al. (2009) . . . 27
5.7.2. Sarkar et al. (2010) . . . 27
5.7.3. Witten et al. (1998) . . . 28
5.7.4. Analiza rezultata . . . 28
5.8. Uklanjanje zavisnosti o kategoriji i korpusu . . . 29
5.9. Primjeri rada sustava . . . 31

6. Programska izvedba **33**
6.1. Programski alati . . . 33
6.1.1. Python . . . 33
6.1.2. Orange paket (Python) . . . 33
6.1.3. Git . . . 33
6.2. Organizacija koda . . . 34

7. Zaključak **36**

Literatura **37**

1. Uvod

Ključne fraze omogućuju kratak pregled sadržaja te brzo i jednostavno dohvaćanje dokumenata unutar zbirke. Kako broj dokumenata u digitalnom formatu raste, javlja se sve veća potreba za automatizacijom vezanja ključnih fraza uz sadržaj. Iako se od autora članka, bilo novinskog ili znanstvenog, očekuje da uz sadržaj veže ključne fraze, taj zahtjev često nije ispunjen. Angažiranje stručnjaka u tu svrhu neisplativo je te je potrebno razviti alate koji će automatizirati postupak vezivanja ključnih fraza uz dokumente.

Ključnu frazu definiramo kao skup od jedne ili više semantički vezanih riječi koje dobro opisuju sadržaj nekog dokumenta. Prije nego što uđemo u detalje postupka, bitno je napomenuti da postoje dvije grublje podjele vezivanja ključnih fraza: *dodjeljivanje ključnih fraza* i *ekstrakcija ključnih fraza*. *Dodjeljivanje ključnih fraza* odnosi se na postupak vezivanja ključnih fraza koje se ne nalaze nužno u dokumentu koji taj algoritam obrađuje. Ujedno, postupak koristi unaprijed definirani tezaurus (listu ključnih fraza, poželjno hijerarhijski organiziranu) kao izvor za preslikavanje fraza. S druge strane *ekstrakcija ključnih fraza* dodjeljuje isključivo one riječi koje se nalaze u obrađivanom tekstu. Označivači ključnih fraza nerijetko vežu fraze koje se ne nalaze u obrađivanom dokumentu. Turney (2000) u svom radu tvrdi da se prosječno oko 75% dodijeljenih ključnih fraza nalazi u sadržaju dokumenta. To znači da u slučaju ekstrakcije prosječna vrijednost mjere preciznosti nije u mogućnosti nadići granicu od 75%.

Ovaj rad opisuje sustav KPEX za automatsku *ekstrakciju ključnih fraza* iz tekstnih članaka na hrvatskom jeziku. Sustav je izgrađen korištenjem već provjerenih postupaka iz prethodnih radova kao što su: stvaranje liste kandidata sa statističkim značajkama, filtriranje liste kandidata i primjena metoda strojnog učenja za određivanje utjecaja ulaznih značajki na ciljnu značajku (kandidat ili je ili nije ključna fraza). Uz već iskušane postupke, u sustav KPEX uvode se i neke specifičnosti. Primjerice, filtriranje kandidata provodi se pomoću rangirane liste kombinacija vrsta riječi. Lista je sastavljena analizom učestalosti pojavljivanja kombinacija vrsta riječi među ključnim

frazama dokumenata za treniranje. Takav način filtriranja jamči usklađeniju listu kandidata u odnosu na inače korištene filtre koji su izgrađeni na ljudskoj procjeni. Sustav je treniran i evaluiran nad Hininim skupom dokumenata gdje primjeri dokumenata za testiranje sadrže ključne fraze dodijeljene od strane osam označivača.

U nastavku rada u poglavlju 2 opisani su srodni radovi gdje su istaknuti različiti pristupi problemu vezivanja ključnih fraza uz dokumente. Poglavlje 3 sadrži analizu skupa za treniranje i evaluaciju. Algoritam za ekstrakciju i pojedine faze algoritma opisane su u poglavlju 4. U poglavlju 5 razrađenja je temeljita evaluacija sustava KPEX nad raznim podskupovima unaprijed dodijeljenih fraza. Na kraju rada u poglavlju 6 opisane su programska izvedba i alati koji su pritom korišteni. U poglavlju 7 nalazi se zaključak koji sadrži osvrt na rad sustava KPEX, njegova ograničenja i moguća poboljšanja. Na kraju je navedena literatura na koju se rad poziva.

2. Srodni radovi

Sustav KPEX izgrađen je prema uzoru na radove sa sličnom motivacijom. Srodni radovi također se koriste za određivanje kvalitete izvedbe sustava KPEX. U ovom poglavlju opisani su najvažniji radovi, a rezultati evaluacije nekih od navedenih radova uspoređuju se sa sustavom KPEX u poglavlju 5.

Krulwich i Burkey (1996) koriste heuristiku za ekstrakciju ključnih fraza. Heuristika se temelji na sintaksnim oznakama kao što su kurzivi, akronimi i postojanje fraze u naslovu odjeljaka. Algoritam odabire veliku listu fraza sa slabom preciznošću.

Muñoz (1997) primjenjuje nenadzirano strojno učenje kako bi dodijelio fraze od dvije riječi. Temelji se na neuronskim mrežama ART (adaptivna rezonantna teorija). Aplikacija se svodi samo na dvorječne fraze.

Turney (2000) koristi dvije različite metode: Quinlanovo 4.5 stablo odluke s *bagging* tehnikom i GenEx. GenEx se sastoji od dva podsustava: Genitora i Extractora. Extractor je zadužen za prevođenje teksta u kandidate, njihovo filtriranje i dodjeljivanje značajki, dok je Genitor genetski algoritam koji služi za podešavanje parametara Extractora. Nakon što je završena faza učenja, parametri koji daju najbolje rezultate dalje se koriste za ekstrakciju. Samo podsustav Extractor ostaje aktivan.

Witten et al. (1998) svoj sustav KEA temelji na naivnom Bayesovom klasifikatoru te koristi samo dvije provjerene značajke: TDFxIF i relativnu prvu poziciju fraze u tekstu. Unatoč jednostavnosti ovaj algoritam postiže rezultate podjednake Turneyevom radu, koji je i bio temelj za KEA-u. KEA je zbog svoje jednostavne izvedbe vrlo koristan primjer načelnog rada algoritma za ekstrakciju ključnih fraza pa je tako uvelike doprinijela ovom radu.

KEA++ razvijena je od Medelyan i Witten (2006) kao nadogradnja na KEA-u. Temelji se na tezaurusu Agrovoc te znatno nadmašuje rezultate svog prethodnika. Ova metoda pripada kategoriji *dodjeljivanja ključnih fraza*.

Wang et al. (2006) predstavlja algoritam temeljen na neuronskim mrežama. Ekstrakciju ključnih fraza tretira kao klasifikacijski problem, što nije zadovoljavajuće u slučajevima kad je broj zatraženih fraza veći od broja onih pozitivno klasificiranih.

Sarkar et al. (2010) također koriste neuronske mreže, no odlučuju se na rangiranje vjerojatnosti da je kandidat ključna fraza umjesto na njegovo klasificiranje. Rezultati nadmašuju one dobivene sustavom KEA.

Sustav KPEX kao i KEA koristi naivan Bayesov klasifikator te značajke TDFxIF i relativnu prvu poziciju fraze. Rad Sarkar et al. (2010) postigao je jako dobre rezultate koristeći višeslojni perceptron što je bila dodatna motivacija za korištenje te metode u ovom radu. KEA, Sarkar et al. (2010) i KPEX omogućuju odabir broja ključnih fraza na izlazu za razliku od sustava koji opisuje Wang et al. (2006) gdje je broj fraza na izlazu ograničen brojem pozitivno klasificiranih fraza. Za razliku od Muñoz (1997) gdje je su ekstrahirane fraze isključivo dvorječne sustav koji ćemo opisati vraća fraze koje mogu sadržavati jednu, dvije ili tri riječi. Niti jedan od srodnih radova nije eksperimentirao s formiranjem filtra vrsta riječi na temelju unaprijed dodijeljenih ključnih fraza iz skupa za učenje što se pokazalo kao dobar odabir u izgradnji KPEX sustava.

3. Skup podataka za učenje i evaluaciju

Učenje i evaluacija rada sustava KPEX provedeni su nad skupom Hininih dokumenata posebno pripremljenim za tu namjenu. Dokumenti su početno formatirani kao XML-ovi koji se kasnije prevode u običan tekstni oblik kako bi se pojednostavilo njihovo dohvaćanje. U pretvorbi formata zadržani su sljedeći podaci: lista kategorija, ključne fraze, naslov i sadržaj. Skup dokumenata za evaluaciju sadrži skupove ključnih fraza dodijeljenih od strane čak osam označivača, što je omogućilo razne operacije nad skupovima kao što su unija i formiranje skupa ključnih fraza oko kojih se složilo N označivača. U slučaju kada je N jednak maksimalnom broju označivača (osam), skup je često prazan što narušava postupak evaluacije. Iz tog razloga evaluacija je provedena nad svima vrijednostima parametra N koji su vraćali samo neprazne skupove.

Tablica 3.1: Statistički podaci Hininih skupova za treniranje i evaluaciju

	Skup za treniranje	Skup za evaluaciju
Broj dokumenata	955	60
Broj riječi	292.000	19.754
Broj riječi po dokumentu	305,76	329,23
Broj riječi nakon obrade*	130.274	8.764
Prosjek riječi nakon obrade*	136,41	146,07
Broj ključnih fraza	4.355	1.202 / 209
Prosjek ključnih fraza	4,56	20,03 / 3,48

* obrada se odnosi na predobradu i filtriranje kandidata

Napomena: U tablici 3.1 broj ključnih fraza u skupu za evaluaciju prikazan je za slučaj unije skupova fraza svih označivača te za skup oko kojeg su se složila četiri označivača (najlošiji slučaj).

