SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 117

OTKRIVANJE PRVE VIJESTI U SLIJEDNIM TEKSTNIM PODACIMA

Marin Akšamović

Zagreb, lipanj 2010.

Sadržaj

Popis oznaka i kratica

TDT	Otkrivanje i praćenje tema (eng. Topic Detection and Tracking)
FSD	Otkrivanje prve vijesti (eng. First Story Detection)
XML	(eng. EXtensible Markup Language)
tf-idf	(eng. Term Frequency x Inverse Document Frequency)
TP	Točno pozitivan sud (eng. True Positive)
FP	Lažno pozitivan sud (eng. False Positive)
TN	Točno negativan sud (eng. True Negative)
FN	Lažno negativan sud (eng. False Negative)

Popis tablica

Popis slika

1. Uvod

Vijesti danas dolaze iz mnogo različitih izvora kao što su televizija, radio, novine i internet. Kako se konstantno povećava broj novih vijesti objavljenih svaki dan pojavljuje se potreba za tehnikama koje automatski pretražuju, organiziraju i strukturiraju tekstne materijale iz raznih izvora vijesti. Upravo iz tog razloga pokrenuta je inicijativa za stvaranje sustava za otkrivanje i praćenje tema (eng. *Topic Detection and Tracking,* TDT).

Sustav za otkrivanje i praćenje tema sastoji se od pet glavnih zadataka [1]:

1. Segmentacija vijesti (eng. *story segmentation*),
2. Otkrivanje prve vijesti (eng. *first story detection,* FSD),
3. Otkrivanje skupine (eng. *cluster detection*),
4. Praćenje tema (eng. *topic tracking*),
5. Otkrivanje povezanosti vijesti (eng. *story link detection*).

Segmentacija vijesti je problem dijeljenja prijepisa vijesti iz radio ili televizijskih izvora u individualne vijesti. Ovaj problem nije prisutan u izravnom tekstnom izvoru vijesti jer je on već podijeljen na individualne vijesti, dakle prisutan je samo u televizijskim i radio vijestima.

Otkrivanje prve vijesti je problem prepoznavanja pojavljivanja nove teme u toku vijesti.

Otkrivanje skupine je problem grupiranja svih vijesti u stvarnom vremenu, temeljeno na temama o kojima raspravljaju.

Praćenje tema je praćenje toka vijesti kako bi se našle dodatne priče o temi koja je već identificirana koristeći nekoliko uzoraka.

Otkrivanje povezanosti vijesti je problem odlučivanja raspravljaju li dvije slučajno odabrane vijesti o istoj temi.

Ovaj rad će pružiti detaljniji uvid u zadatak otkrivanja prve vijesti, tj. FSD-zadatak i ponuditi nekoliko konkretnih rješenja zadatka za slijedne tekstne podatke na hrvatskom jeziku prikupljene s internet portala.

1.1. Osnovni pojmovi

Pojam *događaj* može se najjednostavnije definirati kao „nešto posebno što se dogodilo u određeno vrijeme na određenoj lokaciji“.

U okviru istraživanja otkrivanja i praćenja tema pojam *tema* definiran je kao sjemenski događaj ili aktivnost koji uključuje sve direktno povezane događaje ili aktivnosti.

Pojam *vijest* ili *priča* definiran je kao tematski povezan odsječak toka vijesti koji uključuje dvije ili više deklarativne nezavisne rečenice o jednom događaju. Vijest se smatra dijelom teme kada raspravlja o događajima ili aktivnostima koji su izravno povezani sa sjemenskim događajem teme.

Uzmimo u obzir sljedeću novinsku vijest:

NASLOV: Potres u Haitiju!

TEKST: Katastrofalan potres, najjači u više od 200 godina, pogodio je Haiti - najsiromašniju zemlju zapadne hemisfere. Razrušene su zgrade u tromilijunskom glavnom gradu Port-au-Princeu, a njegovi stanovnici zatrpani su u ruševinama. Još nema ni grubih procjena broja žrtava, no svjetske agencije prenose izjave spasitelja i izvjestitelja koji govore kako bi moglo biti više stotina, pa i tisuća mrtvih.

LOKACIJA: Port-Au-Prince, Haiti

VRIJEME: 12.01.2010.

Slika 1 - Primjer novinske vijesti

Događaj koji vijest opisuje je potres u Haitiju. Budući da je to prva vijest o tom događaju, taj događaj postaje sjemenski događaj nove *teme*. *Tema* sadrži sve *vijesti* o potresu u Haitiju i izravno povezanim događajima (u ovom slučaju to su primjerice spasilački napori i uzroci nepogode). Ako se naprimjer pojavi nova vijest koja opisuje potres u Čileu, ona postaje sjemenski događaj nove teme jer iako su vijesti naočigled slične, one nisu izravno povezane. Također, u obzir se uzima i vrijeme vijesti, pa bi vijest o novom potresu u Haitiju objavljena 14.01.2010. ušla u ovu temu, dok bi recimo ista vijest objavljena 14.01.2012. bila novi sjemenski događaj.

Teme se dijele pomoću 13 različitih pravila o tumačenju [2]:

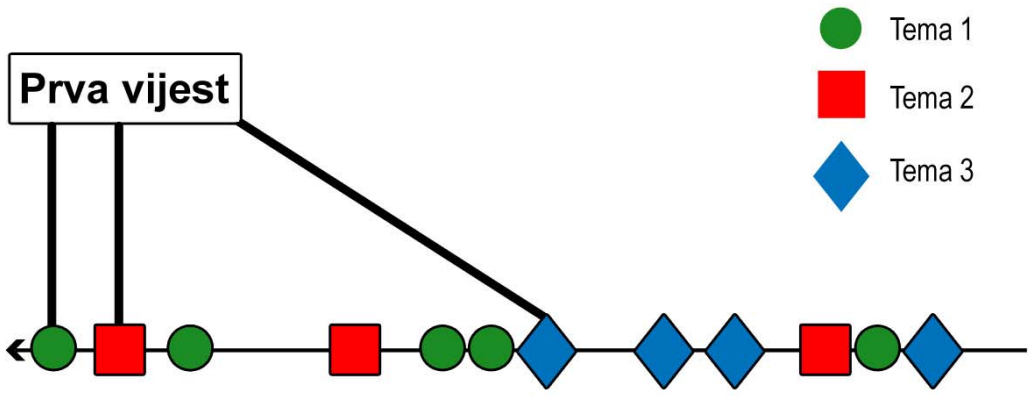
1. *Izbori*
    a. *Sjemenski događaj*: određena politička kampanja, inauguracija, rezultati izbora itd.
    b. *Tema*: potpuni proces izbora, od objave kandidature određenog kandidata kroz kampanju, nominacije, proces izbora do inauguracije
2. *Skandali i saslušanja*
    a. *Sjemenski događaj*: medijski prijenos određenog skandala ili saslušanja, prikupljanje dokaza, istraga itd.
    b. *Tema*: sve od početnog prijenosa skandala kroz istragu do rješenja
3. *Zakonski i kriminalni slučajevi*
    a. *Sjemenski događaj*: zločin, uhićenja, istrage, osude itd.
    b. *Tema*: potpuni proces od prijenosa početnog zločina kroz cijelu istragu i suđenje do ishoda.
4. *Prirodne nepogode*
    a. *Sjemenski događaj*: poremećaji u vremenu (tornado, poplava, suša itd.), drugi prirodni poremećaji poput vulkanskih erupcija i požara, spasilački napori, prijenos utjecaja katastrofe na ljude itd.
    b. *Tema*: uzrok nepogode uključujući i moguća predviđanja, sama nepogoda, žrtve i drugi gubici, evakuacije i spasilački napori
5. *Nesreće*
    a. *Sjemenski događaj*: prometne nesreće, gradski požari, eksplozije itd.
    b. *Tema*: uzrok i sve nezaobilazne posljedice poput broja poginulih, ozlijeđenih, novčanog gubitka, istrage, sve pravne akcije i odštete žrtvama.
6. *Djela nasilja ili rata*
    a. *Sjemenski događaj*: određeni čin nasilja ili terorizma ili niz direktno povezanih incidenata
    b. *Tema*: izravni uzrok i posljedice određenog čina nasilja kao što su pripreme (razvoj tehnologije i oružja), prijenos određene akcije, broj stradalih, pregovori za riješavanje sukoba, izravne posljedice uključujući uzvratni napad.
7. *Novosti u znanosti i nova otkrića*
    a. *Sjemenski događaj*: objava otkrića, tehnološko unapređenje, nagrade ili priznanja za znanstveno dostignuće itd.
    b. *Tema*: bilo koji aspekt otkrića, utjecaj na svakodnevni život, istraživači i znanstvenici uključeni u rad, opisi istraživanja i tehnologije direktno korištene u otkriću
8. *Financijske vijesti*
    a. *Sjemenski događaj*: određena ekonomska ili financijska priopćenja; reakcije na događaj; izravni utjecaj na ekonomski ili poslovni svijet.
    b. *Tema*: određeni događaj, njegovi izravni uzroci, utjecaj na financije, vladine intervencije ili istrage, reakcije javnog ili poslovnog svijeta, medijski prijenos i analiza događaja.

