Sveučilište u Zagrebu
Fakultet organizacije i informatike, Varaždin

"Sustav za automatsko indeksiranje i kategorizaciju Web stranica na hrvatskoj domeni Interneta"

Jasminka Dobša
Danijel Radošević
Zlatko Stapić
Marinko Zubac

U Varaždinu, 30. studenog 2004.

Uvod

U vremenu neprestanog i eksponencijalnog porasta količine informacija koje nas okružuju, sve je važnije snaći se i znati plivati u tom moru koje bezdušno guta slobodni prostor na milionima tvrdih diskova računala povezanih u globalnu mrežu. Tako za razliku od mnogih zemalja kojima i na ovom polju bezuvjetno pomažu računala u automatskom indeksiranju, kategorizaciji i semantičkom pretraživanju teksta, u Hrvatskoj takav sustav (software) još ne postoji.

Cilj je ovog rada napraviti sustav koji će automatski indeksirati i kategorizirati Web stranice koje se nalaze u katalogu hijerarhijske strukture hrvatskih Web stranica smještenom na adresi http://www.hr/. Ovaj katolog razvijen je na Zavodu za telekomunikacije Fakulteta elektrotehnike i računarstva u Zagrebu, gdje se i danas ažurira. Za razliku od poznatih pretraživača (Google, Altavista,Yahoo) koji za prikupljanje i indeksiranje stranica koriste posebne programe koji samostalno pretražuju mrežu i indeksiraju linkove, www.hr je katalog stranica u koji se stranica dodaje isključivo na zahtijev korisnika, tj. autora stranice. Pri tome korisnik sam određuje kategorije u koje smatra da bi njegova stranica trebala biti uvrštena, nakon čega administrator ima pravo korigirati taj prijedlog. Stranice su organizirane u 14 osnovnih kategorija (na najvišoj hijerarhijskoj razini) , a ukupan broj kategorija je 650. Broj stranica kataloga svakodnevno raste za 10-20 i do sada ih ima oko 12 000.

Naš je zadatak razviti sustav za automatsko indeksiranje i kategorizaciju koji ćemo onda testirati na katlogu. U prvoj točci objašnjen je način reprezentacije hijerarhijske strukture kataloga pomoću xml datoteke. Ova datoteka sadrži osnovne podatake o stranici: identifikacijski broj, adresu, naziv, opis stranice, itd. Automatsko indeksiranje i kategorizacija vrši se na osnovi riječi sadržanih u opisu stranice koji daje autor. Takav opis često nije dovoljno informativan i naša je pretpostavka da nije pogodan za indeksiranje. Stoga smo odlučili izraditi proširenu xml datoteku koja će umjesto opisa stranice sadržavati tekst sadržan na stranici. Izvedba proširenog xml-a objašnjena je u drugoj točci.

Važno je napomenuti da će se automatska indeksiranje vršiti isključivo **analizom teksta (text mining)** na postojećim stranicama (a ne i slika). Stoga je potrebno razviti programsku podršku koja će svaku sakupljenu stranicu reprezentirati u obliku koji će omogućiti primjenu klasičnih algoritama za automatsku klasifikaciju teksta. Naša programska podrška omogućuje reprezentaciju stranice u obliku matrice koja sadrži informaciju o frekvenciji pojavljivanja indeksnih riječi (bag of words representation). To je osnovni oblik reprezentacije i on može imati brojne nedostatke. Npr. tu se javlja problem sinonima (više riječi s istim ili sličnim značenjem) i homonima (jedna riječ s više značenja). Slične stranice koje ne sadrže iste pojmove, već sinonime, neće biti prepoznate kao slične, a bitno različite stranice koje sadrže homonime biti će prepoznate kao slične na osnovi reprezentacije koja uvažava samo frekvenciju pojavljivanja riječi na pojedinoj stranici. Ovo je problem koji se intenzivno znanstveno proučava i implementacija nekih od postojećih rješenja biti će predmet budućeg rada.

Poseban problem za automatsku indeksiranje predstavlja velika morfološka složenost hrvatskog jezika. Za engleski jezik razvijeno je mnoštvo algoritama za svođenje riječi na njen korijenski oblik (stemming), dok se u Hrvatskoj ovakva istraživanja još uvijek provode.

1. Reprezentacija stranica

Matematički model za prikaz stranica jest **matrica pojmova i dokumenata (term-document matrix, eng.)**. Dokumenat je u ovom slučaju stranica, tj. njen prikaz u xml datoteci. Pojam je riječ sadržana u kolekciji dokumenata koja se nalazi na listi indeksnih pojmova. Lista indeksnih pojmova formirana je ovdje sljedećom procedurom:

1) formira se lista od svih riječi sadržanih u kolekciji dokumenata
2) izbace se riječi iz datoteke stop riječi; to su riječi koje se pojavljuju vrlo često i nisu pogodne za indeksiranje dokumenta, npr. ali, ovaj, i itd.
3) izbace se riječi koje se javljaju u manje od n_1 i više od n_2 dokumenata.

Matrica pojmova i dokumenata jest matrica $A = [a_{ij}]$, gdje je a_{ij} težina *i*-tog pojma u *j*-tom dokumentu. Težina pojma u dokumentu određuje se na osnovi frekvencije pojavljivanja pojma u dotičnom dokumentu i u cijeloj kolekciji. U matrici pojmova i dokumenata stupci predstavljaju vektorske prikaze dokumenata.