4. Ekstrakcija ključnih fraza

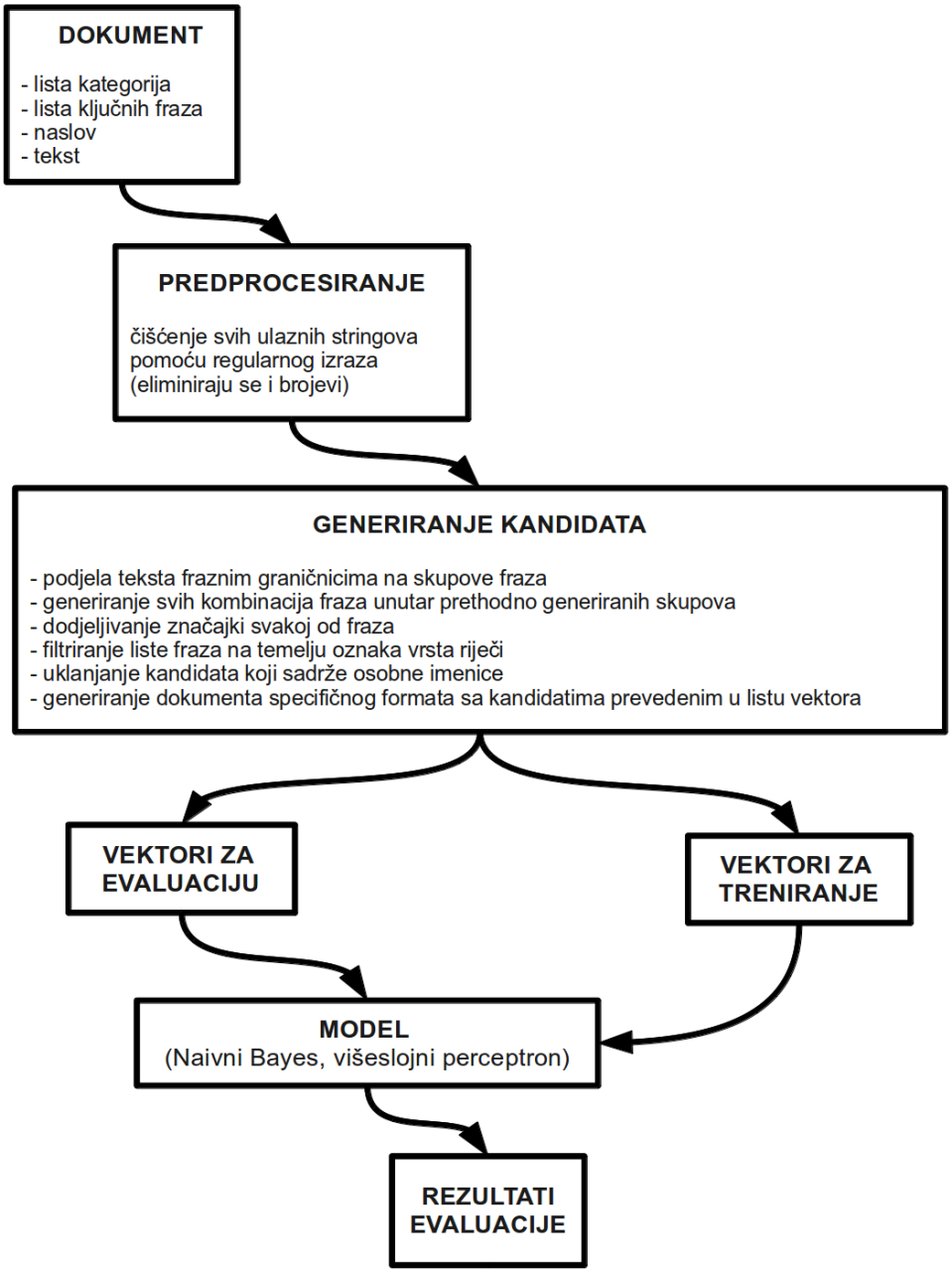
U poglavlju o srodnim radovima navedeni su različiti načini automatizacije vezivanja ključnih fraza uz dokument. Ovisno o cilju, odabire se odgovarajuća metoda strojnog učenja. Automatizira se ekstrakcija ključnih fraza i pritom se korisniku omogućuje kontrola nad brojem automatski dodijeljenih fraza.

Na ulazu sustava očekuje se sadržaj, naslov te kategorije kojima su dokumenti pridruženi unutar neke zbirke. Prilikom izgradnje i testiranja modela potrebno je priložiti i listu ključnih fraza kako bi se omogućilo učenje i evaluacija. Dakle, ideja je stvoriti model koji aproksimira ljudsku sposobnost određivanja relevantnih fraza.

Algoritmu nerazumljivi parametri kao što su sadržaj, naslov i lista kategorija prevode se u listu ulaznih vektora. Vektori predstavljaju sve fraze iz teksta koje prolaze inicijalne filtre, a sastoje se od niza značajki koje je potrebno pomno odabrati jer su one jedina modelu dostupna informacija. Izlazna značajka može poprimiti dva stanja što upućuje na problem klasifikacije. Klasifikator određuje pripada li vektor skupu ključnih fraza. Uzevši u obzir mogućnost biranja broja fraza (K) na izlazu, postavlja se zahtjev da sustav vraća vjerojatnost pripadnosti vektora svakoj od klasa. Na temelju te informacije stvara se nova lista vektora. Lista je poredana prema većoj vjerojatnosti pripadnosti klasi ključnih fraza. Prvih K elemenata te liste predstavlja najrelevantnije automatski ekstrahirane fraze.

Faze izgradnje sustava za ekstrakciju mogu se podijeliti na: *predobradu*, *generiranje kandidata* i *izgradnju modela*. Za optimizaciju rada sustava koriste se različite metode evaluacije (opisane u poglavlju 5) kako bi se ustanovilo koji dijelovi sustava narušavaju odnosno pospješuju rad sustava. U slučaju praktične primjene najbolji dobiveni model koristi se za predlaganje relevantnih fraza korisniku.

Slika 4.1 prikazuje načelan rad sustava te sadrži kratke opise faza izgradnje.



Slika 4.1: Shematski prikaz rad sustava

4.1. Predobrada

Prije nego se pristupi bilo kakvim operacijama nad samim dokumentom potrebno je ukloniti sve suvišne informacije. KPEX odbacuje sve znakove koji nisu opisani sljedećim regularnim izrazom:

```
[a-zA-Z:,.!?;()čćžšđČĆŽŠĐ]
```

Dakle, prihvaćaju se isključivo velika i mala slova te frazni graničnici kojima se tekst dijeli na manje cjeline. Brojevi su odbačeni budući da se nedovoljno često pojavljuju među prihvatljivim ključnim frazama. Također, uklanjaju se sve višestruke oznake razmaka i novog reda te se zamjenjuju s jednom oznakom razmaka.

4.2. Generiranje kandidata

Nakon uklanjanja suvišnih znakova dokument se dijeli na skupove fraza uz pomoć sljedećeg izraza:

```
[:,.!?;()]
```

Ovisno o maksimalnom definiranom broju riječi, generiraju se sve kombinacije fraza unutar svakog od dijelova teksta koji smo prethodno podijelili fraznim graničnicima. KPEX dopušta maksimalno tri riječi po frazi. Na primjer, ako se radi o sljedećem nizu riječi:

predmet znanstvenog interesa

i ograničenju od tri riječi moguće je izvesti sljedeće kombinacije fraza:

predmet
znanstvenog
interesa
predmet znanstvenog
znanstvenog interesa
predmet znanstvenog interesa

Nad skupom fraza provodi se dodjeljivanje značajki što je detaljno opisano u sljedećoj cjelini.

Prije izgradnje modela potrebno je, ako je moguće, optimizirati primjere za učenje. Kako broj negativnih primjera u velikoj mjeri nadilazi pozitivne, primjenjuje se filtar

vrsta riječi. Filtar je izgrađen analizom raspodjele ključnih fraza po kombinacijama vrsta riječi.

Deset vršnih kombinacija vrsta riječi dobivenih analizom Hininog skupa dokumenata za treniranje:

N	AN	NN	NF	NAN	V	F	A	NXN	VN
1276	1272	490	141	129	126	113	108	101	92

gdje vrijedi:

N - imenica
A - pridjev
V - glagol
X - zaustavna riječi
F - vrsta riječi nije poznata

Naime, lista je dodatno smanjena pomoću evaluacije te je na kraju svedena na samo pet najzastupljenijih kombinacija. Imajući na umu da je imenica najučestalija vrsta riječi te činjenicu da lematizator nije u mogućnosti razvrstati sve riječi, osnovnom skupu pridružene su kombinacije FN,FNN,NFN,NNF,FFN,NFF. U tablicama 4.1 i 4.2 prikazana je raspodjela vektora po klasama prije i nakon filtriranja.

Tablica 4.1: Raspodjela vektora po klasama prije filtriranja na temelju vrste riječi

	Ključna fraza	Nije ključna fraza
Skup za treniranje	0,7%	99,3%
Skup za evaluaciju (unija označivača)	2,3%	97,7%

Tablica 4.2: Raspodjela vektora po klasama nakon filtriranja na temelju vrste riječi

	Ključna fraza	Nije ključna fraza
Skup za treniranje	3,7%	96,3%
Skup za evaluaciju (unija označivača)	11,9%	88,1%

Fraze s manje od tri znaka odbacuju se jer su vjerojatno nastale kao posljedica nepravilnosti u samom načinu zapisa teksta. U fazi evaluacije primjena tog kriterija pokazala je dobre rezultate. Odbacuju se i fraze koje sadrže osobne imenice, čime je poboljšana učinkovitost.
Također, u svrhu optimizacije, odlučeno je uklanjati duplikate odnosno fraze s istom

lematiziranom formom. Duplikatne fraze se uklanjaju iz skupa kandidata nakon što je izračunata učestalost njihovog pojavljivanja u dokumentu. Zadržana je samo ona fraza koja se prva pojavljuje kao predstavnik. Tim postupkom smanjena je složenost ulaznih podataka.

4.3. Značajke

4.3.1. Kandidat je u kategoriji

Značajka koja vrednuje pojavljivanje kandidata u nazivu kategorije. Provjera se temelji na uspoređivanju lematizirane forme i kombinacije oznaka vrsta riječi svih kategorija dokumenta s lematiziranom formom kandidata i njenim kombinacijama oznaka vrsta riječi. U slučaju podudaranja, značajki se dodjeljuje vrijednost 1, inače 0. Pretpostavlja se da će se model izgraditi u obliku koji preferira fraze koje se spominju u listi kategorija. Značajka je binarna.

4.3.2. Kandidat je u naslovu

Značajka koja vrednuje pojavljivanje kandidata u naslovu. Provjera se temelji na uspoređivanju lematizirane forme i kombinacije oznaka vrsta riječi svih fraza naslova s lematiziranom formom kandidata i njenim kombinacijama oznaka vrsta riječi. U slučaju podudaranja, značajki se dodjeljuje vrijednost 1, inače 0. Pretpostavlja se da će se model izgraditi u obliku koji preferira fraze koje se pojavljuju u naslovu. Značajka je binarna.