9. *Novi zakoni*
    a. *Sjemenski događaj*: najava novih zakona ili prijedloga, prihvaćanje ili odbijanje zakona, reakcije
    b. *Tema*: potpuni proces, najava prijedloga, lobiranje, glasanje, reakcija javnog i političkog svijeta, analiza i mišljenje o legislaciji
10. *Sportske novosti*
    a. *Sjemenski događaj*: određeni sportski događaj ili turnir, sportske nagrade, ozlijeda ili umirovljenje određenog sportaša itd.
    b. *Tema*: trening ili priprema za natjecanje, sam događaj, rezultat.
11. *Politički i diplomatski sastanci*
    a. *Sjemenski događaj*: pripreme za sastanak, sam sastanak, odluke, ishodi, reakcije
    b. *Tema*: potpuni proces od priprema i putovanja, kroz sam sastanak, medijski prijenos i reakcije javnosti do izravnog ishoda sastanka.
12. *Novosti u životu slavnih osoba*
    a. *Sjemenski događaj*: najčešće značajni događaji u životu slavne osobe poput braka ili smrti
    b. *Tema*: određeni događaj, uzroci ili posljedice, reakcija javnosti ili medijski prijenos, retrospektive ili životna povijest koje su izravna posljedica sjemenskog događaja
13. *Raznovrsne vijesti*
    a. *Sjemenski događa*j: svi događaji ili aktivnosti koji se ne mogu svrstati u jednu od prethodnih kategorija
    b. *Tema*: sam događaj, izravni uzroci i nezaobilazne posljedice

2. Otkrivanje prve vijesti

Zadatak otkrivanja prve vijesti jest prepoznavanje nove teme o kojoj se nije ranije raspravljalo. Postupak kojim se rješava ovaj zadatak prima tok vijesti kao svoj ulaz, a kao izlaz uobičajeno svakoj vijesti dodjeljuje brojčani rezultat koji označava pouzdanost sustava da je ta vijest stvarno prva vijest o nekoj temi. Ako pouzdanost neke vijesti prijeđe određeni prag, tada se ta vijest označava kao prva vijest.

U slučaju sustava za otkrivanje prve vijesti, *prva vijest* je pojam relativan trenutnom toku vijesti. To jest, to je prva vijest o novoj temi koja se pojavi u toku vijesti koji je predan sustavu iako ta vijest možda nije apsolutna prva vijest koja se pojavila u svim medijima. Dakle, moguće je da *prva vijest* u stvarnosti bude jedna od kasnije objavljenih vijesti, ali u predanom toku ona je uistinu prva vijest.



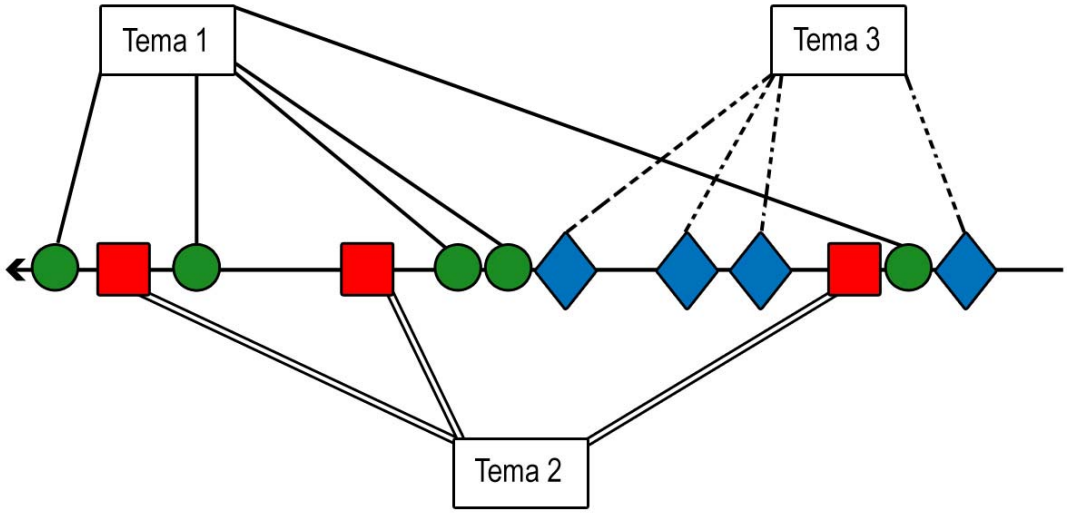
Slika 2 - Otkrivanje prve vijesti

Rad FSD-sustava ilustriran je na slici 1. Vijesti prilaze sustavu zdesna na lijevo i sustav označava prve vijesti. FSD-sustav radi u stvarnom vremenu, kao i cijeli TDT-sustav, tj. vijesti se obrađuju trenutačno pri njihovom dolasku.

Otkrivanju prve vijesti se uglavnom pristupa svođenjem vijesti na skup značajki, prikazanog kao vektor ili kao distribucija vjerojatnosti. Kada se pojavi nova vijest, njen skup značajki se uspoređuje sa svim prošlim vijestima. Ako postoji dovoljna razlika između nove vijesti i prijašnjih vijesti, tada se nova vijest označava kao prva vijest.

FSD-zadatak je uobičajeno TDT-zadatak s najmanjom uspješnošću i najvećom cijenom izvedbe, ali je ipak učinkovitiji od ručnog označavanja.

Poopćeni FSD-zadatak postaje zadatak otkrivanja skupine. U tom slučaju, uz otkrivanje prve vijesti sustav grupira sve vijesti iste teme u istu grupu. Rad tog sustava ilustriran je na slici 2. Nekoliko pristupa rješavanja FSD u isto vrijeme rješava i zadatak otkrivanja skupine, te će i jedan od njih biti detaljnije opisan u ovom radu.



Slika 3 - Otkrivanje skupine

Glavna razlika između zadatka otkrivanja skupine i zadatka otkrivanja prve vijesti jest način na koji ih TDT-sustav ocjenjuje. Zadacima otkrivanja skupine bitnije je skupiti sve vijesti o jednoj temi u grupu, ako promaše početak neke nove teme dobivaju samo malu kaznu uspjeha. Za razliku od njih zadatak otkrivanja prve vijesti strogo se kažnjava ako ne uspije naći početak nove teme i ne mari o tome što se događa unutar neke teme [1].