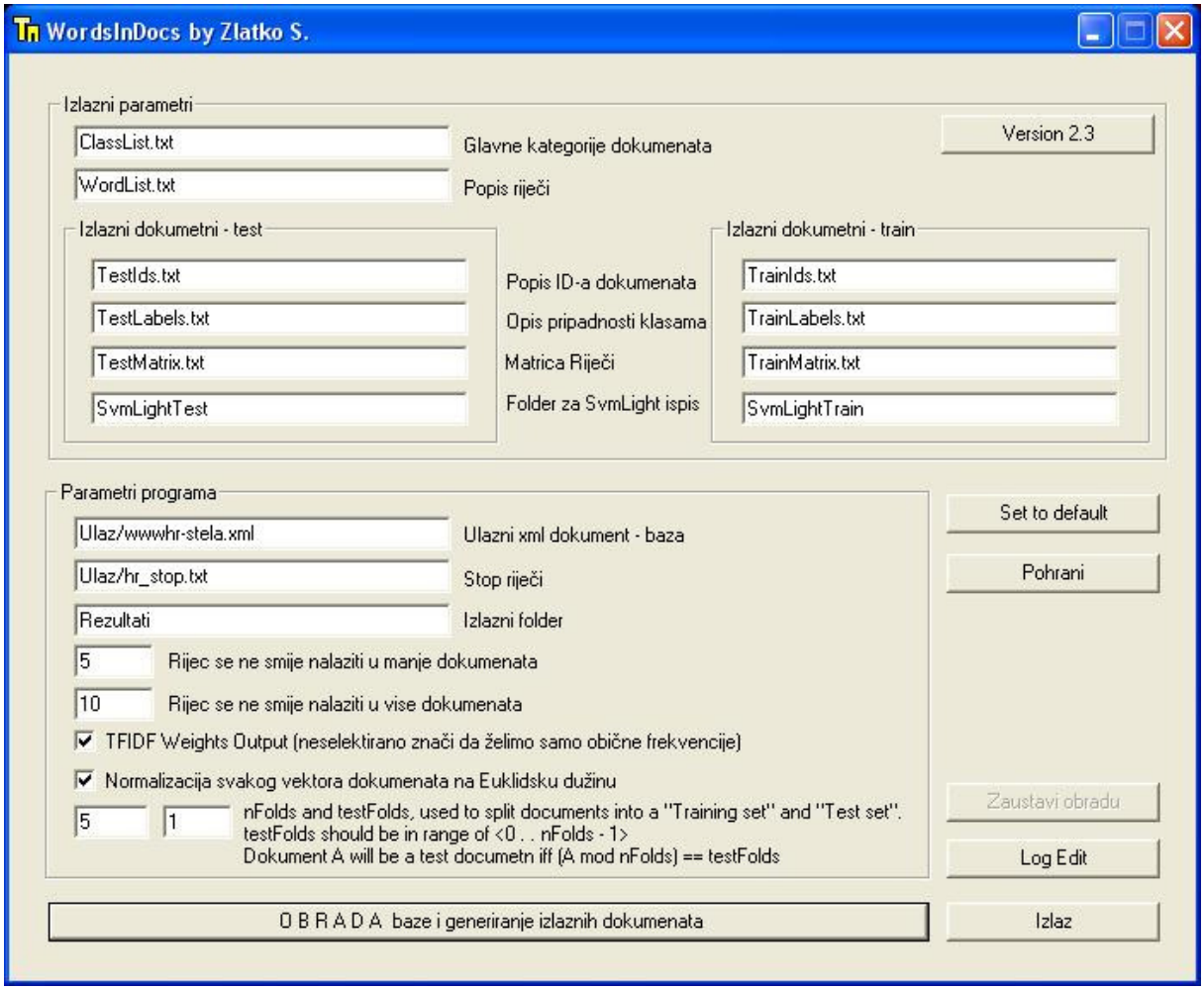
Ulazni parametri

Maska za unos ulaznih parametara i pokretanje obrade prikazana je Slikom 1. Obrada je uvjetovana ulaznim parametrima programa, i pored parametara kao što su imena ulaznih i izlaznih datroteka bitno je spomenuti sljedeće ulazne parametre:

- datoteka stop riječi – mi smo koristili datoteku stop riječi formiranu na Filozofskom fakultetu u Zagrebu
- najmanji broj dokumenata u kojima se treba nalaziti tražena riječ (n_1)
- najveći broj dokumenata u kojima se treba nalaziti tražena riječ (n_2)
- TFIDF – Ako je ova opcija selektirana onda se težine pojmova u dokumentima računaju korištenjem TFIDF formule, pa se težina *i*-tog pojma u *j*-tom dokumentu računa se prema formuli:

$$tezina_{ij} = f_{ij} * \ln\left(\frac{ukupan\ broj\ dokumenata}{broj\ dokumenata\ koji\ sadrze\ pojam\ i}\right)$$

  pri tome je f_{ij} frekvencija pojavljivanja *i*-tog pojma u *j*-tom dokumentu.
  Ako ova opcija nije selektirana onda su težine pojmova u dokumentima jednostavno njihove frekvencije.
- Normalizacija – Ako je ova opcija selektirana onda se vektorski prikazi dokumenata, tj. stupci matrice pojmova i dokumenata normaliziraju tako da budu jedinične norme.
- nFolds i testFolds parametri – određuju podjelu kolekcije na dokumenate za učenje i dokumenate za testiranje. Naime, svaki dokument gdje je A mod nFolds = testFolds bit će testni dokument dok su ostali dokumenti za učenje. Pri tome je A redni broj dokumenta u bazi, ili njegov ID.



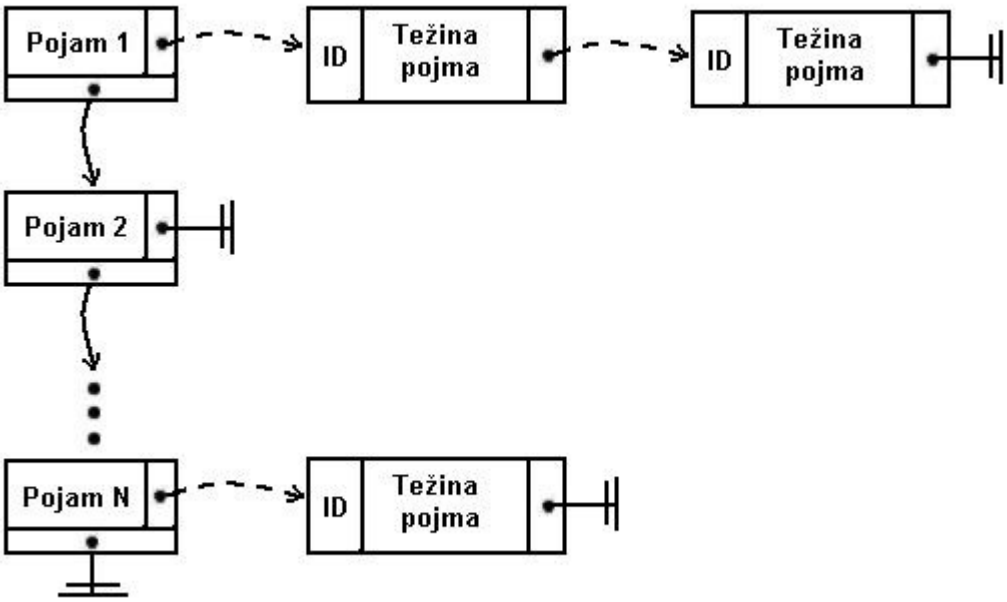
Slika 1. Maska za unos ulaznih parametara programa

Strukture podataka korištene za implementaciju matematičkog modela

Obrada podataka se vrši čitanjem xml datoteke i generiranjem određenih struktura koje sadrže popis riječi, popis dokumenata u kojima se te riječi nalaze i sl. Na taj način se implementira matematički model opisan gore. Obradom generirane strukture računaju se ostali potrebni podaci koji se potom zapisuju u izlazne datoteke.