4.3.3. TFxIDF

Značajka koja vrednuje frazu prema učestalosti pojavljivanja u tekstu i korpusu. Relacija **TFxIDF** glasi:

$$TF \times IDF = \frac{freq(P, D)}{size(D)} \times \log_2 \frac{N}{df(C)} \tag{4.1}$$

gdje $freq(P, D)$ označava broj pojavljivanja fraze u dokumentu, $size(D)$ je broj riječi u dokumentu, $df(C)$ predstavlja broj dokumenata u kojima se fraza pojavljuje barem jednom i N je ukupan broj algoritmu dostupnih dokumenata.

4.3.4. Relativna prva pozicija lematizirane fraze

Relativna prva pozicija lematizirane fraze izračunava se kao minimalna vrijednost pozicije svih identičnih lematiziranih fraza u tekstu podjeljenja s brojem riječi u dokumentu. Izračunava se sljedećom relacijom:

$$first_R(C) = \frac{pos(C, D)}{size(D)} \tag{4.2}$$

gdje je $pos(C, D)$ redni broj prve riječi kandidata C u dokumentu D, dok je $size(D)$ broj riječi u dokumentu D. Značajka je predstavljena kao realan broj.

4.3.5. Broj riječi u frazi

Broj riječi kandidata. Temelji se na pretpostavci da će u graničnim slučajevima duljina fraze biti indikativna za odluku je li fraza ključna ili nije. Pretpostavimo da su modelu u fazi učenja predstavljena dva vektora s gotovo identičnim značajkama gdje se u prvom slučaju fraza koju vektor predstavlja sastoji od jedne riječi, dok druga sadrži dvije. Također, pretpostavimo da je drugi vektor označen kao ključna fraza, a prvi nije. Model će u tom slučaju davati prednost frazama s dvije riječi. Značajka je predstavljena kao cijeli broj.

4.3.6. Frekvencija najfrekventnije lematizirane riječi u frazi

Karakteristika dobivena izračunavanjem maksimalne vrijednosti frekvencije pojavljivanja pojedinih lematiziranih formi riječi (unutar cijelog teksta) u promatranoj frazi. Primjerice za frazu:

predmet znanstvenog interesa

frekvencije pojavljivanja pojedinih riječi u dokumentu su:

predmet 2
znanstven 2
interes 5

pa se značajka izračunava pomoću *max(2,2,5)*. Dakle, vrijednost je *5*. Značajka je izražena kao cijeli broj.

4.3.7. Minimalan broj znakova

Minimalan broj znakova u riječima analizirane fraze. Ova značajka trebala bi omogućiti modelu aproksimiranje donje prihvatljive granice minimalnog broja znakova riječi u frazi. Značajka je cijeli broj.

4.3.8. Prosječan broj znakova

Ova se vrijednost izračunava na razini same fraze kao prosjek znakova svih riječi. Značajka koja nije trebala proći u konačnu selekciju pokazala se kao korisna. Moguće je da uklanja ekstremne slučajeve. Značajka je predstavljena kao realan broj.

4.3.9. Fraza je ključna fraza

Jedina izlazna značajka. Određuje se usporedbom lematizirane forme kandidata sa svim unaprijed dodijeljenim ključnim frazama, također u lematiziranoj formi. Ako se kandidat nalazi u listi, značajki se dodjeljuje vrijednost 1, inače 0. Značajka je binarna.

4.3.10. Odbačene značajke

Iako se početna lista značajki nije mnogo mijenjala neke su značajke ipak odbačene u konačnoj verziji sustava. Radi se o *maksimalnom broj znakova među riječima fraze* i *poziciji fraze u tekstu*.

Maksimalan broj znakova odbačen je kao značajka u fazi evaluacije pošto je narušavao rad algoritma. Moguć razlog je u tome što ne postoji prihvatljiva aproksimacija gornje granice maksimalnog broja znakova riječi u frazi.

Značajka *pozicija fraze u tekstu* pokazala se suvišnom nakon uklanjanja duplikata gdje su zadržani samo kandidati koji se prvi pojavljuju u tekstu kao predstavnici fraze te su vrijednosti značajki *pozicija fraze u tekstu* i *prva pozicija lematizirane fraze* bile jednake.

4.4. Izgradnja modela

Filtrirani kandidati predaju se algoritmu za izgradnju modela u obliku vektora s gore navedenim značajkama. Cilj je izgraditi klasifikator koji aproksimira relaciju ulaznih i izlaznih značajki. Prije primjene algoritma nad značajkama se provodi diskretizacija, čime se vrijednosti značajki smještaju u ograničeni broj intervala.

Diskretizacija pospješuje rad sustava smanjivanjem složenosti tako što se različite kontinuirane vrijednosti koje padaju u isti interval tretiraju kao jedna vrijednost. Utjecaj diskretizacije na rezultate evaluacije te definicije različitih metoda diskretizacije opisani su u odjeljku 5.5. KPEX je evaluiran s dva klasifikatora: naivnim Bayesovim klasifikatorom i višeslojnim perceptronom.

4.4.1. Naivan Bayesov klasifikator

Naivan Bayesov klasifikator koristi pretpostavku da su značajke nezavisne te računa utjecaj vrijednosti pojedine ulazne značajke na izlazne značajke. Algoritam dodjeljuje vjerojatnost pripadnosti vektora svakoj od mogućih klasa. Vektor pripada klasi s najvećom vjerojatnošću. Naivan Bayesov klasifikator pokazao se vrlo uspješnim u rješavanju ove vrste problema.

4.4.2. Višeslojni perceptron

Višeslojni perceptron rijetko se koristi u sustavima za ekstrakciju ključnih fraza, no u našem se slučaju pokazao boljim od naivnog Bayesovog klasifikatora. Prilikom učenja višeslojnog perceptrona zaključeno je da nije potrebno raditi više od 100 iteracija jer je već na tako malom broju iteracija postignuta gotovo konstantna vrijednost kvadratne pogreške. U ovom radu neće se detaljnije obrađivati funkcioniranje višeslojnog perceptrona. Za izgradnju modela korišteni su parametri prikazani u tablici 4.3.

Tablica 4.3: Višeslojni perceptron s povratnom propagacijom pogreške (parametri za SNNS implementaciju[1])

ulazni sloj	broj neurona odgovara broju ulaznih značajki (8 neurona)
skriveni sloj	8 neurona
izlazni sloj	broj neurona odgovara broju izlaznih značajki (1 neuron)
broj iteracija	100
brzina učenja	0,1
minimalna kvadratna pogreška	0

[1] http://www.ra.cs.uni-tuebingen.de/SNNS/

5. Evaluacija i analiza rezultata

5.1. Svrha, nedostaci i definicija evaluacije

Evaluacija je način mjerenja učinkovitosti izvedbe algoritma strojnog učenja nad nekim skupom podataka. Svrha tog postupka jest optimizacija rezultata algoritma te odabir optimalnih značajki koje se predaju algoritmu za strojno učenje. Također, evaluacija omogućuje usporedbu rezultata srodnih radova (odjeljak 5.7). Usporedba često nije pouzdana i rezultati koji su brojčano bolji često se ne bi trebali uzimati kao takvi. Nepouzdanost je uvjetovana varijacijom skupova podataka pojedinih izvedbi. Evaluacija se temelji na različitim relacijama odnosno mjerama od kojih sustav KPEX koristi tri najzastupljenije kada govorimo o problemu ekstrakcije ključnih fraza. Korištene mjere opisane su u nastavku teksta.

Za evaluaciju su korištene mjere: *preciznost* (engl. *precision*), *odaziv* (engl. *recall*) i *F1*. Termin *broj pogodaka* označava broj podudaranja automatski generiranih fraza s frazama označivača.

Preciznost na korpusu $prec_{eval}$ izračunavamo na temelju relacije 5.1 kao *broj pogodaka* NP_{eval} u svim dokumentima skupa za evaluaciju podijeljenim s umnoškom broja zatraženih fraza K i broja dokumenata ND_{eval} u trenutnom skupu za evaluaciju.

$$prec_{eval} = \frac{NP_{eval}}{K \times ND_{eval}} \tag{5.1}$$

Prema relaciji 5.2 *odaziv na korpusu* $odaziv_{eval}$ dobiven je kao *broj pogodaka* NP_{eval} u svim dokumentima skupa za evaluaciju podijeljenim s ukupnim brojem fraza označivača u trenutnom skupu za evaluaciju $Npreassigned_{eval}$.

$$odaziv_{eval} = \frac{NP_{eval}}{Npreassigned_{eval}} \tag{5.2}$$

Mjera *F1* predstavlja najrelevantniju mjeru postupka evaluacije pošto dovodi *preciznost* i *odaziv* u protutežu. Ta je relacija neobično važna jer se vrijednost preciznosti može povisiti na štetu odaziva. Primjerice, povećanjem $Npreassigned_{eval}$

povećava se vrijednost NP_{eval} što će nad relativno manjim K dati veću vrijednost preciznosti $prec_{eval}$. S druge strane, $odaziv_{eval}$ će se povećanjem broja fraza označivača $Npreassigned_{eval}$ prema relaciji 5.2 smanjivati. Relacija 5.3 za mjeru $F1_{eval}$ sprječava dominaciju parametra $prec_{eval}$ odnosno $odaziv_{eval}$.

$$F1_{eval} = \frac{2 \times prec_{eval} \times odaziv_{eval}}{prec_{eval} + odaziv_{eval}} \tag{5.3}$$

Prvobitne postavke sustava KPEX prilikom evaluacije

Ako nije posebno navedeno za evaluaciju se koristi:
- diskretizacija MDL koja je opisana u poglavlju 5.5
- unija skupova ključnih fraza svih označivača
- značajke navedene u poglavlju 4.3
- filtar temeljen na najboljih pet kombinacijama vrsta riječi proširen s dodatnim kombinacijama vrsta riječi koje lematizator nije razriješio istaknutim u poglavlju 4.2.