Zadatak otkrivanja skupina uglavnom se ostvaruje kao zadatak grupiranja. *Grupiranje* (eng. *clustering*) je zadatak nenadziranog učenja koji pokušava naći određenu strukturu u zbirci nepoznatih objekata. Jednostavnija definicija jest da je to proces koji organizira objekte u grupe čiji su članovi na neki način međusobno slični.

Grupa (eng. *cluster*) je zbirka objekata koji su međusobno slični i u isto vrijeme različiti od objekata u drugim grupama. U slučaju zadatka otkrivanja skupina jedna grupa sadrži sve vijesti o istoj temi.

2.1. Pregled područja

U prvoj TDT-inicijativi sudjelovalo je nekoliko različitih sveučilišta i komercijalnih tvrtki. Ta činjenica je doprinjela razvoju nekoliko različitih pristupa rješavanja FSD-zadatka.

U prvom kompletnom TDT-sustavu FSD-zadatak ostvaren je na potpuno jednak način kao zadatak praćenja tema, koji je fundamentalno sličan zadatku filtriranja informacija iz područja pretraživanja informacija (eng. *information retrieval*). FSD je tada postigao vrlo loše rezultate i zaključeno je kako je prihvatljiv FSD ostvaren na taj način moguć samo ako je sustav za praćenje savršen, što nije moguće u praktičnoj uporabi [3][4].

Jedna od metoda za rješenje FSD-zadatka je metoda najbližih susjeda. Ova metoda sprema sve vijesti u jednoelementne grupe i zatim spaja grupe, ako sličnost između grupa prelazi određeni prag. Sustav radi tako da čeka određeno vrijeme odgode (eng. *deferral period*) koje je dano kao neki broj vijesti, tj. kada se skupi određeni broj vijesti provjerava se sličnost među svim grupama, prošlim i sadašnjim, i vrši se spajanje. Ako nakon procesa spajanja još uvijek postoji jednoelementna grupa, tada se vijest koja je u toj grupi smatra prvom vijesti [5].

Jedan od najuspješnijih pristupa koristi vektorski prostorni model za predstavljanje svake vijesti, a zatim koristi tradicionalne metode grupiranja za predstavljanje događaja. Vijest je predstavljena kao vektor čija je veličina broj jedinstvenih izraza iz skupa podataka i čiji su elementi težina izraza u vijesti. Računanje težina izraza provodi se tako da izrazi koji se često pojavljuju u različitim vijestima imaju manju težinu od izraza koji se pojavljuju u trenutnoj vijesti ali ne i u ostatku skupa podataka.

Pristup grupira sve nadolazeće vijesti u grupe i vraća prvu vijest u svakoj grupi kao rezultat. Svaka nova vijest uspoređuje se s centroidom svake grupe i ako je dovoljno slična, vijest postaje dio te grupe. Inače, ako nije dovoljno slična ni jednoj drugoj grupi, stvara se nova grupa [6].

3. Pristup problemu

Za ostvarenje FSD-sustava odabran je pristup inkrementalnog vektorskog prostornog modela. Taj pristup se spominje kao najučinkovitiji pristup FSD-zadatku nakon prve TDT-inicijative [7]. Vektorski prostorni model je algebarski model koji predstavlja pojedine tekstne dokumente kao vektore. Model se smatra inkrementalnim kada se povećava nakon svake iteracije algoritma.

Postoji nekoliko različitih načina za pretvaranje dokumenta u vektor, a za ovaj sustav je odabran najpopularniji način, *mjera tf-idf*. Vektore je također potrebno međusobno uspoređivati, a tome služi *kosinusna sličnost*.

3.1. Tf-idf

Mjera tf-idf je statistička mjera koja označava koliko je određeni izraz važan jednom dokumentu u zbirci ili korpusu [8] . Važnost se proporcionalno povećava s brojem pojava izraza u dokumentu, ali se smanjuje s brojem pojava istog izraza u cijelom korpusu. Definicija pojma *izraz* ovisi o primjeni modela. Uglavnom se radi o jednoj riječi, ključnoj riječi ili dužim frazama.

Frekvencija pojave izraza (eng. *term frequency*, tf) uglavnom se normalizira kako bi se spriječila pristranost prema većim dokumentima, pa se definira kao

$$tf_{i.j} = \frac{n_{i.j}}{\sum_k n_{k,j}} \tag{1}$$

, gdje je $n_{i,j}$ broj pojava i-tog izraza u j-tom dokumentu, a u nazivniku je suma broja pojava svih izraza u j-tom dokumentu.

Inverzna frekvencija dokumenta (eng. *inverse document frequency,* idf) je mjera općenite važnosti izraza definirana kao

$$idf_i = \log \frac{N}{df_i} \tag{2}$$

, gdje je N broj svih dokumenata u korpusu, a df_i broj pojava i-tog izraza u svim dokumentima. U nazivnik se uobičajeno dodaje i jedna jedinica kako bi se izbjeglo moguće dijeljenje s nulom.

Konačna formula tf-idf mjere korištena za ovaj sustav je

$$(tf * idf)_{i.j} = tf_{i.j} * idf_i = \frac{n_{i.j}}{\sum_k n_{k,j}} * \log \frac{N}{1+df_i} \quad (3)$$

3.2. Kosinusna sličnost

Kosinusna sličnost je mjera sličnosti između dva *n*-dimenzionalna vektora. Računa se kao kosinus kuta između dva vektora i često je korištena za usporedbu dokumenata u vektorskom prostornom modelu [8]. Za vektore A i B kosinusna sličnost je definirana kao

$$cossim(A, B) = \frac{A \cdot B}{\|A\|\|B\|} \quad (4)$$

, gdje je u brojniku skalarni produkt vektora, a u nazivniku umnožak njihovih normi.

Vrijednost sličnosti između dva vektora koji predstavljaju dokumente biti će u intervalu [0,1], gdje 0 označava dva potpuno različita dokumenta, a 1 dva potpuno ista dokumenta.

3.3. Ostvarenje

Sustav iz ulaznog toka vijesti uzima jednu po jednu vijest i svakom iteracijom proširuje vektorski prostorni model. Vektorski model se gradi pomoću osnovne mjere tf-idf. Svaka nova vijest, tj. njen vektor tf-idf vrijednosti, uspoređuje se sa svim prethodnim vijestima preko kosinusne sličnosti vektora. Ako je nova vijest dovoljno različita od svih ostalih vijesti, tj. ako je sličnost s ostalim vijestima dovoljno malena, onda se ta vijest označava kao prva vijest neke teme. Optimalna vrijednost praga mjere kosinusne sličnosti biti će određena eksperimentalno.

Način rada bez grupiranja:

1) *Vijesti se učitavaju određenim redoslijedom.*
2) *Novo učitana vijest stvara novi vektorski model i briše stari.*
3) *Vektor trenutne vijesti pojedinačno se uspoređuje sa svim prethodnim vektorima u modelu.*
4) *Pamti se najveća sličnost, ako je vijest prva pamti se nula.*
5) *Ako je zapamćena sličnost manja od praga, sustav označava vijest kao prvu vijest. Ako je sličnost veća ili jednaka trenutna vijest nije prva vijest*.

Slika 4 - Algoritam načina rada bez grupiranja

Ostvareno je i proširenje prethodnog sustava koje omogućava i rješavanje zadatka otkrivanja skupina u isto vrijeme. Ovaj pristup stvara grupe za svaku novu prvu vijest i umjesto sa svim prethodnim vektorima, uspoređuje novu vijest samo s predstavnicima svake grupe. Predstavnik grupe je srednja vrijednost svih vektora u grupi, tj. centroid.