Pojmovi se upisuju u dinamički kreirane liste podataka koje su u memoriji složene na način prikazan Slikom 2. Dakle, pojmovi su povezani u vezanu listu, gdje se pri kreiranju same liste slažu po abecednom redosljedu. Pored toga, svaki pojam je ujedno pokazivač na novu listu, listu u kojoj elementi sadrže popis ID-a dokumenata u kojima se taj pojam nalazi kao i težina pojma u spomenutom dokumentu.

Tzv. lista riječi prikazana Slikom 2 jest pojedostavleni prikaz. Lista pojmova ima sljedeću strukturu:



Slika 2. Dinamički kreirana lista podataka za implementaciju matrice pojmova i dokumenata

```
struct t_rijec{
bool obrisana;
char naziv_rijeci[150];
float TFIDF;
struct t_rijec *sljedeca;
struct t_rijec *prethodna;
struct t_u_dokumentu *prvi_dokument;
};

struct t_u_dokumentu{
int ID_dokumenta;
float broj_rijeci;
struct t_u_dokumentu *sljedeci;
};
```

Pri prvom čitanju baze se kreira i lista dokumenata. Struktura kojom se ona realizira je dana sljedećim fragmentom:

```
struct t_dokument{
int ID_dokumenta;
char naziv_dokumenta[500];
int ID_kategorija[200];
float normalize;
struct t_dokument *sljedeci;
};
```

Bitno za spomenuti je atribut '`int ID_kategorija[200]`'. U ovaj se vektor zapisuje kojim sve kategorijama dotični dokument pripada.

Treća je lista kategorija u kojoj su zapisani podaci o kategorijama.

Izlazne datoteke

Program ima 8 izlaznih datoteka. To su sljedeće datoteke:

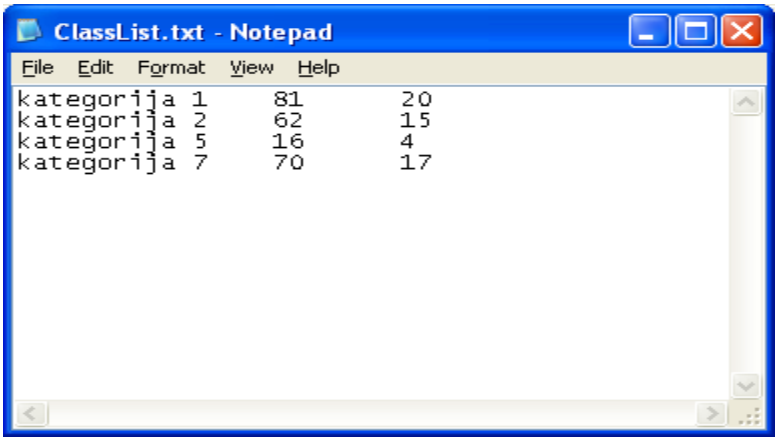
- ClassList.txt
- WordList.txt

- TrainIDs.txt TestIDs.txt
- TrainLabels.txt TestLabels.txt
- TrainMatrix.txt TestMatrix.txt

Objasniti ćemo izgled i funkciju svake od njih, te prikazati primjer ispisa izlaznih datoteka za ulaznu xml datoteku koja je dio xml datoteke pridružene katalogu hrvatskih Web stranica www.hr. Ova ulazna xml datoteka odnosi se na podskup od 100 stranica sa kataloga. Pri tome će ulazni parametri biti kao oni prikazani na maski za unos na Slici 1.

ClassList.txt

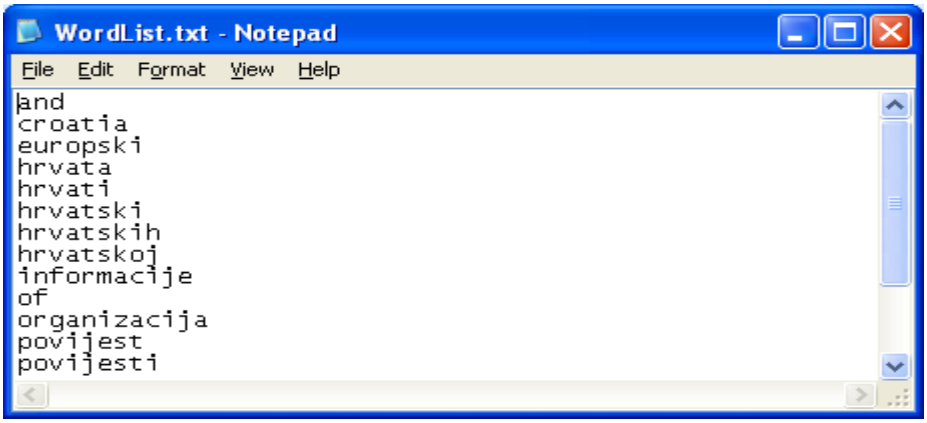
Kolekcija dokumenata (Web stranica) podijeljena je u dva dijela: učne dokumente (train documents, eng.) i testne dokumente (test documents, eng.). ClassList.txt jest izlazna datoteka u kojoj postoji popis svih kategorija kao i broj učnih i testnih dokumenata koji pripadaju navedenim kategorijama. Na Slici 3. prikazan je primjer ispisa izlazne datoteke .



Slika 3. Primjer ispisa izlazne datoteke ClassList.txt

WordList.txt

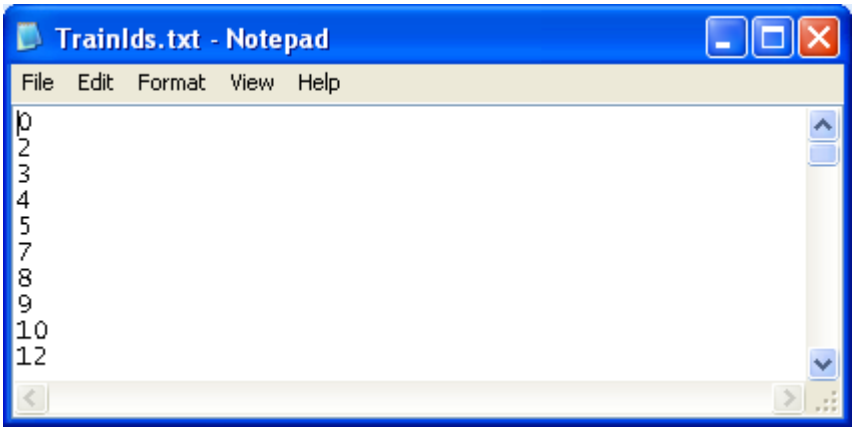
Datoteka u kojoj se nalazi popis svih obrađivanih riječi. Naime ovaj popis se kreira pri učenju i samo one riječi koje se nalaze u onoliko dokumenata koliko je zadano u intervalu ulaznih parametara ulaze u navedenu listu, a testiranje se vrši upravo na osnovu ovog popisa. Upravo zbog toga se ova lista pri testiranju zamjenjuje sa listom stop riječi pri učenju. Dok su riječi koje su se nalazile u listi stop riječi 'ispadale iz igre', pri testiranju je uvjet da se riječ nalazi u listi riječi koje su naučene. Slika 4 daje primjer ispisa datoteke WordList.txt.