5.2. Unija označivača

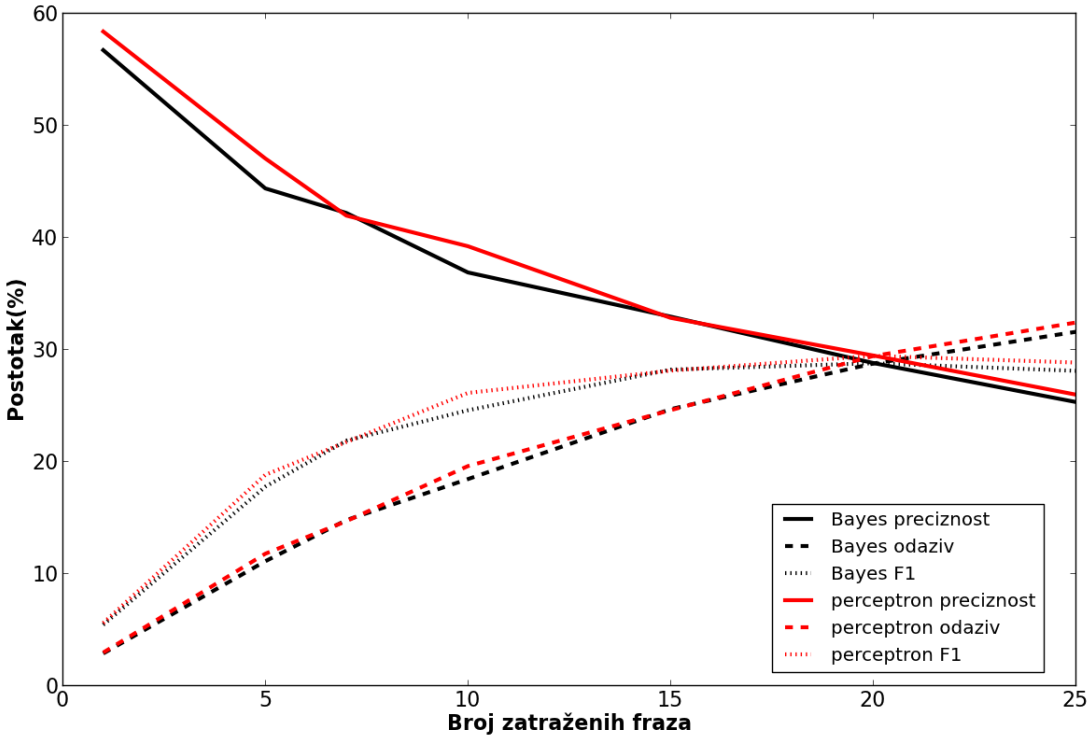
Pod *unijom označivača* podrazumijeva se skup unaprijed dodijeljenih ključnih fraza dobiven unijom skupova fraza svih osam označivača. U ovom slučaju nailazimo na veliki broj unaprijed dodijeljenih ključnih fraza po dokumentu te su stoga mjere preciznost, odaziv i F1 vrlo visoke. Rezultati dobiveni ovim načinom evaluacije mogli bi se smatrati pretjerano dobrim te je stoga potrebno objasniti značenje mjere F1. Spomenuta mjera uzima u obzir preciznost i odaziv te ih stavlja u relaciju protuteže. Tako odaziv prigušuje značaj preciznosti i obrnuto. Također, ova evaluacija pokazuje da je sustav u mogućnosti odabrati velik broj fraza koje su barem jednom označivaču prihvatljive kao ključne fraze.

Tablica 5.1: Karakteristike dokumenata za uniju označivača

Ukupno dokumenata	60
Broj KF po dokumentu	20,03
Broj kandidata po dokumentu	99,35
Broj pozitivnih kandidata po dokumentu	11,85

Tablica 5.2: Rezultati evaluacije za uniju označivača

K	Naivan Bayesov klasifikator			Višeslojni perceptron		
	Preciznost (%)	Odaziv (%)	F1 (%)	Preciznost (%)	Odaziv (%)	F1 (%)
1	56,67	2,83	5,39	58,33	2,91	5,55
5	44,33	11,06	17,71	47,00	11,73	18,77
7	42,14	14,73	21,82	41,90	14,64	21,70
10	36,83	18,39	24,53	39,17	19,55	26,08
15	32,89	24,63	28,16	32,78	24,54	28,07
20	28,75	28,70	28,73	29,42	29,37	29,39
25	25,27	31,53	28,05	25,93	32,36	28,79



Slika 5.1: Grafički prikaz rezultata evaluacije za uniju označivača

5.3. Prosjek označivača

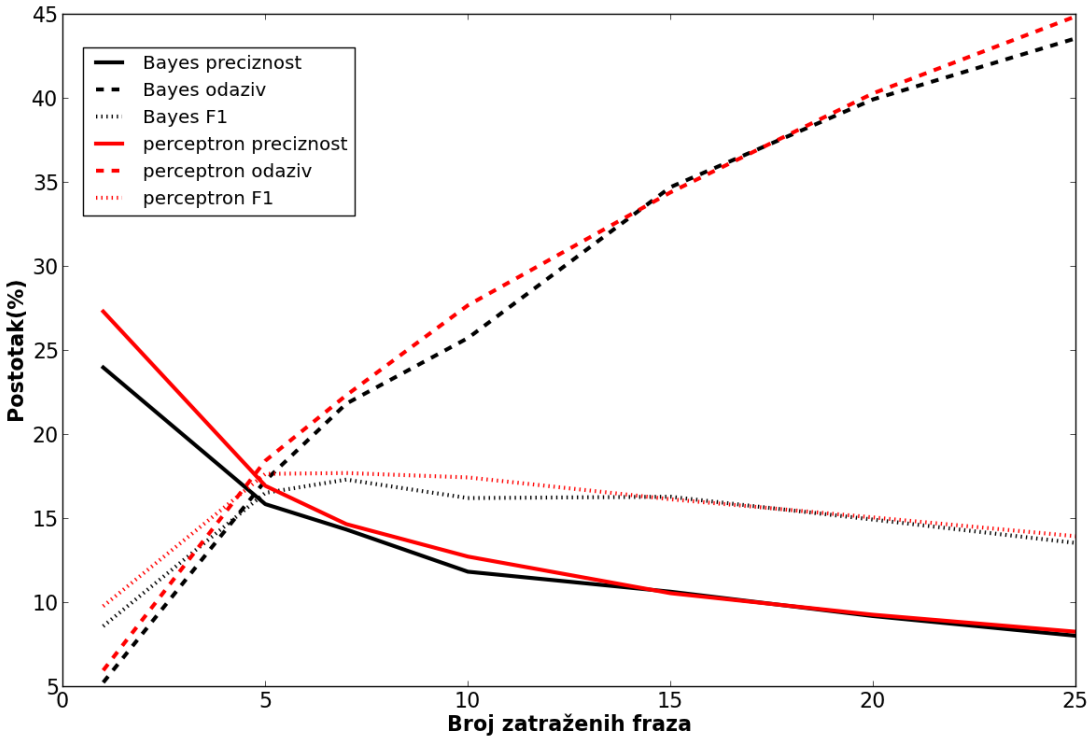
U metodi koju smo nazvali *prosjek označivača* evaluacija se izvodi posebno za svakog od označivača (8) nad svim dokumentima (60) te se računa prosječna vrijednost svih pojedinih slučajeva. Brojimo, dakle, 480 dokumenata. Od svih metoda evaluacije, upravo ova najviše odgovara klasičnom načinu ispitivanja rada sustava. Prema tome, u slučaju usporedbe s drugim radovima, u obzir bi se trebala uzimati ova evaluacija.

Tablica 5.3: Karakteristike dokumenata za prosjek označivača

Ukupno dokumenata	480
Broj KF po dokumentu	4,59
Broj kandidata po dokumentu	99,35
Broj pozitivnih kandidata po dokumentu	3,30

Tablica 5.4: Rezultati evaluacije za prosjek označivača

	Naivan Bayesov klasifikator			Višeslojni perceptron		
K	Prec. (%)	Odaziv (%)	F1 (%)	Prec. (%)	Odaziv (%)	F1 (%)
1	23,96	5,22	8,57	27,29	5,94	9,76
5	15,83	17,23	16,50	16,92	18,41	17,63
7	14,32	21,81	17,29	14,64	22,31	17,68
10	11,81	25,71	16,19	12,71	27,66	17,42
15	10,62	34,69	16,27	10,53	34,38	16,12
20	9,17	39,91	14,91	9,25	40,27	15,04
25	8,00	43,54	13,52	8,24	44,85	13,92



Slika 5.2: Grafički prikaz rezultata evaluacije za prosjek označivača

5.4. Varijacija podskupova fraza prema označivačima

Slijede rezultati evaluacije dobiveni nad skupovima ključnih fraza oko kojih su se složila barem dva, tri, odnosno četiri označivača. Na kraju ovog odjeljka priložen je graf koji uspoređuje rezultate za prosjek, uniju te slaganje N označivača za slučaj višeslojnog perceptrona. Prikazan je samo višeslojni perceptron jer daje bolje rezultate od naivnog Bayesovog klasifikatora, a takav prikaz smanjuje zasićenost grafa.

5.4.1. Evaluacija nad skupovima ključnih fraza oko kojih se slažu barem dva označivača

Tablica 5.5: Karakteristike dokumenata za skupove ključnih fraza oko kojih se slažu barem dva označivača

Ukupno dokumenata	60
Broj KF po dokumentu	9,23
Broj kandidata po dokumentu	99,35
Broj pozitivnih kandidata po dokumentu	5,88

Tablica 5.6: Rezultati evaluacije za skupove ključnih fraza oko kojih se slažu barem dva označivača

K	Naivan Bayesov klasifikator			Višeslojni perceptron		
	Preciznost (%)	Odaziv (%)	F1 (%)	Preciznost (%)	Odaziv (%)	F1 (%)
1	41,67	4,51	8,14	41,67	4,51	8,14
5	27,33	14,80	19,20	28,33	15,34	19,91
7	24,76	18,77	21,36	25,71	19,49	22,18
10	20,67	22,38	21,49	23,33	25,27	24,26
15	18,44	29,96	22,83	18,67	30,32	23,11
20	15,75	34,12	21,55	15,92	34,48	21,78
25	13,87	37,55	20,25	14,60	39,53	21,32

5.4.2. Evaluacija nad skupovima ključnih fraza oko kojih se slažu barem tri označivača

Tablica 5.7: Karakteristike dokumenata za skupove ključnih fraza oko kojih se slažu barem tri označivača

Ukupno dokumenata	60
Broj KF po dokumentu	5,62
Broj kandidata po dokumentu	99,35
Broj pozitivnih kandidata po dokumentu	3,47

Tablica 5.8: Rezultati evaluacije za skupove ključnih fraza oko kojih se slažu barem tri označivača

K	Naivan Bayesov klasifikator			Višeslojni perceptron		
	Preciznost (%)	Odaziv (%)	F1 (%)	Preciznost (%)	Odaziv (%)	F1 (%)
1	33,33	5,93	10,08	33,33	5,93	10,08
5	21,00	18,69	19,78	21,33	18,99	20,09
7	18,10	22,55	20,08	17,38	21,66	19,29
10	14,50	25,82	18,57	16,17	28,78	20,70
15	12,56	33,53	18,27	13,33	35,61	19,40
20	10,67	37,98	16,66	11,00	39,17	17,18
25	9,33	41,54	15,24	9,53	42,43	15,57

5.4.3. Evaluacija nad skupovima ključnih fraza oko kojih se slažu barem četiri označivača

Tablica 5.9: Karakteristike dokumenata za skupove ključnih fraza oko kojih se slažu barem četiri označivača