Način rada s grupiranjem:

1) *Vijesti se učitavaju određenim redoslijedom.*
2) *Prva vijest stvara prvu grupu.*
3) *Svaka sljedeća vijest stvara novi vektorski model i uspoređuje se sa predstavnicima svih grupa.*
4) *Pamti se najveća sličnost.*
5) *Ako je zapamćena sličnost veća ili jednaka pragu, vijest se dodaje u odgovarajuću grupu. Ako je sličnost manja, vijest stvara novu grupu i sustav je označava kao prvu vijest.*

Slika 5 - Algoritam načina rada s grupiranjem

4. Podatkovni uzorak

Vijesti su ručno prikupljene sa sljedećih internet portala: *index.hr, tportal.hr, net.hr, jutarnji.hr, vecernji.hr, dalje.com, monitor.hr, nacional.hr, bussiness.hr, glas-slavonije.hr, seebiz.eu, slobodnadalmacija.hr*.

Prikupljanje vijesti je obavljeno tokom dva različita petodnevna razdoblja, a ukupan broj prikupljenih vijesti je 1032. Raspodjela prikupljenih vijesti po vremenskim razdobljima i izvorima je vidljiva u tablici 1, odnosno 2.

Tablica 1 - Raspodjela vijesti prema vremenskim razdobljima

Vremensko razdoblje	Količina
06.04. – 10.04. 2010.	498
03.05. – 07.05. 2010.	534

Pri odabiru vijesti vodilo se računa o tome da je prikupljena vijest uistinu vijest, a ne komentar na trenutne događaje kako bi se izbjegle vijesti koje raspravljaju o više tema. Sve prikupljene vijesti su na hrvatskom jeziku.

Tablica 2 - Raspodjela vijesti prema izvoru

Izvor	Količina
bussiness.hr	49
dalje.com	104
glas-slavonije.hr	30
index.hr	131
jutarnji.hr	123
monitor.hr	39
nacional.hr	90
net.hr	109
seebiz.eu	31
slobodnadalmacija.hr	29
tportal.hr	146
vecernji.hr	151

Vijest je prikupljena tako da se izvorišna adresa vijesti, naslov i tekst vijesti ručno upišu u odgovarajući XML-dokument čije je ime naslov vijesti.

Primjer XML-dokumenta:

```
<document>

<sourceURL> Izvorišna adresa </sourceURL>

<title> Naslov </title>

<text> Tekst vijesti </text>

<date>
      <day> Dan </day>
      <month> Mjesec </month>
      <year> Godina </year>
</date>

<first_story_real> TRUE ili FALSE </first_story_real>

<first_story_classified> TRUE ili FALSE </first_story_classified>

<first_story_confidence> Pouzdanost </first_story_confidence>

<cluster_ID_real> TRUE ili FALSE </cluster_ID_real>

</document>
```

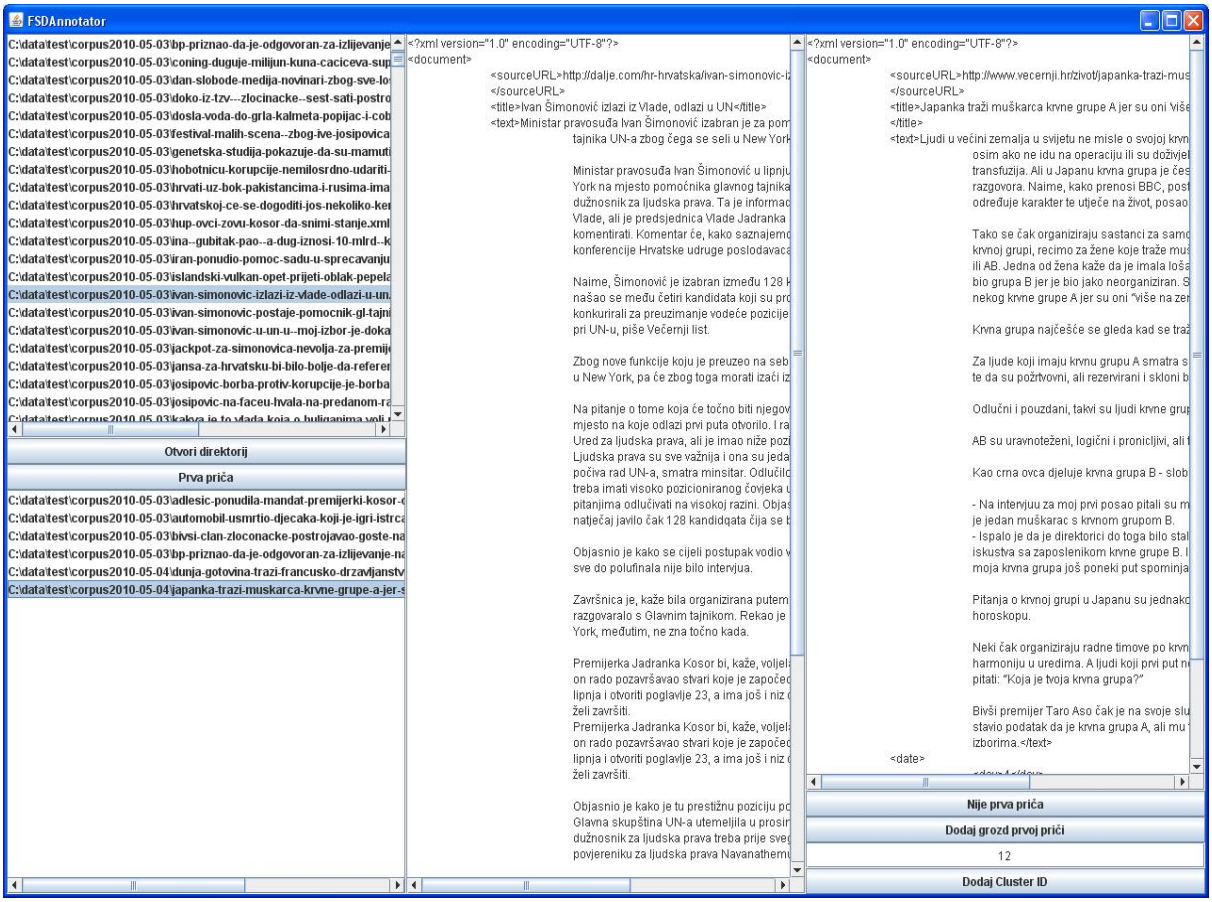
Element `first_story_real` označava je li vijest stvarno prva vijest i dobiven je ručnim označavanjem. Vrijednost TRUE označava vijest kao prvu vijest, a vrijednost FALSE označava da vijest nije prva vijest. Taj element je bitan za evaluaciju sustava, a sam sustav ga ne koristi.

Element `first_story_classified` je Booleova vrijednost koju sustav dodjeljuje svakoj vijesti na temelju elementa unutar `first_story_confidence`.

Element `cluster_ID_real` označava kojoj grupi pripada vijest, i taj element je također dobiven ručnim označavanjem i služi za evaluaciju sustava.

Nakon što su svi uzorci prikupljeni, njihovi XML-dokumenti su djelomično vremenski raspoređeni, tj. svi uzorci skupljeni u jednom danu se nalaze u jednom direktoriju koji predstavlja taj dan (npr. uzorci prikupljeni 6.4.2010. nalaze se u direktoriju *corpus2010-04-06*) dok su unutar tog direktorija poredani abecedno.