Slika 4. Primjer ispisa izlazne datoteke WordList.txt

TrainIds.txt i TestIds.txt

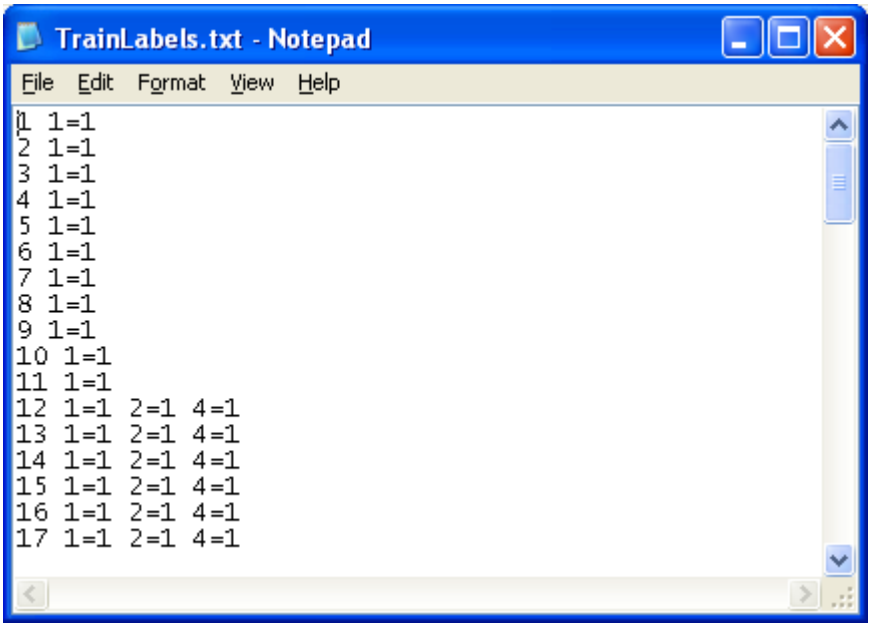
TrainIds.txt i TestIds.txt su izlazne datoteke koje sadrže popis identifikacijskih brojeva onih dokumenata koji su čitani pri učenju, odnosno testiranju. Slika 5 daje primjer ispisa izlazne datoteke TrainIds.txt.



Slika 5. Primjer ispisa izlazne datoteke TrainIds.txt.

TrainLabels.txt i TestLabels.txt

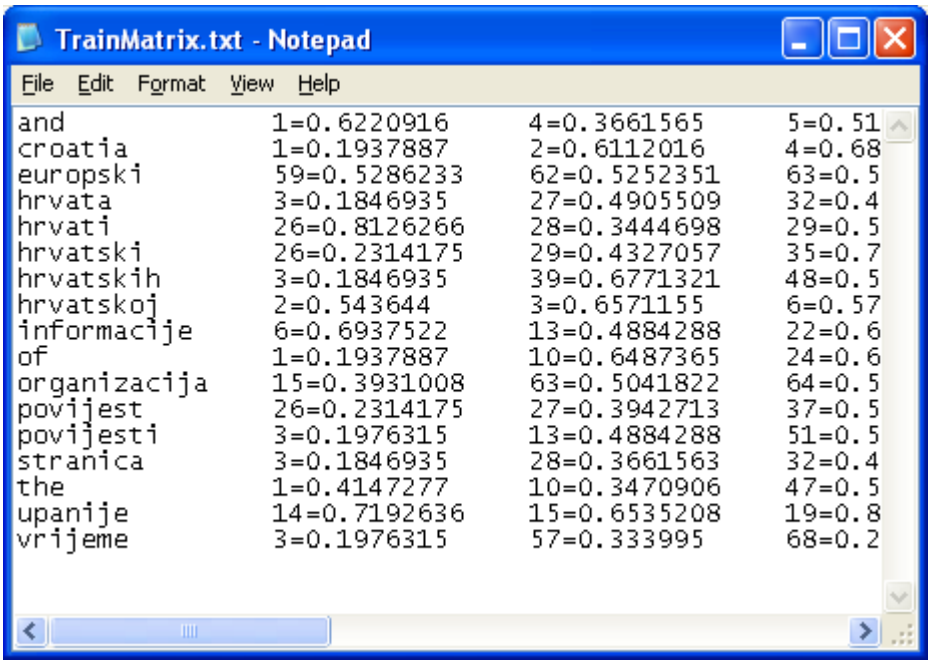
Datoteke koja sadrži informacije o pripadnosti učnih, odnosno testnih dokumenata određenim kategorijama. Primjer ispisa izlazne datoteke TrainLabels.txt dan je Slikom 6. U ovom primjeru 1. dokument pripada samo 1. kategoriji , dok na 12. dokument pripada 1., 2., i 4. kategoriji. Bitno je spomenuti da ovdje navedeni 12. dokument nije dokument čij je identifikacijski broj 12, već je to 12. učni dokument dok se njegov ID čita u datoteci TrainIDs.txt .