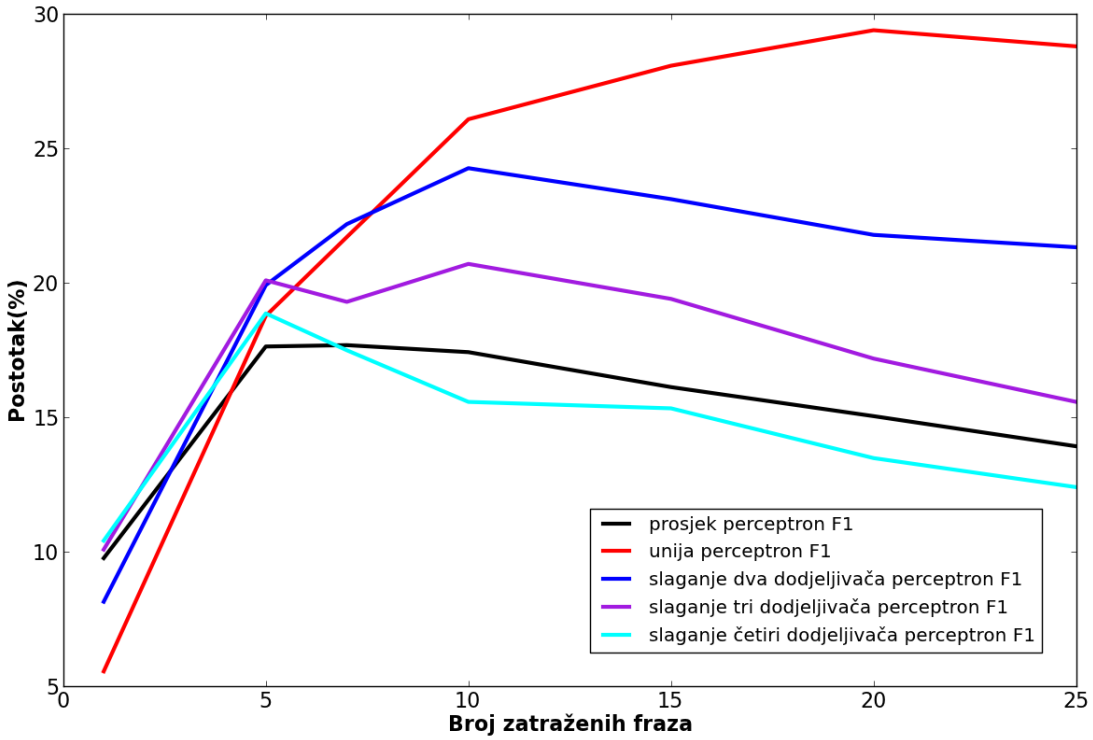
Ukupno dokumenata	60
Broj KF po dokumentu	3,48
Broj kandidata po dokumentu	99,35
Broj pozitivnih kandidata po dokumentu	2,43

Tablica 5.10: Rezultati evaluacije za skupove ključnih fraza oko kojih se slažu barem četiri označivača

K	Naivan Bayesov klasifikator			Višeslojni perceptron		
	Preciznost (%)	Odaziv (%)	F1 (%)	Preciznost (%)	Odaziv (%)	F1 (%)
1	20,00	5,74	8,92	23,33	6,70	10,41
5	13,67	19,62	16,11	16,00	22,97	18,86
7	12,14	24,40	16,22	13,10	26,32	17,49
10	9,33	26,79	13,84	10,50	30,14	15,57
15	8,78	37,80	14,25	9,44	40,67	15,33
20	7,75	44,50	13,20	7,92	45,45	13,48
25	6,73	48,33	11,82	7,07	50,72	12,40

5.4.4. Analiza evaluacije nad podskupovima ključnih fraza

Nakon vrlo temeljite evaluacije različitih podskupova dodijeljenih ključnih fraza, sastavljen je graf na slici 5.3 koji omogućuje praćenje ovisnosti F1 mjere u odnosu na različite podskupove fraza. Lako je zaključiti da površina ispod krivulje raste s veličinom skupa unaprijed dodijeljenih fraza što ukazuje na očekivanu zavisnost vrijednosti mjere F1 o veličini skupa unaprijed dodijeljenih fraza. Kao još jednu zanimljivost treba spomenuti da maksimalna vrijednost mjere F1 uglavnom teži prosječnoj vrijednosti unaprijed dodijeljenih ključnih fraza. U slučaju potrebe za usporedbom ovog rada s ostalim radovima, predlaže se razmatranje rezultata prosjeka svih označivača.



Slika 5.3: Usporedba rezultata za prosjek, uniju, te slaganje dva, tri i četiri označivača za slučaj višeslojnog perceptrona

5.5. Varijacija metoda diskretizacije

Prilikom evaluacije korištene su tri različite metode diskretizacije sa svrhom da se najbolja koristi u konačnoj verziji sustava. Korištene metode su: *Minimum Description Length (MDL)*, *Equal Width (EQW)* i *Equal Frequency (EQF)*.

Minimum Description Length (MDL) - diskretizacija koja se temelji na minimizaciji entropije unutar pojedinih diskretnih intervala čime se postiže specifičan i odgovarajući broj intervala za zadani skup kontinuiranih vrijednosti.

Equal Width (EQW) - ovisno o zadanom parametru *N* ova metoda diskretizacije dijeli skup kontinuiranih vrijednosti na *N* diskretnih intervala podjednake duljine. Prilikom usporedbe s drugim metodama korištena je vrijednost *N=200*.

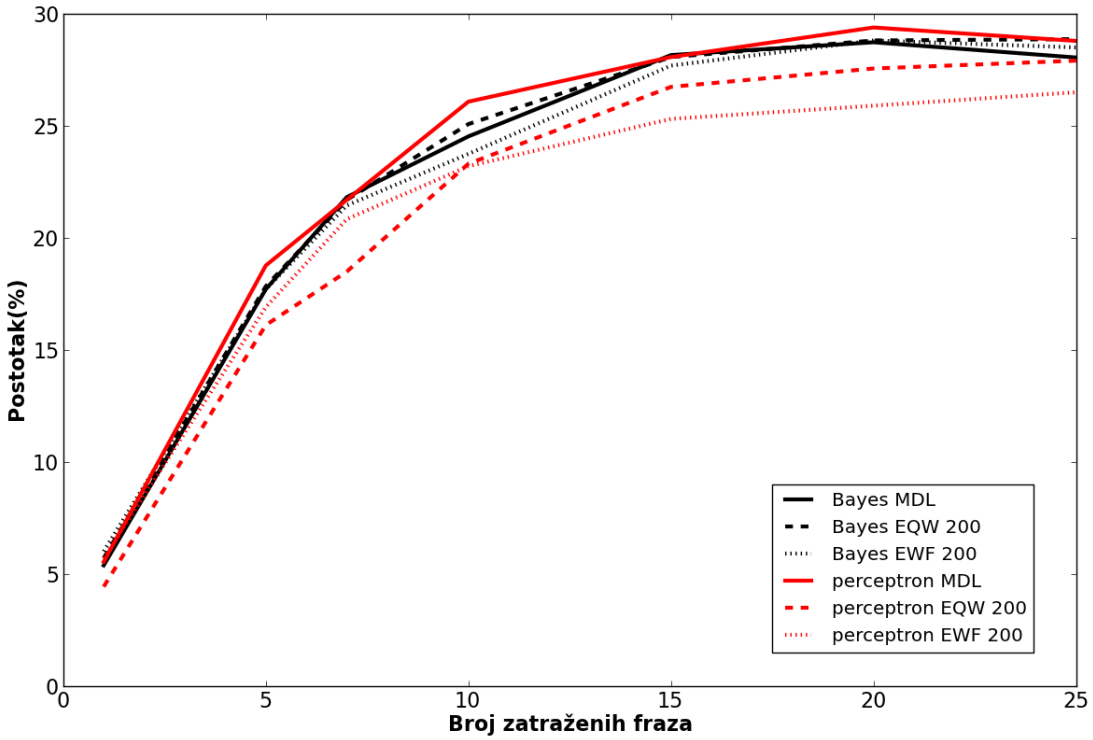
Equal Frequency (EQF) - ovisno o zadanom parametru *N*, ova metoda diskretizacije dijeli skup kontinuiranih vrijednosti na *N* diskretnih intervala u kojima se nalazi podjednak broj elemenata iz kontinuiranog skupa. Prilikom usporedbe s drugim metodama korištena je vrijednost *N=200*.

Primjenom navedenih metoda s naivnim Bayesovim klasifikatorom i višeslojnim perceptronom zaključeno je da MDL i višeslojni perceptron daju najbolje rezultate. Ostale kombinacije diskretizacije i algoritama za učenje također daju prihvatljive rezul-

tate, što je moguće vidjeti u tablici 5.11 te na slici 5.4.

Tablica 5.11: Tablični prikaz vrijednosti F1 za varijaciju diskretizacije

K	Naivan Bayesov klasifikator			Višeslojni perceptron		
	MDL	EQW 200	EQF 200	MDL	EQW 200	EQF 200
1	5,39	5,71	6,02	5,55	4,44	5,71
5	17,71	17,84	17,71	18,77	16,11	16,91
7	21,82	21,70	21,45	21,70	18,50	20,84
10	24,53	25,08	23,75	26,08	23,31	23,20
15	28,16	28,07	27,69	28,07	26,74	25,31
20	28,73	28,81	28,81	29,39	27,56	25,90
25	28,05	28,87	28,50	28,79	27,91	26,50



Slika 5.4: Grafički prikaz F1 vrijednosti za varijaciju diskretizacije

5.6. Optimizacija

5.6.1. Filtriranje na temelju vrste riječi

Filtriranje na temelju vrste riječi veoma je bitan dio procesa u izgradnji sustava za automatsku ekstrakciju ključnih fraza. Analizom učestalosti pojavljivanja svih kombinacija vrsta riječi nad ključnim frazama dokumenata za treniranje, moguće je zaključiti koje su kombinacije vrsta riječi najučestalije. Nakon sastavljanja poredane liste kombinacija potrebno je omogućiti ugađanje parametra N koji predstavlja N najučestalijih kombinacija (prvih N elemenata poredane liste).

Nadalje, evaluacijom je potrebno pronaći optimalnu vrijednost N. Optimalna vrijednost je ona koja odbacuje dovoljan broj lažno pozitivnih i lažno negativnih kandidata i pritom zadržava dovoljan broj istinito pozitivnih i istinito negativnih kandidata. Također, ovim se postupkom uravnotežuje nejednakost raspodjele kandidata po klasama (negativnih kandidata znatno je više nego pozitivnih) i uvodi se u određenoj mjeri leksičko znanje za razliku od ostalih značajki koje su isključivo statističke prirode.