Svi uzorci su ručno označeni poštivajući pravila iznesena u TDT priručniku za označavanje [2]. Ta pravila su pobliže objašnjena u poglavlju 1.1. Za potrebe označavanja razvijen je jednostavni pomoćni program *FSDAnnotator*.



Slika 6 - FSDAnnotator

Program *FSDAnnotator* omogućava jednostavno dodavanje elemenata `first_story_real` i `cluster_ID_real` u odabrani XML-dokument preko grafičkog sučelja. Potrebno je dodati direktorij s željenim uzorcima nakon čega će se pojaviti lista uzoraka u gornje lijevom stupcu. Odabirom jednog uzorka iz tog stupca, u srednjem elementu se pojavljuje njegov sadržaj. Ako se taj uzorak označi kao prva priča, on se dodaje u donji lijevi stupac kako bi se omogućila jednostavnija usporedba sa sljedećim uzorcima. Desni element prikazuje sadržaj odabrane prve priče. Moguće je i izbrisati prvu priču i vratiti `first_story_real` na FALSE.

Također je moguće i upisati broj koji predstavlja pripadnost određenoj grupi i upisati ga u `cluster_ID_real`. To se odnosi na dokumente iz gornje lijevog stupca dok je za prve priče moguće dodati slijedno rastući broj (od 0 pa nadalje) svakoj sljedećoj prvoj priči jer svaka prva priča mora pripadati različitoj grupi.

Ručno označavanje je obavljao jedan sudac jer je prosuđeno da označavanje prve vijesti i označavanje pripadnosti grupa nisu subjektivni problemi.

5. Programsko ostvarenje

FSD-sustav ostvaren je u programskom jeziku Java zbog jednostavnije prenosivosti, neovisnosti o operacijskom sustavu i raznim korisnim bibliotekama otvorenog koda. Od biblioteka su korištene *Jdom* biblioteka [9] za rad s XML-dokumentima, *Jama* biblioteka za rad s matricama [10] i biblioteka *Lemmatizer.dll* za lematizaciju vijesti.

Sustav se sastoji od glavnog programa *FSDMain* koji dodaje vrijednosti u `first_story_confidence` i po potrebi stvara grupe i dodatnih programa *FSDEval* za evaluaciju samog sustava koji dodaje vrijednost u `first_story_classified` i *FSDClusEval* za evaluaciju grupiranja.

Glavni program slijedno obrađuje XML-dokumente iz zadanog direktorija i njegovih poddirektorija. Pri pokretanju je moguće odabrati način rada, tj. način bez grupiranja ili način s grupiranjem. Način rada bez grupiranja pokreće se bez definiranja praga za odluku jer se odluka o pripadnosti priče može odgoditi za evaluaciju, dok je za način rada s grupiranjem potrebno definirati prag jer je potrebno donijeti odluku o pripadnosti grupi.

Dokument se obrađuje na sljedeći način. Iz dokumenta se pročita naslov i tekst vijesti i oni se spremaju u jedan tekstni niz. Zatim se na tom tekstnom nizu provodi *lematizacija*. *Lematizacija* (eng. *lemmatisation*) je svođenje riječi iz korpusa na njihove natukničke oblike, tj. svođenje različitih riječi na zajedničku lemu. Na primjer, riječi *stol*, *stolova* ili *stolu* bile bi svedene na lemu *stol* [11]. Pri tome je *lema* kanonski oblik neke riječi. U poglavlju 3.1. spomenuto je kako pojam *izraz* ovisi o primjeni modela. U ovom slučaju *izraz* je upravo *lema*, tj. jedna vijest će biti predstavljena skupom lema koje se nalaze u njoj. Lematizacija se provodi uporabom automatski pribavljenog flektivnog leksikona [12]. Važno je spomenuti i da se pri postupku lematizacije iz tekstnog niza izbacuju stop riječi (eng. *stop words*). To su riječi koje se vrlo često pojavljuju u raznim dokumentima, a pri tome nemaju bitno značenje i ne pridonose opisu nekog dokumenta i stoga se često filtriraju iz dokumenata. U ovom slučaju stop riječi su svi prijedlozi, zamjenice, veznici, čestice i brojevi.

Nakon što je tekstni niz lematiziran, pamti se pripadnost svake leme svakoj vijesti i broj pojava leme unutar određene vijesti. Zatim se gradi tf-idf težinska matrica, tj. vektorski model, gdje stupci matrice predstavljaju dokument, retci leme, a elementi tf-idf vrijednosti lema u dokumentima. Budući da se matrica proširuje s dodatnim stupcem i nepoznatim brojem dodatnih redaka nakon svake nove vijesti, radi se o inkrementalnom modelu. Nakon toga se zadnji stupac, tj. vektor trenutne vijesti, uspoređuje sa svim prethodnim stupcima matrice (ili sa svim predstavnicima grupa ako je odabran način rada s grupiranjem) i u trenutni XML-dokument se upisuje najviša vrijednost u `first_story_confidence`. Ako je odabran način s grupiranjem, dokument se odmah klasificira kao TRUE ako je ta vrijednost manja od zadanog praga i kao FALSE ako je veća ili jednaka. Ako je FALSE vijest se također dodaje u grupu kojoj je najbliža. Nakon što je jedan XML-dokument, tj. vijest, obrađen kreće se sa obradom sljedećeg dokumenta dok nisu obrađeni svi dokumenti u cijelom zadanom direktoriju i njegovim poddirektorijima.

Slijedi primjer rada sustava za dvije vijesti. Radi jednostavnosti koristit će se samo jedan par rečenica za predstavljanje vijesti, a ne cijeli XML-dokument.

Primjer rada sustava:

Prag = 0,3 , neki slučajno odabrani prag za potrebe primjera.

Prvi prolaz:

Vijest 1 = *„Ovo je prva prilazeća vijest. Ona će biti označena kao prva vijest.“*

Lematizacija 1 => *„prilazeći vijest biti označen vijest“*

Tablica 3 - Vektorski model, prvi prolaz

	Vijest 1
biti	0,2
prilazeći	0,2
vijest	0,4
označen	0,2

Usporedba nije moguća jer se radi o prvom dokumentu, tj. vijesti, u sustavu pa je pouzdanost vijesti jednaka nuli.

Vrijedi (pouzdanost < prag) pa je vijest 1 označena kao prva vijest.

Drugi prolaz:

Vijest 2 = *„Sustavu prilazi sljedeća vijest. Uspoređuje se s prethodnom viješću.“*

Lematizacija 2 => *„sustav prilaziti sljedeći vijest uspoređivati prethodan vijest“*

Tablica 4 - Vektorski model, drugi prolaz

	Vijest 1	Vijest 2
sustav	0	0,1618
prethodan	0	0,1618
biti	0,2392	0
uspoređivati	0	0,1618
prilaziti	0	0,1618
prilazeći	0,2392	0
vijest	0,2825	0,1911
označen	0,2392	0
sljedeći	0	0,1618

Usporedba između vijesti 2 i vijesti 1 = 0,2632

Pouzdanost vijesti 2 je 0,2632.

Vrijedi (pouzdanost < prag) pa je i vijest 2 označena kao prva vijest.

Za više vijesti proces se analogno nastavlja.

6. Evaluacija

Nakon što je FSD-zadatak programski ostvaren, provedena je evaluacija oba načina rada i usporedba rezultata. Evaluacijom će se saznati uspješnost programskog ostvarenja.

6.1. Tablica zabune

Evaluacija FSD-zadatka izvedena je pomoću tablice zabune i osnovnih evaluacijskih mjera. Primjer tablice zabune za FSD-sustav dan je u tablici 4, gdje doneseni sudovi ovise o vrijednostima (TRUE ili FALSE) unutar elemenata XML-dokumenta `first_story_real` i `first_story_classified`.