Slika 6. Primjer ispisa izlazne datoteke TrainLabels.txt

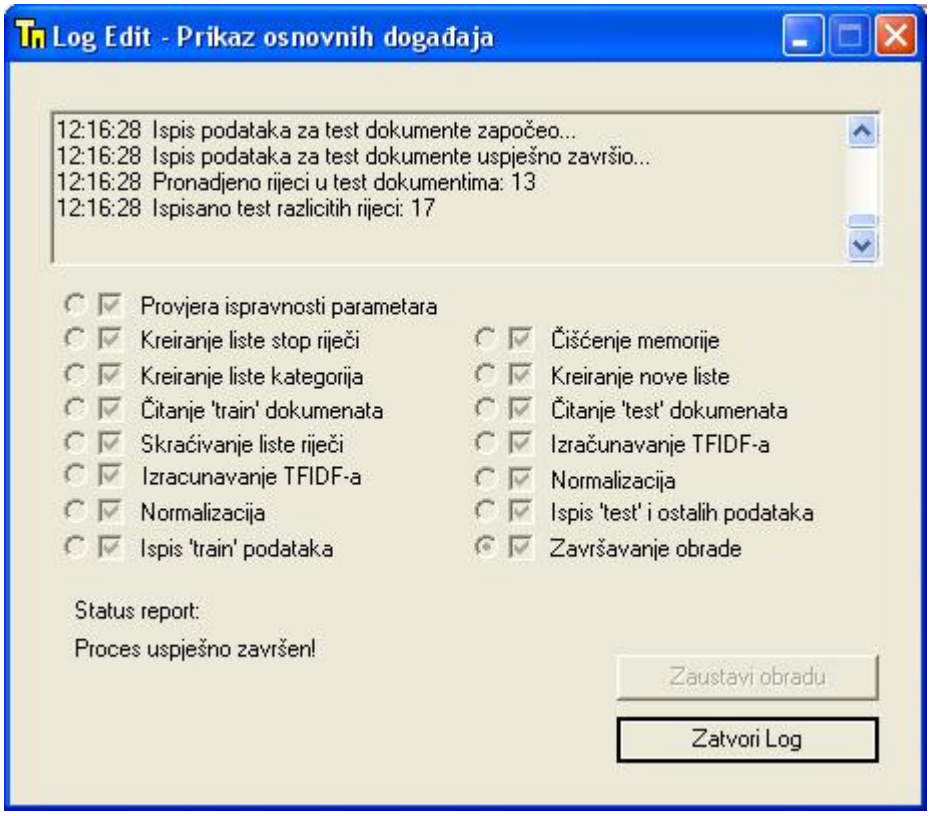
TrainMatrix.txt i TestMatrix.txt

Ovo su glavne datoteke koju generira program. U ovoj datoteci se nalazi popis svih riječi te popis ID dokumenata i težine riječi u navedenom dokumentu za učne, odnosno testne dokumente. Primjer ispisa izlazne datoteke TrainMatrix.txt dan je Slikom 7. Iz prikazanog se može vijdjeti da se pojam *stranica* nalazi u 3. učnom dokumentu s težinom 0.197, u 57. učnom dokumentu s težinom 0.33 itd. Primijetimo da lista indeksnih pojmova za primjer ove kolekcije od 100 dokumenata sadrži samo 17 pojmova koji zadovoljavaju sve uvijete postavljene ulaznim parametrima.



Slika 7. Primjer ispisa izlazne datoteke TrainMatrix.txt

Slika 8 prikazuje prozor za prikaz osnovnih događanja tokom obrade. Ovaj prozor omogućuje praćenje obrade.



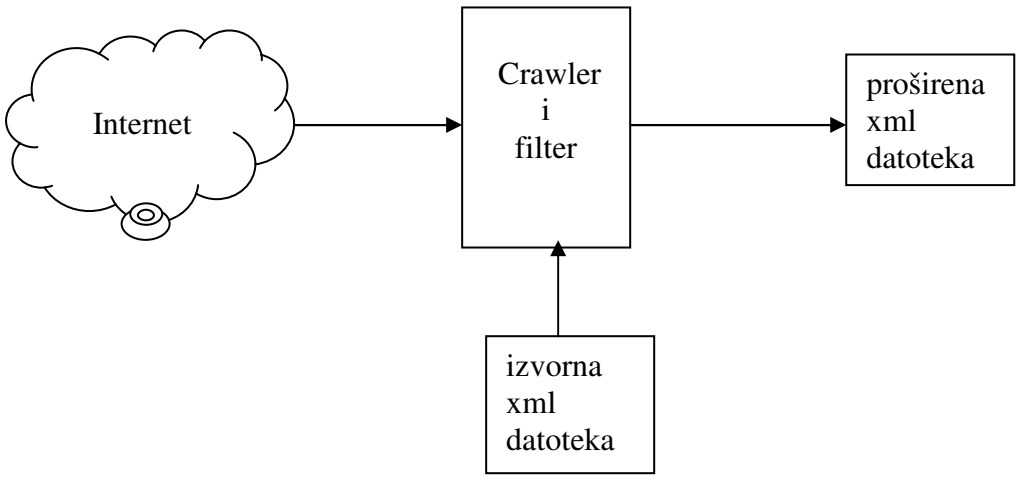
Slika 8. Prozor za prikaz osnovnih događanja tokom obrade

2. Prikupljanje sadržaja s Interneta (Crawler)

Izvorna xml datoteka koja korespondira Internet stranici http://www.hr svaki link iz kataloga opisuje vrlo šturo, samo nazivom stranice na koju upućuje link i kratkim opisom danim od korisnika. Zbog toga smo odlučili izraditi poseban program (crawler) koji bi s Web mjesta navedenih u izvornoj xml datoteci prikupljao tekst i uvrštavao ga u proširenu xml datoteku. Program je trebalo izraditi tako da omogućuje dvije funkcije

- osnovnu funkciju, a to je prikupljanje samog sadržaja s Web mjesta i
- pomoćnu funkciju, tj. filter koji bi pročišćavao prikupljeni tekst s Web mjesta.

Princip rada ovakvog programa prikazan je Slikom 9. Za spajanje na Web mjesta i prikupljanje sadržaja iskoristili smo skup Windows Internet (WinInet) API funkcija. Crawler poziva dretvu koja pomoću Internet funkcija prikuplja podatke s Web mjesta. Vrijeme izvođenja dretve je ograničeno i može se mijenjati iz sučelja programa. S time se ubrzava proces prikupljanja podataka i izbjegava se mogućnost da program "blokira" čekajući podatke od servera. Ako server ne odgovori programu u predviđenom vremenu, za to Web mjesto se uzimaju podaci iz izvorne xml datoteke, a u logu se bilježi adresa Web mjesta koje je izazvalo problem. Slika 10 prikazuje masku za unos ulaznih parametara programa.



Slika 9. Princip rada Crawlera

Program omogućuje upis sljedećih parametara:

- ulazna xml datoteka iz koje će se čitati ulazni podaci kao npr. kategorije, Web mjesta, naziv i opis dokumenata itd.
- naziv i lokacija proširene xml datoteke u koji će se upisivati novi podaci prikupljeni s Interneta
- napravi izvještaj o radu, ako je opcija izabrana izrađuje se poseban log u koji se bilježi napredak obrade, također se bilježi i količina podataka prikupljena za svaki dokument
- prati linkove u html-u, ako je opcija izabrana sa Internet stranica "izvlače" se linkovi, poveznice na druge stranice i bilježi ih se u poseban dokument; ti bi se linkove kasnije trebali iskoristiti za još potpunije proširenje xml datoteke
- prikaži napredak obrade
- čekaj n sekundi na odgovor od servera, omogućuje da se programu zada vrijeme koje će čekati na odziv od servera; ako u tom roku ne primi odgovor od servera, u novi xml datoteku će zabilježiti sadržaj iz izvorne xml datoteke, a program će preći na dohvatanje podataka sa sljedećeg Web mjesta