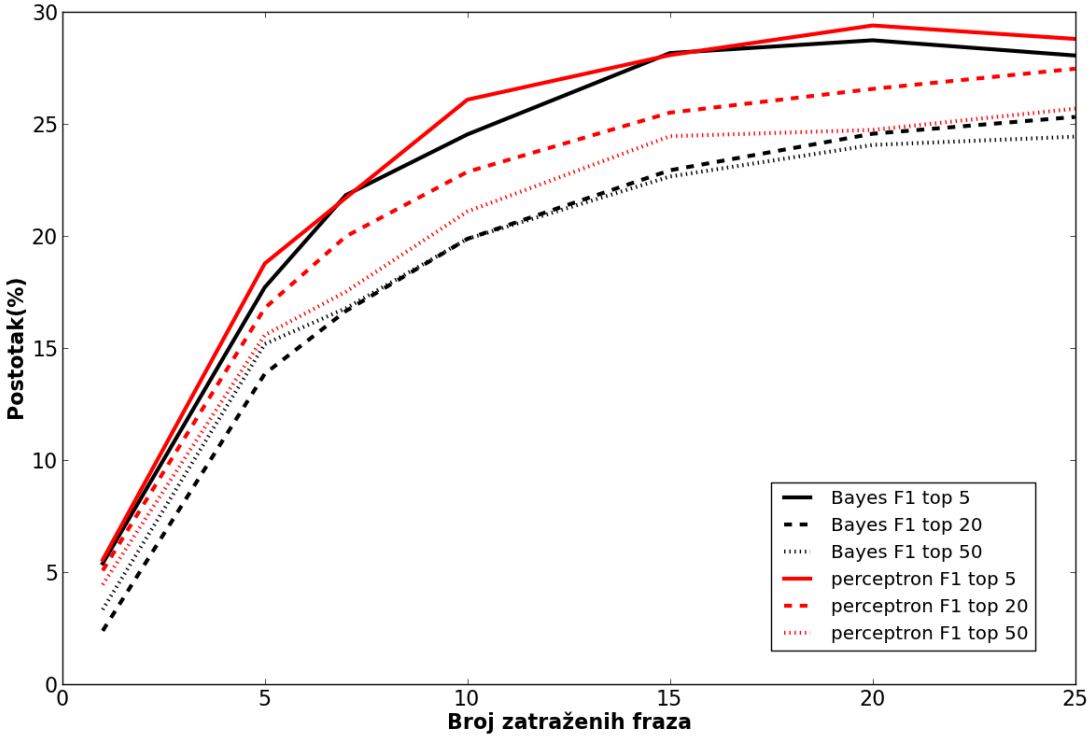
U tablici 5.12 prikazani su rezultati za 20 vršnih kombinacija, tablica 5.13 sadrži rezultate evaluacije za 50 vršnih kombinacija, a na slici nalazi se grafički prikaz 5.5 na kojem su navedeni rezultati uspoređeni s optimalnom vrijednošću parametra N (N=5).

Tablica 5.12: Varijacija podskupova kombinacija vrsta riječi za uniju označivača (20 najboljih)

K	Naivan Bayesov klasifikator			Višeslojni perceptron		
	Preciznost (%)	Odaziv (%)	F1 (%)	Preciznost (%)	Odaziv (%)	F1 (%)
1	25,00	1,25	2,38	53,33	2,66	5,07
5	34,67	8,65	13,85	42,00	10,48	16,78
7	32,14	11,23	16,65	38,57	13,48	19,98
10	29,83	14,89	19,87	34,33	17,14	22,86
15	26,78	20,05	22,93	29,78	22,30	25,50
20	24,58	24,54	24,56	26,58	26,54	26,56
25	22,80	28,45	25,31	24,73	30,87	27,46

Tablica 5.13: Varijacija podskupova kombinacija vrsta riječi za uniju označivača (50 najboljih)

K	Naivan Bayesov klasifikator			Višeslojni perceptron		
	Preciznost (%)	Odaziv (%)	F1 (%)	Preciznost (%)	Odaziv (%)	F1 (%)
1	35,00	1,75	3,33	46,67	2,33	4,44
5	38,00	9,48	15,18	39,00	9,73	15,58
7	32,38	11,31	16,77	33,81	11,81	17,51
10	29,83	14,89	19,87	31,67	15,81	21,09
15	26,44	19,80	22,65	28,56	21,38	24,45
20	24,08	24,04	24,06	24,75	24,71	24,73
25	22,00	27,45	24,43	23,13	28,87	25,68



Slika 5.5: Grafički prikaz rezultata evaluacije za varijacija podskupova kombinacija vrsta riječi uniju označivača

5.7. Rezultati srodnih radova

U ovom odjeljku predstavljeni su rezultati srodnih radova te su uz njih priloženi podaci o okolini u kojima je provođena njihova evaluacija. Rezultate srodnih radova nije moguće pouzdano usporediti te neke od njih proglasiti boljima od drugih. Razlog nemogućnosti provođenja takve usporedbe navodimo na početku poglavlja 4. Ipak, rezultati srodnih radova priloženi su kako bi se stekao dojam okoline u kojoj su provođena mjerenja sličnih radova te kako bi se dobio uvid u raspon vrijednosti mjera za evaluaciju.

5.7.1. Ahel et al. (2009)

Evaluirano je 200 dokumenata s prosječno 6,5 ključnih fraza te 370 kandidata po dokumentu. Učinkovitost algoritma mjerena je kao prosječna vrijednost desetorostruke unakrsne provjere.

Tablica 5.14: Rezultati evaluacije za rad Ahel et al. (2009)

K	Preciznost (%)	Odaziv (%)	F1 (%)
1	**22,00**	3,40	5,90
5	x	x	x
7	13,40	14,50	13,90
10	12,50	19,30	**15,10**
15	10,40	**24,10**	14,50

5.7.2. Sarkar et al. (2010)

Validacija ovog algoritma izvedena je na principu trostruke unakrsne provjere 150 dokumenata, od kojih je 100 korišteno za treniranje, a 50 za testiranje.

Tablica 5.15: Rezultati evaluacije za rad Sarkar et al. (2010)

K	Preciznost (%)	Odaziv (%)	F1 (%)
1	x	x	x
5	**34,00**	35,00	**34,50**
7	x	x	x
10	22,00	46,00	29,80
15	17,00	**51,00**	25,50

5.7.3. Witten et al. (1998)

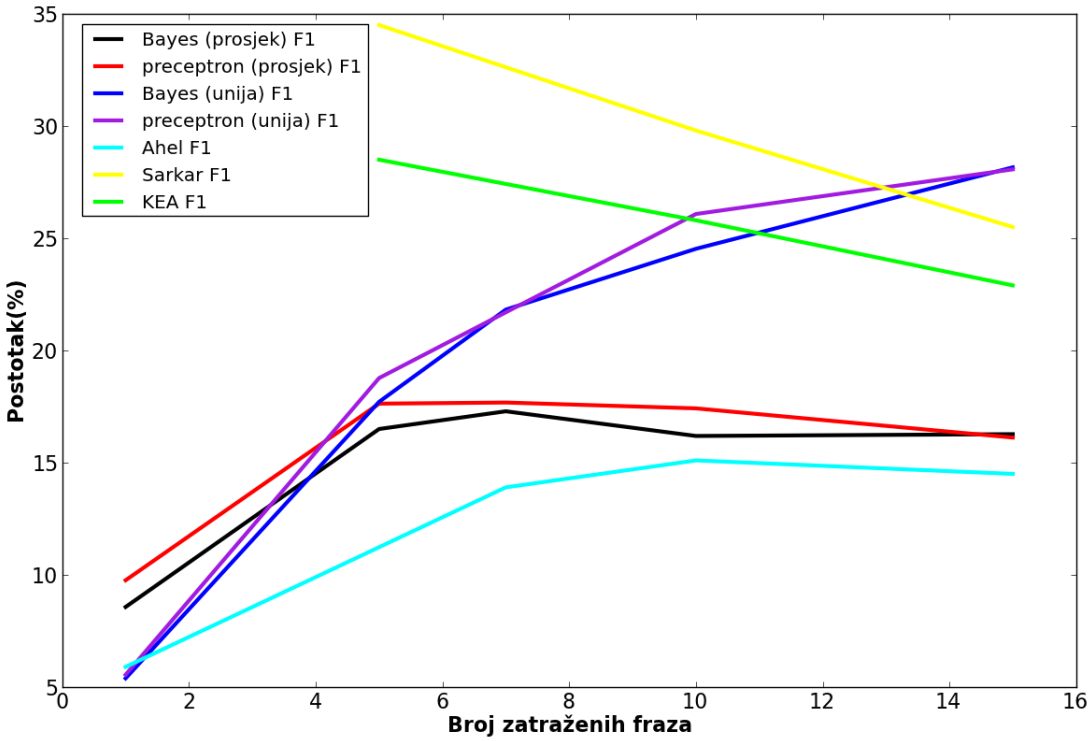
Rezultati evaluacije ovog algoritma navedeni su u radu Sarkar et al. (2010), a dobiveni su u istim uvjetima kao i gore navedeni.

Tablica 5.16: Rezultati evaluacije za rad Witten et al. (1998)

K	Preciznost (%)	Odaziv (%)	F1 (%)
1	x	x	x
5	**28,00**	29,00	**28,50**
7	x	x	x
10	19,00	40,00	25,80
15	15,00	**48,00**	22,90

5.7.4. Analiza rezultata

Na slici 5.6 uspoređeni su rezultati evaluacije srodnih radova s rezultatima sustava KPEX koristeći mjeru F1. Treba napomenuti da rezultati s većom vrijednosti mjere F1 ne predstavljaju nužno bolje rezultate. Primjerice, metode višeslojni perceptron i naivan Bayesov klasifikator za skup *"unija označivača"* te vrijednost *K=15* daju bolje rezultate od ostalih radova (uz jednake uvjete) gdje su ti rezultati opravdani velikim brojem unaprijed dodijeljenih ključnih fraza specifičnim za slučaj skupa *"unija označivača"*. Također je zanimljivo primjetiti da vrijednost mjere F1 za radove Witten et al. (1998) i Sarkar et al. (2010) pada s porastom broja zatraženih ključnih fraza *K* što je pokazatelj da ti radovi opisuju sustave koji dobro rade sa malim brojem zatraženih ključnih fraza te da je njihova evaluacija izvođena nad malim brojem ključnih fraza po dokumentu. U slučaju ostalih radova ovisnost parametra F1 o broju zatraženih fraza ne može se opisati kao monotona funkcija.