Tablica 5 - Tablica zabune

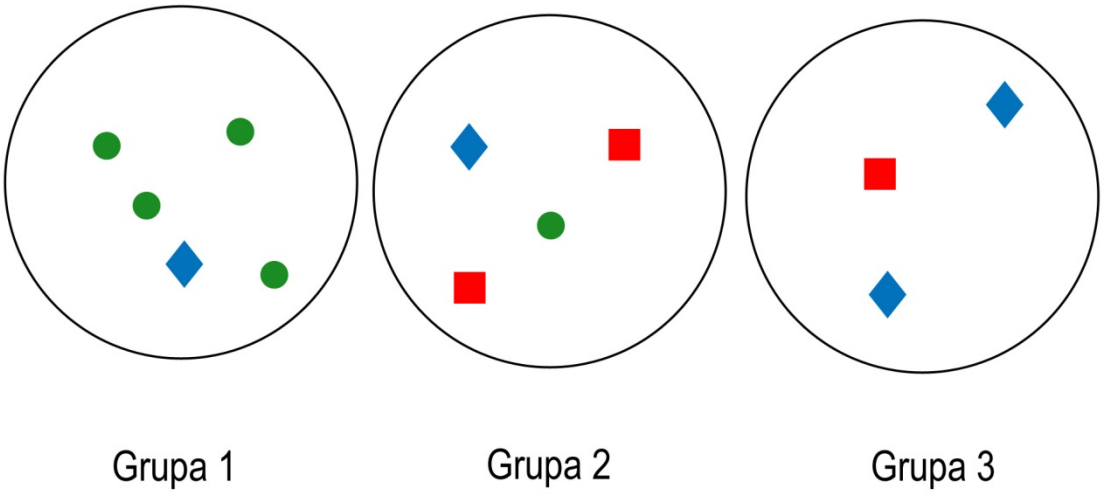
		Stvarno stanje, `first_story_real`	
		TRUE	FALSE
Predviđeno stanje, `first_story_classified`	TRUE	Točno pozitivan True Positive = TP	Lažno pozitivan False Positive = FP
	FALSE	Lažno negativan False Negative = FN	Točno negativan True Negative = TN

Osnovne evaluacijske mjere [13]:

- *Točnost* (eng. *accuracy*) je udio točno klasificiranih primjera u skupu svih primjera.
  - $Točnost = \frac{TP+TN}{TP+FP+FN+TN}$
- *Preciznost* (eng. *precision*) je udio točno klasificiranih primjera u skupu pozitivno klasificiranih primjera.
  - $Preciznost = \frac{TP}{TP+FP}$
- *Odziv* (eng. *recall*) je udio točno klasificiranih primjera u skupu svih pozitivnih primjera.
  - $Odziv = \frac{TP}{TP+FN}$

- *Specifičnost* (eng. *specificity*) je udio točno klasificiranih primjera u skupu svih negativnih primjera.
    - $Specifičnost = \frac{TN}{TN+FP}$
- *F1-mjera* je harmonijska sredina preciznosti i odziva, koja jednaku važnost pridaje preciznosti i odzivu.
    - $F1\text{-}mjera = \frac{2*preciznost*odziv}{preciznost+odziv}$

6.2. Randov indeks



Slika 7 - Primjer grupiranja

Za evaluaciju grupiranja korišten je Randov indeks i F1-mjera. Evaluacijske mjere koje se koriste u ovoj vrsti evaluacije su identične osnovnim evaluacijskim mjerama uz činjenicu da je vrijednost Randova indeksa analogna vrijednosti točnosti, ali je zato različit način izračuna vrijednosti tablice zabune.

Grupiranje se interpretira kao niz odluka, po jedna za $\frac{N*(N-1)}{2}$ parova entiteta, gdje je N ukupan broj entiteta. Cilj grupiranja u tom slučaju je svrstavanje dvaju entiteta u istu grupu ako i samo ako su slični. Istinito pozitivna odluka (TP) dodjeljuje dva slična entiteta u istu grupu, dok istinito negativna odluka (TN) dodjeljuje dva neslična entiteta, u različite grupe. Uz ove točne odluke moguće je ostvariti dvije vrste pogrešaka – lažno pozitivna odluka (FP) pri kojoj se dva neslična entiteta dodijele istoj grupi te lažno negativna odluka

(FN) koja dva slična entiteta dodjeljuje dvama grupama. Randov indeks mjeri postotak odluka koje su istinite, tj. vrijedi *Randov indeks* = $\frac{TP+TN}{TP+FP+FN+TN}$ [8][14].

Izračun vrijednosti odluka biti će prikazan na primjeru sa slike 4. Prvo se računa ukupan broj pozitivnih odluka kao općeniti broj parova u grupama. Budući da grupe sa slike redom sadrže 5,4 i 3 entiteta izračun je sljedeći

$$P = TP + FP = \binom{5}{2} + \binom{4}{2} + \binom{3}{2} = 19 \tag{5}$$

Zatim je potrebno izračunati broj parova u entitetima koji u danim grupama tvore parove entiteta određene istinosne vrijednosti. U ovom slučaju radi se o 4 entiteta iste klase u prvoj grupi, odnosno 2 u drugoj i 2 u trećoj grupi.

$$TP = \binom{4}{2} + \binom{2}{2} + \binom{2}{2} = 8 \tag{6}$$

Sada je jednostavno izračunati iz (5) i (6)

$$FP = P - TP = 19 - 8 = 11 \tag{7}$$

Ukupan broj negativnih odluka računa se preko broja kombinacija entiteta iz različitih grupa.

$$N = TN + FN = \ 5 * 4 + 5 * 3 + 4 * 3 = 47 \tag{8}$$

Broj lažno negativnih odluka je izračunljiv preko broja kombinacija entiteta u pogrešnim grupama sa svim preostalim entitetima te klase, odnosno

$$FN = 4 * 1 + 1 * 1 + 1 * 2 + 1 * 2 + 2 * 1 = 11 \tag{9}$$

Iz (8) i (9) slijedi

$$TN = N - FN = 47 - 11 = 36 \tag{10}$$

Sada je moguće izračunati sve osnovne evaluacijske mjere, a vrijednost Randovog indeksa je

$$\textit{Randov indeks} = \frac{TP+TN}{TP+FP+FN+TN} = \ \frac{8+36}{8+11+11+36} = 0{,}666 \tag{11}$$

6.3. Metodologija i rezultati

Budući da je potrebno eksperimentalno odrediti prag ispod kojeg će vijest biti označena kao prva, evaluacija je provedena metodom *holdout*. Ukupan skup uzoraka je podijeljen na dva podskupa. U ovom slučaju skup je podijeljen vremenski tako da je prvih pet dana činilo skup za učenje *train*, a drugih pet dana skup za ispitivanje *test*. Točna podjela i količina uzoraka u svakom skupu analogna je onoj prikazanoj u tablici 1.