Za analizu rada dobro je uzeti primjer jednog Web mjesta iz xml datoteke za koje je crawler proširio količinu podataka. Tako npr. u izvornoj xml datoteci, u prvoj kategoriji koja nosi naslov *abouthr* i u kojoj se nalaze Web mjesta koja sadrže opće informacije o Republici Hrvatskoj pronalazimo Web mjesto *Croatia in English* koje je u izvornoj xml datoteci opisano na sljedeći način

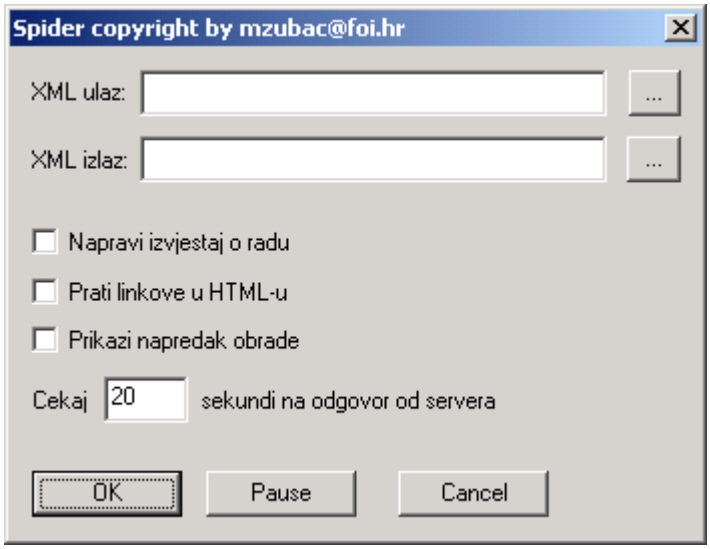
<link id="L0" numVisits="4830" dateAdded="2002-aug-23" rating="11.17">
Croatia-In-English
<desc>Purpose: This site is for English-speaking people who have an interest in Croatia, and especially for people of Croatian descent who were born into other cultures and who are now trying to learn more about their ethnic roots. The focus is primarily genealogy, but there is also some information on translation, travel, customs, and encouragment to visit the homeland.
</desc>
</link>

Nakon pokretanja crawlera, opisni tag *<desc>* za Web mjesto *www.croatia-in-english.com* sadrži mnogo više informacija i izgleda ovako:

<link id="L0" numVisits="4830" dateAdded="2002-aug-23" rating="11.17">
Croatia-In-English
<desc> croatian genealogy, geneology, travel, translation; home page this is the home page croatia-in-english.com editor@croatia-in-english.com purpose: this site is for english-speaking people who have an interest in croatia -- and especially for people of croatian descent who were born into other cultures and who are now trying to learn more about their ethnic roots. the focus is primarily genealogy, but there is also some information on translation, travel, customs, and encouragment to visit the homeland.
to robert jerin's links croatian immigrant history project if your family comes from near dubrovnik, please participate in this special project. new! forgotten faces. if you have some connection to watsonville, ca, please help identify these photos. new! all surnames of konavle. a rare, complete list from the southern tip of croatia. adriatic fact-finding service (genealogy) index at bottom. how the croatian alphabet is shown on this site. index at bottom. topics new! travel tips many topics: visiting
the family, using the telephone, driving in croatia, is it safe in croatia?, finding accommodations, gift bearing, drinking the water, etc. genealogy/family history this includes how to find the home village, how to find the church, how to read church documents, success stories, surname variants, tips on spelling the original name, social customs that can lead to proper identification, writing to the church, glossary of genealogy words, etc. this is the heart of this website. there is also a tutorial
series in this section. registry of surnames (genealogy) a list of surnames being researched and how to contact the researcher. nosquot;nja (national dress). photos of croatian national dress and embroidery from various regions. maps. only a few right now. more later. photos of villages. panoramas, overviews. [to be expanded] croatian communities outside croatia. places where croatians have congregated. sources and listings of genealogy information from these places. old postcards there are many
beautiful old postcards of croatia dating from the 1890s to around 1920. many of these are of the small villages -- just right for family histories. we are looking for examples to show. please send us digital images. see here first. translation translation problems (post your problem here), a registry of translators. links.... ....to other sites of interest and to robert jerin's links page. written material wantedwe are looking for helpful information -- articles, monographs, lists of terms and...
</desc>
</link>

Nakon što program prikupi sadržaj s određenog Web mjesta, sadržaj prolazi kroz filter koji odstranjuje html tagove i ostale nepotrebne sadržaje. Program u sadržaju pronalazi početak (<) i kraj (>) html taga, te taj dio briše iz prikupljenog sadržaja. Također je potrebno izbrisati i ostale html specifične oznake kao što su , &gt, &lt itd. Osim html oznaka, program mora prepoznati i izbrisati sintaksu određenog programskog jezika koji se izvršava na strani klijenta, kao npr. JavaScript, VBScript. To program obavlja na taj način da u sadržaju pronađe početni (<script>) tag i završni (</script>) tag, te sav sadržaj koji se nalazi između ta dva taga izbriše. Primijećen je veliki broj Internet stranica koje sadrže oznake od nekoliko slova u nizu koje ne predstavljaju određeno značenje, kao npr. aaaaaa >>>>> ..... ---------- abcdefg... Program za sada još nije u mogućnosti "pročistiti" takvu vrstu sadržaja, te tu nastaju mali problemi.

Kod prvog rada crawler je postojeću xml datoteku sa 3.9 MB proširio na 9.1 MB. Očekivali smo da će proširena xml datoteka biti veća u odnosu na izvornu, ali pogled u log crawlera otkriva da mnoga Web mjesta iz izvorne xml datoteke više ne postoje na Internetu ili je prikupljeni sadržaj vrlo kratak, te zato ne ulazi u novu xml datoteku. Tu smo se suočili i s novim problemom. Naime mnoga Web mjesta, pogotovo ona koja se nalaze na besplatnim servisima poput Yahoo! Geocities, Fortunecity, Lycos itd.., ne sadrže one podatke koji su opisani u izvornoj xml datoteci.



Slika 10. Maska za unos ulaznih parametara za crawler

Tako je npr. sadržaj Web mjesta http://www.geocities.com/SiliconValley/Grid/2816
welcome to yahoo! geocities - your home on the web
sorry, the site you requested is inactive. this geocities site has been deactivated due to inactivity. are you the site owner? click here to reactivate your site. are you a visitor? try a search below. search yahoo! geocities advanced geocities search options option 1 intelligent default an exact phrase match matches on all words (and) matches on any word (or) option 2 yahoo! geocities categories yahoo! geocities web sites addres....