Slika 5.6: Usporedba s rezultatima srodnih radova

5.8. Uklanjanje zavisnosti o kategoriji i korpusu

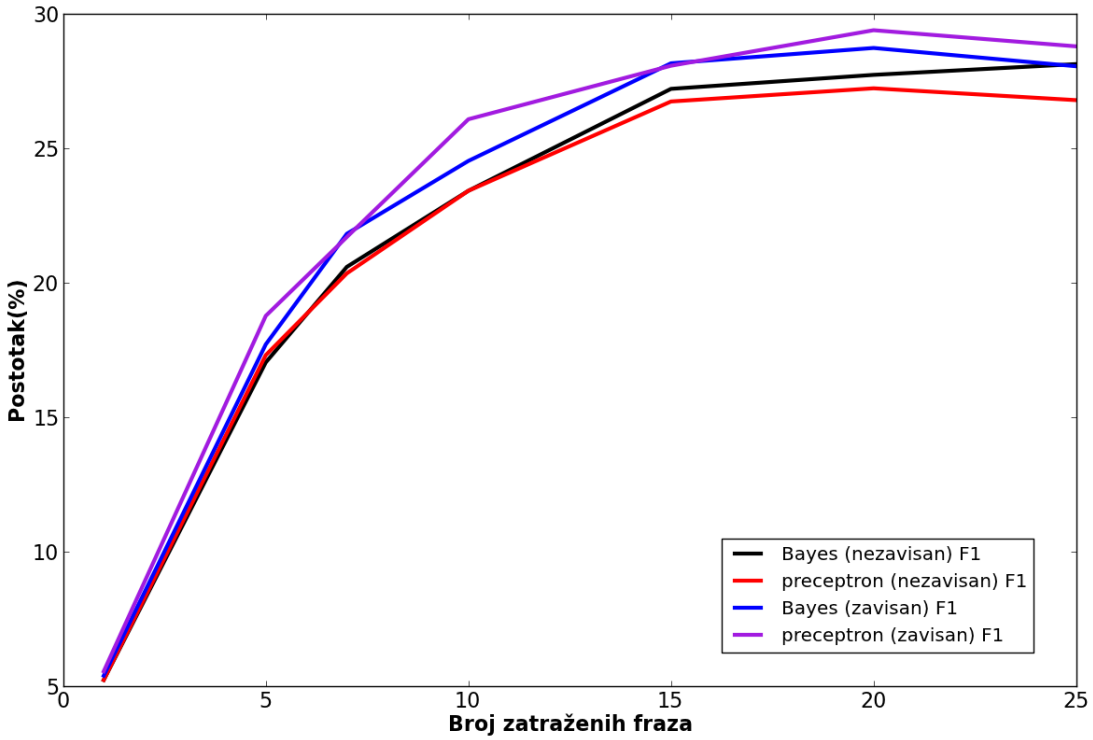
Budući da je KPEX moguće koristiti i kao alat za dodjelu ključnih fraza izoliranim dokumentima (ne pripadaju nekom korpusu), koji ne posjeduju informaciju o kategoriji, provedena je i evaluacija značajki koje ne sadrže tu zavisnost.

Značajka *TFxIDF* svedena je na *TF* jer *IDF* nosi zavisnost o korpusu. Uklonjena je i značajka *kandidat je u kategoriji*, zbog pretpostavke da ulazni dokumenti ne sadrže te podatke.

Evaluacija je provedena nad skupom *"unija označivača"* i uspoređena je s rezultatima evaluacije provedene nad primarnim značajkama. Temeljem tablice 5.17 i grafa na slici 5.7 zaključujemo da karakteristike sustava nisu značajno narušene u odnosu na primarne te je prihvatljivo koristiti sustav u slučaju kada dokumenti ne pripadaju nikakvom korpusu i nemaju dodijeljenu kategoriju.

Tablica 5.17: Uklanjanje zavisnosti o kategoriji i korpusu

K	Naivan Bayesov klasifikator			Višeslojni perceptron		
	Preciznost (%)	Odaziv (%)	F1 (%)	Preciznost (%)	Odaziv (%)	F1 (%)
1	55,00	2,75	5,23	55,00	2,75	5,23
5	42,67	10,65	17,04	43,33	10,82	17,31
7	39,76	13,89	20,59	39,29	13,73	20,35
10	35,17	17,55	23,42	35,17	17,55	23,42
15	31,78	23,79	27,21	31,22	23,38	26,74
20	27,75	27,70	27,73	27,25	27,20	27,23
25	25,33	31,61	28,13	24,13	30,12	26,79



Slika 5.7: Usporedba rezultata nakon uklanjanja zavisnosti o kategoriji i korpusu unija označivača

5.9. Primjeri rada sustava

U tablicama 5.18 i 5.19 prikazan je primjer dobre i loše ekstrakcije sustava za skup *"unija označivača"* i *K=10*. Fraze su navedene u lematiziranoj formi, a one koje se nalaze u oba skupa (unaprijed i automatski dodijeljene), istaknute su masnim slovima.

Zanimljivo je da frazu *"adriatica.net"* sustav eliminira još u fazi generiranja kandidata odnosno prilikom diobe teksta fraznim graničnicima. Kako bi sustav prihvatio frazu *"adriatica.net"* kao kandidata za konačnu listu fraza, trebali bismo isključiti znak "." iz skupa fraznih graničnika. No time bismo narušili rad sustava. Drugo rješenje je da oznaku "." zamijenimo s ". " u skupu fraznih graničnika, pri čemu uzimamo u obzir da su tekstovi ispravno formatirani i da se nakon točke nalazi razmak. Ipak, to često nije slučaj te bismo zbog takve pretpostavke narušili učinak algoritma. Ovo je jedan od primjera u kojima se odričemo pogotka jedne fraze kako ne bismo narušili ostale pogotke.

Tablica 5.18: Primjer dobre ekstrakcije za K=10

unaprijed dodijeljene ključne fraze (lematizirana forma)	**turistički agencija**, dopisan birotehnički škola, hit, **turizam**, karijera u turizam, **pjevački karijera**, turistički agencija adriatica.net, **korporativan poslovanje**, **karijera**, **prodajan predstavnik**, pjevačica
automatski dodijeljene ključne fraze (lematizirana forma)	**karijera**, **turistički agencija**, **turizam**, **korporativan poslovanje**, **prodajan predstavnik**, pjevačica rent, rent končić, poslovanje turistički, **pjevački karijera**, birotehnički škola
broj unaprijed dodijeljenih ključnih fraza	11
broj zatraženih ključnih fraza K	10
broj pogodaka	6

Tablica 5.19: Primjer loše ekstrakcije za K=10

unaprijed dodijeljene ključne fraze (lematizirana forma)	ured, predsjednik odbor za izbor i imenovanje, regionalizacija, najam, prostor za udruga, napustiti, prostor, klub nezavisan zastupnik, predsjednik odbor, tajništvo sabor, **parlamentaran većina**, isključen iz hdza, udruga
automatski dodijeljene ključne fraze (lematizirana forma)	otimanje zastupnik, špelunku dokidanje, prostran ured, **parlamentaran većina**, nezavisan zastupnik, većina otimanje, odgovarajući ured, nekretnina, gornjogradski nekretnina, slijep hodnik
broj unaprijed dodijeljenih ključnih fraza	13
broj zatraženih ključnih fraza K	10
broj pogodaka	1

6. Programska izvedba

6.1. Programski alati

6.1.1. Python

Python je programski jezik koji podržava objektno orijentirano, strukturno i aspektno orijentirano programiranje. Sintaksa jezika diktira uredno pisanje koda, a kako se relativno brzo usvaja, sve se više koristi za edukaciju početnika. Dolazi s velikom bibliotekom, a zbog velike zastupljenosti u *open source* zajednici dodatno je obogaćen mnogim korisnim paketima.

Sustav KPEX izgrađen je u Pythonu zbog lake prenosivosti među platformama i njegove jednostavnosti. Sustav koristi samo jedan vanjski paket - *Orange*[1]. Paket služi za implementaciju strojnog učenja.

6.1.2. Orange paket (Python)

Orange je Python paket koji nudi velik broj operacija za implementaciju strojnog učenja. U samom projektu korišten je za diskretizaciju ulaznih podataka i evaluaciju pomoću naivnog Bayesovog klasifikatora. Implementacija višeslojnog perceptrona ostvarena je pomoću *OrangeSNNS*[2] modula, koji omogućuje tok podataka između *Orange* modula i *SNNS*-a (Stuttgart Neural Network Simulator[3]).

6.1.3. Git

Git je distribuirani sustav za kontrolu revizija. Koristi se za verzioniranje koda, njegovo lakše dijeljenje te lakše razumijevanje tijeka izgradnje, čime se pojednostavljuje pridruživanje novih članova postojećem projektu.