Tablica 6 - Raspodjela uzoraka za evaluaciju

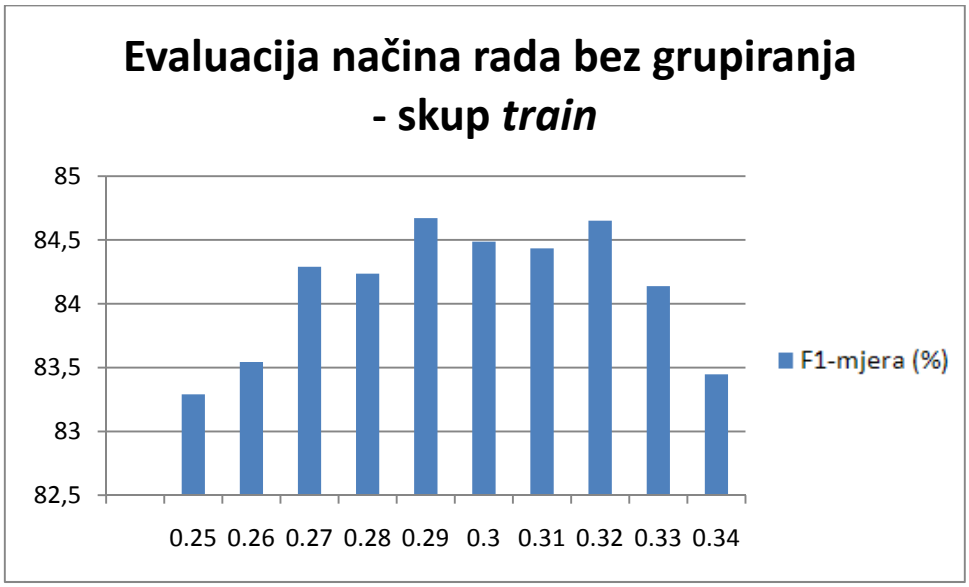
Skup	Količina uzoraka	Broj označenih prvih priča
train	498	196
test	534	215
Ukupni skup	1032	411

Osnovna ideja je određeni broj puta vršiti evaluaciju skupa *train* i pri tome svaki put mijenjati prag. Mjera koja će služiti kao dobrota nekog praga biti će F1-mjera. Kad je pronađen prag sa najboljom F1-mjerom, pomoću tog praga se evaluira skup *test* i tako se evaluira rad sustava.

Za početni prag odabrana je vrijednost 0,3. Ta vrijednost je odabrana nakon ručne provjere pouzdanosti nekoliko vijesti koje bi trebale biti označene kao prva vijest. Prag će biti mijenjan dodavanjem ili oduzimanjem višekratnika vrijednosti 0,01. Očekivano je da veći prag daje veći odziv (više vijesti će biti označeno kao prva vijest i stoga povećati broj FP sudova, a smanjiti broj FN sudova), dok manji prag daje veću preciznost (manje vijesti označeno kao prva vijest iz čega slijedi manji broj FP sudova i veći broj FN sudova). Idealan prag bi imao podjednako balansiranu preciznost i odziv pa je upravo zbog toga F1-mjera odabrana kao dobrota praga. Kada prag postane prevelik ili premalen biti će očit pad vrijednosti F1-mjere. Odluka o prestanku mijenjanja praga u određenom smjeru će biti donesena nakon što se primjeti stalan pad vrijednosti F1-mjere pri kretanju u tom smjeru. Nakon što je odlučeno prestati sa mijenjanjem praga u oba smjera, izabire se dotadašnji prag s najvećom F1-mjerom.

6.3.1. Evaluacija načina rada bez grupiranja

Način rada bez grupiranja evaluira se pomoću programa *FSDEval* uz dodatni parametar koji označava željeni prag budući da se na taj način odluka o prvoj vijesti može odgoditi do procesa evaluacije, tj. glavni program samo upisuje vrijednosti u `first_story_confidence` dok evaluacijski program upisuje vrijednosti u `first_story_classified`.



Slika 8 - Dobrota različitih pragova za način rada bez grupiranja

Slika 8 prikazuje dobrote, odnosne F1-mjere različitih pragova za skup train. Na apscisi su vrijednosti pragova, a na ordinati vrijednosti F1-mjere u postocima. Iz slike je vidljivo kako su najbolji pragovi sa vrijednostima 0,29 i 0,32, ali je teško vidljivo koji je bolji jer su jako slični. F1-mjera praga 0,29 je 84.67%, dok F1-mjera praga 0,32 iznosi 84,65%. Dakle, za idealan prag odabrana je vrijednost 0,29 i slijedi evaluacija podskupa *test*.

Tablica 7 - Tablica zabune za skup *test* uz prag=0,29

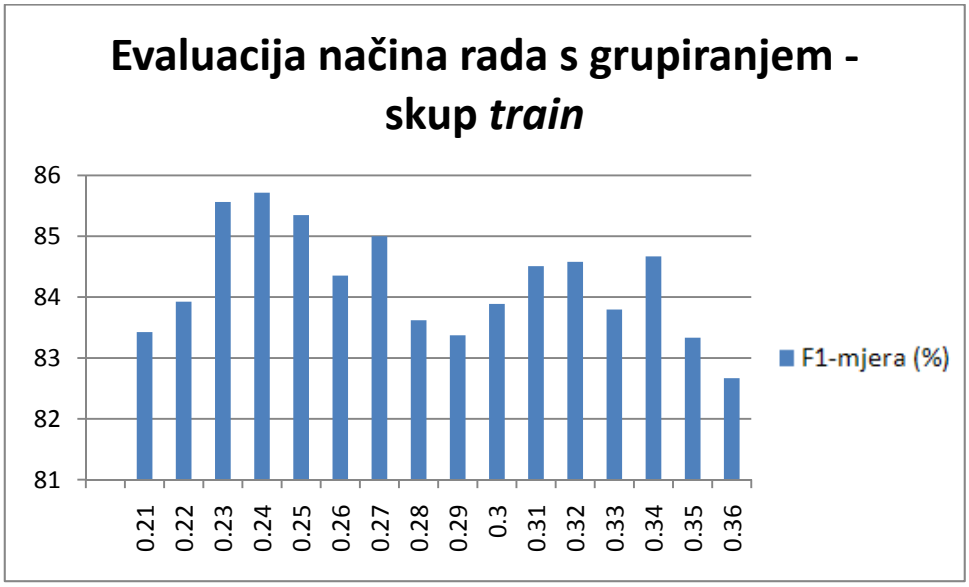
Predviđeno stanje, `first_story_classified`	Stvarno stanje, `first_story_real` TRUE	Stvarno stanje, `first_story_real` FALSE
TRUE	TP = 197	FP = 46
FALSE	FN = 19	TN = 272

Tablica 8 - Osnovne evaluacijske mjere za skup *test* uz prag=0,29

Evaluacijska mjera	Vrijednost (%)
Preciznost	81,0699
Odziv	91,2037
Specifičnost	85,5346
Točnost	87,8277
F1-mjera	85,8388

6.3.2. Evaluacija načina rada s grupiranjem

Način rada s grupiranjem također se evaluira pomoću programa *FSDEval*, ali bez dodatnog parametra koji predstavlja prag. Razlog tomu je što je odluku o prvoj vijesti, tj. pripadnosti određenoj grupi potrebno donijeti trenutačno nakon svake obrađene vijesti. Dakle u načinu rada s grupiranjem, glavni program puni elemente `first_story_classified` i `first_story_confidence` dok evaluacijski program samo računa tablicu zabune i osnovne evaluacijske mjere. FSD evaluacija se vrši kao i u načinu bez grupiranja.



Slika 9 - Dobrota različitih pragova za način rada s grupiranjem

Slika 9 prikazuje dobrote različitih pragova za skup *train*. Vidljivo je kako je potrebno evaluirati skoro dvostruko više pragova dok se vrijednosti dobrota ne stabiliziraju, ali je zato najviša dobrota veća od one iz načina bez grupiranja, a to je prag vrijednosti 0,24 sa vrijednošću F1-mjere 85,71%. Slijedi evaluacija skupa *test*.