Do rješenja tog problema još nismo uspjeli doći. Crawler za sada obavlja osnovnu funkciju prikupljanja i "pročišćavanja" sadržaja s Web mjesta navedenih u izvornoj xml datoteci. Za ubuduće je planirano da se poboljša način "pročišćavanja" te ugradi mehanizam koji bi omogućio praćenje poveznica koje se nalaze na određenim Internet stranicama.

3. Klasifikacija Web stranica

Za potrebu klasifikacije koristili smo metodu podupirućih vektora. **Metoda podupirućih vektora (support vector machines, eng.)** je je algoritam za klasifikaciju koji nalazi hiperravninu koja dijeli pozitivne od negativnih učnih primjera s maksimalnom mogućom marginom. To znači da je udaljenost od hiprravnine do najbližih pozitivnih, odnosno negativnih primjera maksimizirana. Sam algoritam je čvrsto teoretski fundiran i sofisticiran. Mi smo koristili programsku podršku SvmLight v.5.0 od Joachimsa (2002) sa uobičajenim parametrima. Za svaku kategoriju problem klasifikacije smo tretirali kao problem klasifikacije na dvije klase, pri čemu elemente dotične kategorije smatramo pozitivnim primjerima, a sve ostale dokumente negativnim primjerima.

Naš je cilj klasifikacija stranica sadržanih u katalogu hrvatskog Web prostora www.hr (stanje u novembru 2003. godine) i usporedba izvedbe klasifikacije za prikaz pomoću izvorne, odnosno proširene xml datoteke. Za reprezentaciju matrica pojmova i dokumenata formiranih na osnovi xml datoteka u obliku pogodnom za automatsku klasifikaciju koristili smo programsku podršku opisanu u 1. poglavlju korištenjem sljedećih parametara: za izvornu xml datoteku lista pojmova je formirana na osnovi

pojmova sadržanih u barem 4 dokumenata, a za proširenu xml datoteku na osnovi pojmova sadržanih u barem 6 dokumenata. U oba slučaja korištena je lista hrvatskih stop pojmova koja je formirana na Filozofskom fakultetu u Zagrebu pod vodstvom prof. dr. Marka Tadića. Za izvornu xml datoteku dobivena je lista od 7162 pojmova, a za proširenu lista od 11747 pojmova. Za potrebe klasifikacije dokumenti (stranice) su podjeljeni u dva dijela: učne dokumente i testne dokumente. Parametri nFold i testFold iz 1. poglavlja izabrani su da budu 5 i 1. Na taj način je određeno da svaki dokument čij identifikacijski broj pri djeljenju s 5 daje ostatak 1 postane testni, a ostali dokumenti su učni. Korištena je TFIDF težinska formula i vektorski prikazi dokumenata normalizirani su na jediničnu normu.

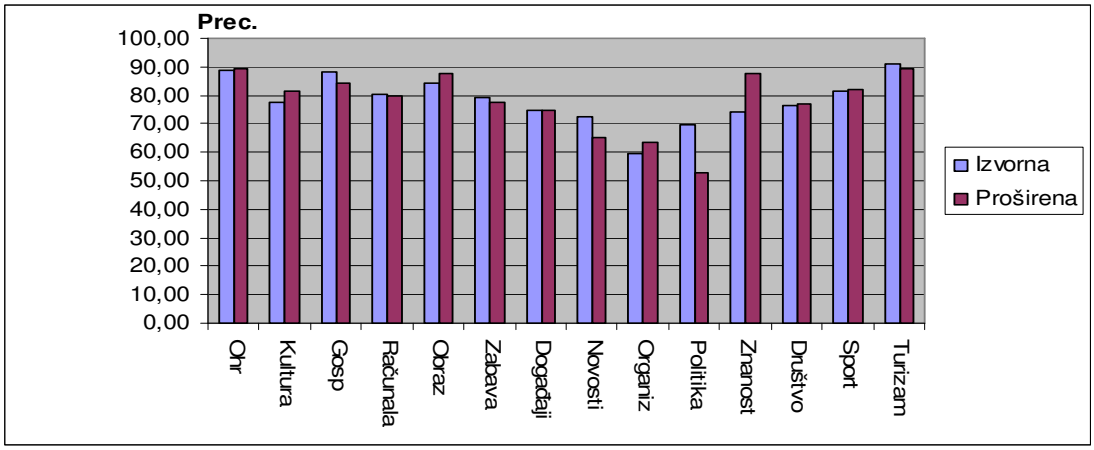
Dokumente smo klasificirali u 14 kategorija na najvišem nivou hijerarhije. To su sljedeće kategorije: O Hrvatskoj (Ohr), Kultura i umjetnost (Kultura), Gospodarstvo (Gosp), Računala i mreže (Računala), Obrazovanje (Obraz), Zabava (Zabava), Događaji (Događaji), Zakonodavstvo i politika (Politika), Novosti, mediji, časopisi (Novosti), Organizacije i udruženja (Organiz), Znanost i istraživanje (Znanost), Sport i rekreacija (Sport), Društvo (Društvo) i Turizam i putovanja (Turizam).

Za evaluaciju su korištene standardne mjere: odaziv (recall, eng.), preciznost (precision, eng.) i F_1 mjera. Odaziv r je definiran kao proporcija dokumenata koji su detektirani kao pozitivni modelom među stvarno pozitivnim dokumentima. Preciznost p je je proporcija stvarno pozitivnih dokumenata među dokumentima pozitivno klasificiranih modelom. F_1 je mjera koja uključuje odaziv i preciznost, a računa se po formuli $F_1 = 2pr/(p+r)$.

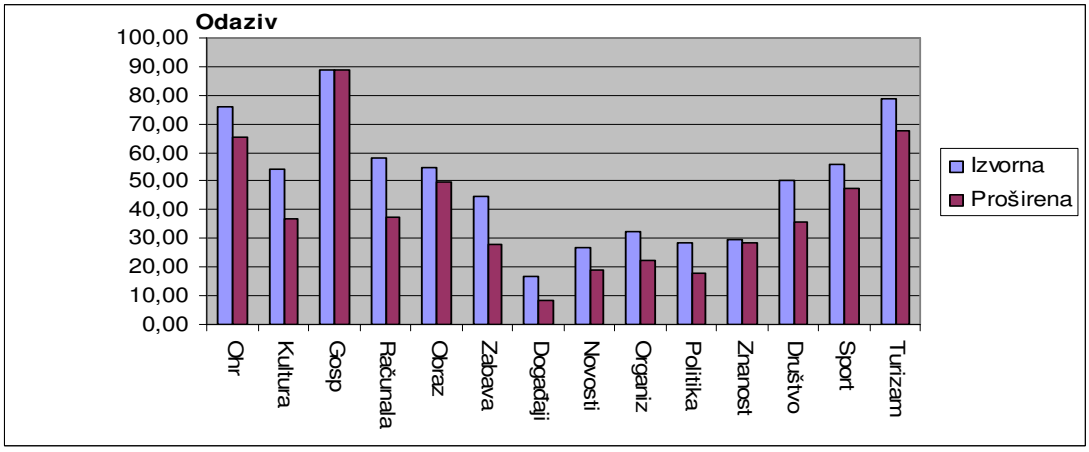
U Tabeli 1 dani su rezultati preciznosti, odaziva i F_1 mjere po kategorijama za obje baze. U zadnjem retku je izračunata makrosredina (macroaverage, eng.). Makrosredina preciznosti, odnosno odaziva izračunava se jednostavno kao aritmetička sredina tih mjera po svim kategorijama. Makrosredina F_1 mjere računa se po formuli navedenoj gore korištenjem makrosredina preciznosti i odaziva.