[1] http://orange.biolab.si/

[2] http://ax5.com/antonio/orangesnns/

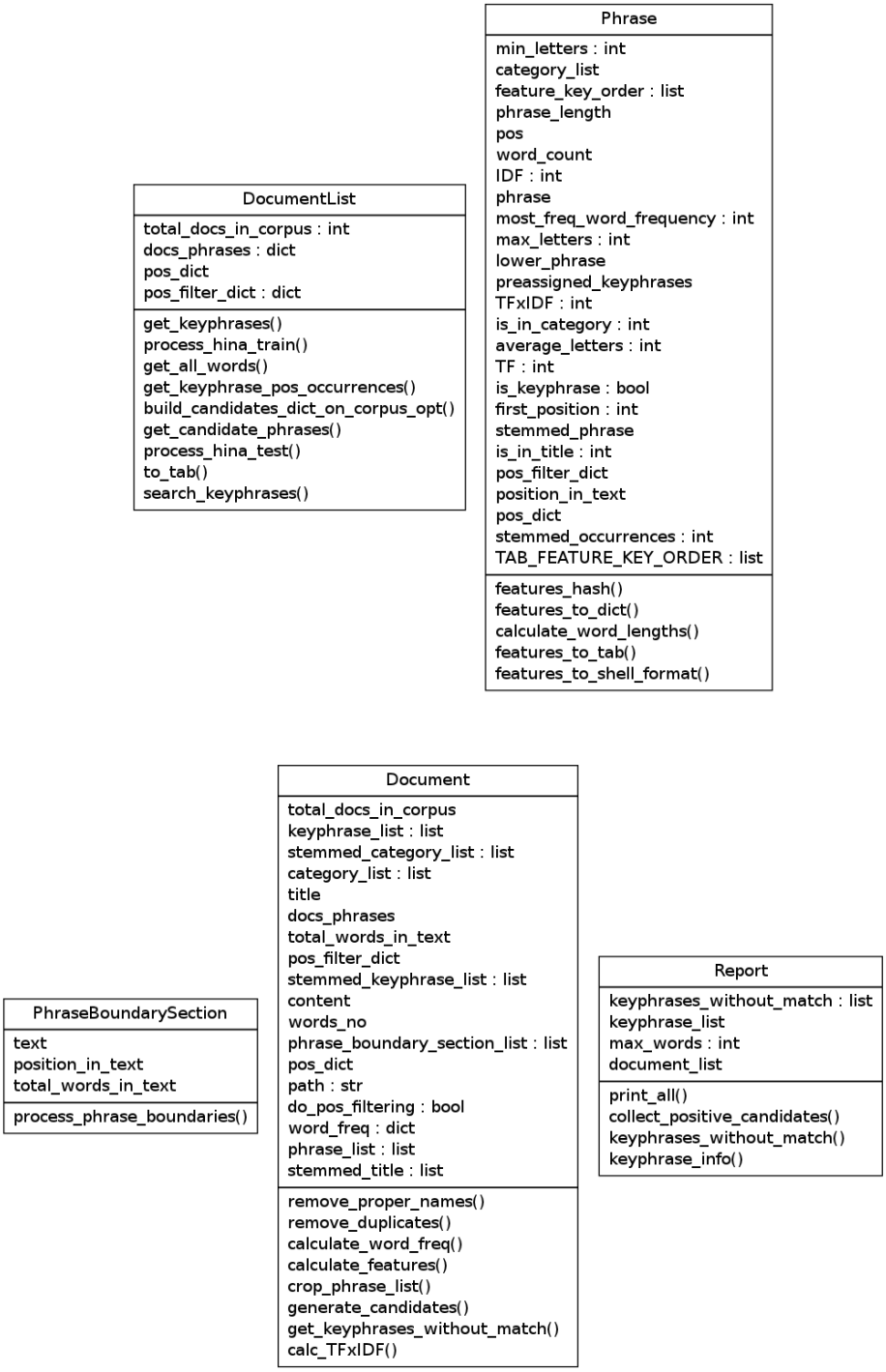
[3] http://www.ra.cs.uni-tuebingen.de/SNNS/

6.2. Organizacija koda

Korištenjem objektno orjentiranih značajki Pythona kod je organiziran u nekoliko klasa. Ulazni dokument preslikava se u klasu *Document*, gdje se posredovanjem *PhraseBoundarySection* generiraju objekti klase *Phrase*. Konačno, svi se objekti *Document* spremaju u objekt klase *DocumentList*. Uz opisane strukture koriste se još neke općenite pomoćne metode. Slika 6.1 sadrži grafički prikaz klasa iz glavnog modula kpex.core.document s njihovim varijablama i metodama.

Za svaku od klasa definirane su metode kao akcije koje se obavljaju na trenutnoj razini apstrakcije. Pa tako *Phrase* izračunava dio značajki (koji je moguće izračunati na toj razini), dok *Document* dodjeljuje one značajke koje se računaju na razini dokumenta.

Kod sadrži neke specifičnosti koje treba ukloniti ukoliko će se sustav koristiti u produkcijskoj okolini. Jedna od specifičnosti jest uska vezanost uz tekstni format dobiven prevođenjem Hininih XML dokumenata. Jedno od rješenja bilo bi korištenje relacijskih baza podataka za definiranje strukture ulaznih podataka.



Slika 6.1: Klase iz modula kpex.core.document s varijablama i metodama

7. Zaključak

Automatska ekstrakcija ključnih fraza predstavlja moderan pristup rješavanju problema vezivanja ključnih fraza uz dokumente u kojima autori nisu zadovoljili taj zahtjev. Ključne fraze omogućavaju brzo i ciljano dohvaćanje dokumenata iz zbirki, a glavna im je uloga izgradnja raznih informacijskih sustava (pretraživači, sustavi za kategorizaciju dokumenata) koji pojednostavljuju dohvaćanje dokumenata te štede korisnikovo vrijeme ubrzavajući pretragu.

Sustav KPEX zamišljen je i izgrađen kao testna okolina za ovaj istraživački rad i kao takav predstavlja jednostavan način ostvarivanja sustava za automatsku ekstrakciju ključnih fraza. Osnovna funkcionalnost sustava KPEX razvijana je prema uzoru na ostale srodne radove (Witten et al. (1998), Ahel et al. (2009), Sarkar et al. (2010)). Ovaj rad oslanjao se na rezultate evaluacije kako bi se postiglo optimalno funkcioniranje algoritma. Konačno, rezultati ovog eksperimentalnog rada mjerljivi su s rezultatima srodnih, objavljenih radova.

U trenutku pisanja rada, KPEX još uvijek nije bio spreman za ozbiljniju praktičnu primjenu. Kako bi se to omogućilo, potrebno je razriješiti specifičnosti vezane uz format Hininog skupa dokumenata te doraditi kod za laku integraciju s vanjskim servisima, primjerice *RESTful web servicea*. Takav servis moguće je prilično jednostavno ostvariti te bi u svojoj osnovnoj verziji primao naslov i tekst dokumenta, a vraćao automatski ekstrahirane fraze.

Daljnje poboljšanje sustava podrazumijeva ispitivanje rada sustava s velikim dokumentima. Ako je to moguće, potrebno je izgraditi model koji dobro radi s velikim i malim dokumentima. U slučaju da takav generalizirani model previše narušava rad sustava, dokumente je potrebno kategorizirati prema veličini u nekoliko intervala, a zatim izgraditi modele za svaku od definiranih kategorija. Prilikom takvog rada, ulazni dokument prvo bi se kategorizirao prema veličini, a zatim bi se nad njim primjenjivao odgovarajući model.

Konačno, sustav bi se mogao unaprijediti boljim razrješavanjem lema te alatom za ispravljanje zatipaka.

LITERATURA

Renee Ahel, Bojana Dalbelo Bašić, i Jan Šnajder. Automatic keyphrase extraction from croatian newspaper articles. 2009.

B. Krulwich i C. Burkey. Learning user information interests through the extraction of semantically significant phrases. *AAAI Spring Symposium on Machine Learning in Information Access, Stanford, CA; March.*, 1996.

Olena Medelyan i Ian H. Witten. Thesaurus based automatic keyphrase indexing. U *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, stranice 296–297. Chapel Hill, NC, USA : ACM Press, 2006.

Alberto Muñoz. Compoundkey word generation from document databases using ahierarchical clustering art model. U *Intelligent Data Analysis*, stranice 25–48. 1997.

Kamal Sarkar, Mita Nasipuri, i Suranjan Ghose. A new approach to keyphrase extraction using neural networks. *International Journal of Computer Science Issues, Vol. 7, Issue 2, No 3, March 2010*, 2010.

Peter D. Turney. Learning algorithms for keyphrase extraction. *INFORMATION RETRIEVAL*, 2:303–336, 2000.

J. Wang, H. Peng, i J.-S. Hu. Automatic keyphrases extraction from document using neural network. U *Advances in Machine Learning and Cybernetics 4th International Conference, ICMLC 2005, Guangzhou, China, August 18-21, 2005, Revised Selected Papers*, stranice 633–641. Springer Berlin / Heidelberg, 2006.

Ian H. Witten, Gordon W. Paynter, Eibe Frank, Carl Gutwin, i Craig G. Nevill-Manning. Kea: Practical automatic keyphrase extraction. U *IN PROCEEDINGS OF THE 4TH ACM CONFERENCE ON DIGITAL LIBRARIES*, stranice 254–255, 1998.

Ekstrakcija ključnih fraza metodama nadziranog strojnog učenja

Sažetak

Ključne fraze omogućuju indeksiranje i kategoriziranje dokumenata te njihovo jednostavno i brzo dohvaćanje unutar zbirke. Broj dokumenata neprestano i ubrzano raste, dok se ključne fraze rijetko dodjeljuju dokumentima. Potrebno je razviti sustave za automatsku ekstrakciju ključnih fraza što je ujedno i motivacija ovog rada. U radu je opisan sustav KPEX koji automatizira ekstrakciju ključnih fraza iz tekstnih dokumenata na hrvatskom jeziku. Algoritam se temelji na nadziranom strojnom učenju, a koriste se naivan Bayesov klasifikator i višeslojni perceptron kako bi se ustanovila bolja metoda. Razrađene su metode optimizacije ugađanjem filtriranja kandidata na temelju vrste riječi, ugađanjem diskretizacije kontinuiranih vrijednosti značajki i izvođenjem podskupa značajki iz proizvoljnog skupa značajki. Analizom je zaključeno da najbolje rezultate daje metoda višeslojnog perceptrona s MDL diskretizacijom, filtar od pet najučestalijih kombinacija vrsta riječi iz skupa unaprijed dodijeljenih ključnih fraza prilikom čega se koristi osam najboljih značajki. U slučaju pet zatraženih ključnih fraza K, sustav ekstrahira 2,2 isprave ključne fraze po dokumentu za 20 unaprijed dodijeljenih ključnih fraza po dokumentu. Ako je prosjek unaprijed dodijeljenih ključnih fraza jednak 4,6 i $K=5$, sustav dodjeljuje 0,8 ispravnih ključnih fraza po dokumentu. Nakon usporedbe rezultata sustava KPEX s rezultatima srodnih radova zaključeno je da je učinkovitost sustava KPEX dovoljno dobra za praktičnu primjenu.

Ključne riječi: naivan Bayesov klasifikator, višeslojni perceptron, ekstrakcija ključnih fraza, strojno učenje, hrvatski jezik

Keyphrase extraction based on supervised machine learning

Abstract

Keyphrases enable indexing and categorization of documents and also their simple and fast fetching from collection. Number of documents is in rapid growth while keyphrases are rarely assigned to documents. There is a need for systems that can automate keyphrase extraction which is also the goal of this project. In this paper we describe system KPEX which can automate keyphrase extraction from Croatian text documents. Procedure is based on supervised machine learning. We compare naive Bayes classifier and multilayer perceptron to find out which performs better. Different optimization methods are used such as adjustment of part-of-speech filtering, choosing the right discretization method for continuous feature values and feature subset selection. Analysis has shown that multilayer perceptron with MDL discretization and top five part-of-speech combinations built on preassigned keyphrases while using eight features performs the best. If five keyphrases are requested system extracts 2.2 correct keyphrases per document with on average 20 preassigned keyphrases. When tested on 4.6 preassigned keyphrases per document and five requested keyphrases system extracts 0.8 correct keyphrases per document. After comparing system KPEX with related projects we have concluded that system performs well enough to be used in practical purposes.

Keywords: naive Bayes classifier, multilayer perceptron, keyphrase extraction, machine learning, Croatian language