	Izvorna			Proširena		
Kategorija	Preciznost	Odaziv	F_1	Preciznost	Odaziv	F_1
Ohr	88,97	76,22	82,10	89,09	65,15	75,26
Kultura	77,53	53,99	63,65	81,63	36,81	50,74
Gosp	88,02	88,96	88,49	84,03	88,82	86,36
Računala	80,16	57,89	67,23	80,00	37,43	51,00
Obraz	84,38	54,55	66,26	87,50	49,49	63,22
Zabava	79,10	44,92	57,30	77,65	27,97	41,13
Događaji	75,00	16,67	27,28	75,00	8,33	14,99
Novosti	72,41	26,58	38,89	65,22	18,99	29,42
Organiz	59,57	32,56	42,11	63,33	22,09	32,75
Politika	69,57	28,57	40,51	52,63	17,86	26,67
Znanost	74,36	29,59	42,33	87,50	28,57	43,08
Društvo	76,68	50,39	60,82	76,97	35,58	48,66
Sport	81,36	56,14	66,44	81,82	47,37	60,00
Turizam	90,84	78,80	84,39	89,25	67,49	76,86
Macroaverage	78,43	49,70	60,84	77,97	39,425	52,37

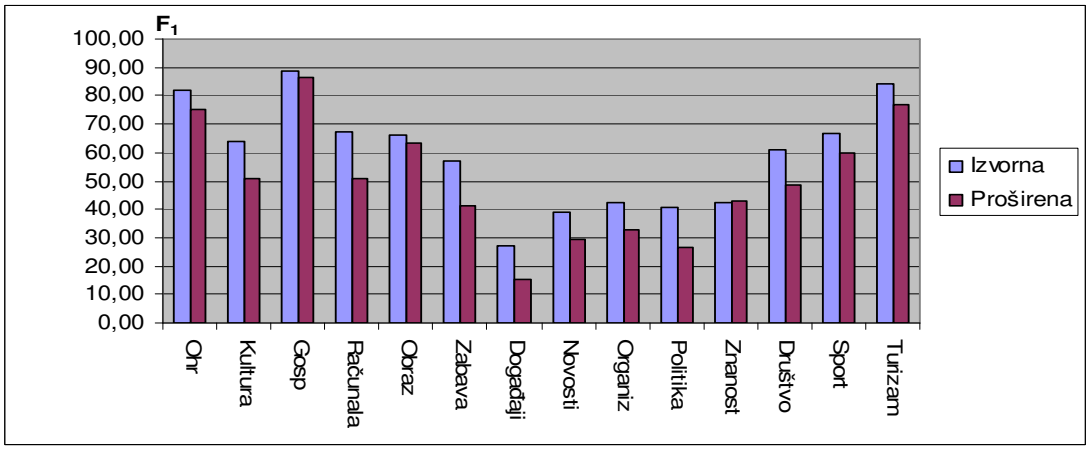
Tabela 1. Preciznost, odaziv i F_1 mjera za klasifikaciju korištenjem izvorne i proširene xml datoteke



Slika 11. Usporedba preciznosti klasifikacije po klasama za izvornu i proširenu xml datoteku



Slika 12. Usporedba odaziva klasifikacije po klasama za izvornu i proširenu xml datoteku



Slika 13. Usporedba F_1 mjere po klasama za izvornu i proširenu xml datoteku

Slika 11, 12 i 13 grafički daju usporedbu preciznosti, odaziva i F_1 mjere za klasifikaciju korištenjem izvorne, odnosno proširene xml datoteke. Možemo primijetiti da je preciznost podjednaka, dok je odaziv slabiji za klasifikaciju korištenjem proširene xml datoteke po svim kategorijama. To se može vidjeti i po makrosredini ovih mjera: makrosredina preciznosti su podjednake, dok je makrosredina odaziva za oko 10% niža za klasifikaciju korištenjem proširene baze. Shodno tome je i makrosredina F_1 mjere niža za klasifikaciju korištenjem proširene baze. Odaziv je posebno nizak za kategorije Događaji (Događaji), Novosti, mediji, časopisi (Novosti), Organizacije i udruženja (Organiz) što je očekivano, ali je isto tako prilično nizak za kategorije Zakonodavstvo i politika i Znanost i istraživanje što nije očekivani rezultat.

4. Diskusija i zaključak

Naš je cilj bio usporediti izvedbu klasifikacije na osnovi izvorne xml datoteke u kojoj su stranice reprezentirane pojmovima iz opisa stranice dane od korisnika i proširene xml datoteke u kojoj su stranice reprezentirane pojmovima sadržanim na samim stranicama. Primjetili smo da su sami opisi stranica dani od korisnika često više senzacionalistički, nego informativni, pa smo smatrali da bi klasifikacija korištenjem proširene xml datoteke mogla dati bolje rezultate. Međutim, ova naša pretpostavka nije se pokazala točnom. Najslabija karika u izvedbi klasifikacije očito je loš odaziv, pa bi u se u budućem radu trebalo težiti popravljanju ove karakteristike. Krajnji cilj ovog posla trebalo bi bili dohvaćanje stranica s Interneta i automatska klasifikacija tih stranica u hijerarhiju hrvatskog Web-a www.hr. Za dohvaćanje stranica trebalo bi adaptirati postojeći crawler. Naravno, za tako dohvaćene stranice neće nam biti dostupan opis stranice od strane korisnika, pa je obrada korištenjem izvorne xml datoteke samo referentna točka za daljni rad.

Literatura

1. N. Cristianini, J. Shave-Taylor, *Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, 2000.
2. F. Sebastiani, Machine learning in automated text categorization*, ACM Computing Surveys,* 34(1), 2002, pp. 1-47.