Tablica 9 - Tablica zabune za skup *test* uz prag=0,24

		Stvarno stanje, `first_story_real`	
		TRUE	FALSE
Predviđeno stanje, `first_story_classified`	TRUE	TP = 185	FP = 31
	FALSE	FN = 20	TN = 298

Tablica 10 - Osnovne evaluacijske mjere za skup *test* uz prag=0,24

Evaluacijska mjera	Vrijednost (%)
Preciznost	90,2439
Odziv	85,6481
Specifičnost	93,7107
Točnost	90,4494
F1-mjera	87,8859

Ono što je specifično za slučaj grupiranja je stvaranje grupa kao direktorija nakon prolaska kroz glavni program. Stvara se jedan glavni direktorij koji se puni poddirektorijima od kojih svaki predstavlja jednu grupu, a poddirektoriji se pune XML-dokumentima koji predstavljaju vijesti. Ovom metodom se rješavaju FSD zadatak i zadatak otkrivanja skupine.

Nakon puštanja skupa *test* kroz glavni program u načinu rada s grupiranjem, stvoreno je 205 grupa, tj. onoliko grupa koliko je klasificirano prvih vijesti. Evaluacija grupiranja je izvedena evaluacijskim programom *FSDClusEval* koji prima lokaciju glavnog direktorija u kojem se nalaze grupe i radi na način opisan u poglavlju 6.2.

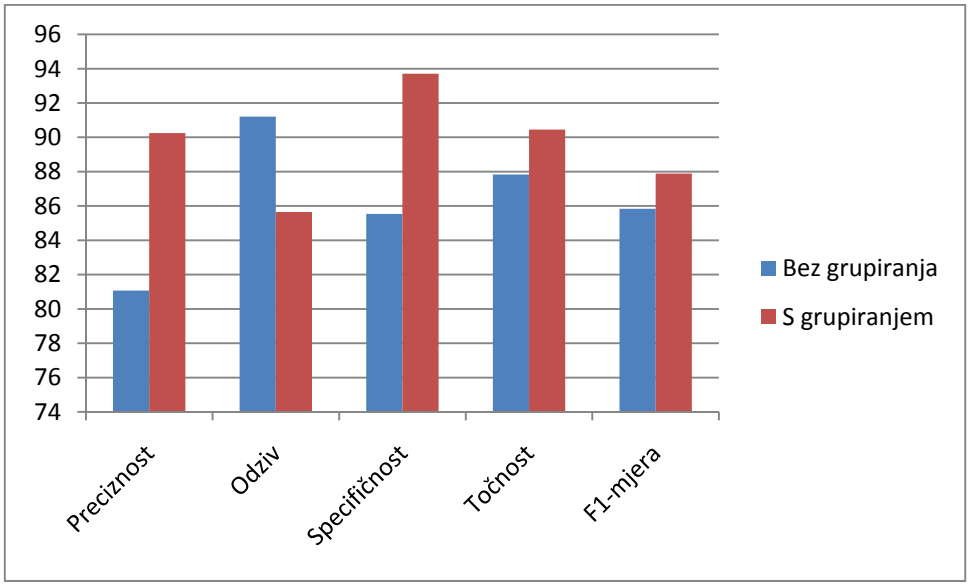
Tablica 11 - Tablica zabune za grupiranje skupa *test*

TP = 2833	FP = 246
FN = 896	TN = 138336

Tablica 12 - Osnovne evaluacijske mjere za grupiranje skupa *test*

Evaluacijska mjera	Vrijednost (%)
Preciznost	92,0104
Odziv	75,9721
Specifičnost	99,8225
Randov indeks	99,1975
F1-mjera	83,2256

6.3.3. Razmatranje rezultata



Slika 10 - Usporedba rezultata različitih načina rada

Oba načina rada su postigla iznimno dobre rezultate, a način rada s grupiranjem je općenito bolji od načina rada bez grupiranja kao što je pokazano na slici 10. Samo grupiranje je također vrlo uspješno s Randovim indeksom od 99,2% i F1-mjerom od 83,2% što pokazuje kako način rada s grupiranjem uspješno rješava i zadatak otkrivanja skupina.

Pristup inkrementalnog vektorskog modela se pokazao vrlo učinkovitim na relativno malenom skupu uzoraka, ali se pretpostavlja kako bi se na znatno većem skupu uzoraka ta učinkovitost smanjila. Naime, budući da svaki jezik ipak rukuje sa određenim ograničenim skupom lema stalno povećanje vektorskog modela će stvarati i mnogo sličnih vijesti, tj. šanse da će nova vijest biti dovoljno različita od ostalih će se smanjivati kako se model povećava.

Jedan od prvotnih pristupa TDT inicijative [6] objašnjen u poglavlju 2.1. je fundamentalno sličan ostvarenom načinu rada s grupiranjem, ali je za razvoj tog pristupa korišten ukupan broj uzoraka jednak 15863. Rezultat F1-mjere za taj sustav je tada bio 56% iz čega se može pretpostaviti kako bi i sustav ostvaren u ovom radu imao lošije rezultate nad većim brojem uzoraka.

Jedno očito rješenje ovog problema je održavanje vektorskog modela na nekoj konstantoj idealnoj veličini, a to povlači smanjivanje modela nakon određenog vremenskog intervala ili nekog drugog kriterija. Uz pretpostavku da nakon određenog vremenskog intervala neka vijest neće biti relevantna moguće je jednostavno obrisati vektor te vijesti iz modela. Jedan od kriterija za izbacivanje vijesti može biti i grupiranje, tj. ako se u neku grupu određeno vrijeme ne dodaju nove vijesti znači da je vjerojatno da ta tema nije više relevantna i moguće je izbaciti sve vijesti koje se nalaze u toj grupi iz vektorskog modela.

7. Zaključak

U radu su pokazani neki osnovni načini rješavanja problema otkrivanja prve vijesti u sklopu sustava otkrivanja i praćenja tema. Jedan od popularnijih postupaka, postupak inkrementalnog vektorskog prostornog modela, programski je ostvaren i prilagođen rješavanju problema nad vijestima na hrvatskom jeziku preuzetih s raznih internet portala. Taj postupak je proširen kako bi mogao rješiti i problem otkrivanja skupina pomoću grupiranja. Postupak se pokazao vrlo učinkovitim na relativno malom broju uzoraka, ali se pretpostavlja pad učinkovitosti s povećanjem broja uzoraka. Jedno od mogućih poboljšanja postupka je održavanje vektorskog modela na idealnoj konstantoj veličini.

U budućem radu bilo bi poželjno testirati programsko ostvarenje sa znatno većim brojem uzoraka kako bi se dokazala pretpostavka o padu učinkovitosti i provesti ručno označavanje s većim brojem sudaca kako bi se dokazala objektivnost problema označavanja prve vijesti i označavanja pripadnosti grupa.

Zbog eksponencijalnog rasta informacija, želja za učinkovitim TDT-sustavom postaje sve izraženija, a FSD-zadatak je jedan od ključnih zadataka TDT-sustava. Iz tog razloga danas se istražuje najbolje rješenje FSD-zadatka za razne izvore informacija i očekuju se nova poboljšanja.

8. Literatura

[1] Allan, James: „*Topic Detection And Tracking:Event-based Information Organization*“, Springer, 2002.

[2] „*TDT 2004: Annotation Manual*“, *Version 1.2,* 2004.

[3] Allan, James; Lavrenko, Viktor; Jin, Hubert: „*First Story Detection In TDT Is Hard“,* University of Massachusetts

[4] Allan, James; Lavrenko, Viktor; Malin, Daniella; Swan, Russell: „*Detections, Bounds And Timelines: Umass And TDT-3“*, University of Massachusetts

[5] Schultz,J. Michael; Liberman, Mark: „*Topic Detection And Tracking Using Idf-Weighted Cosine Coefficient“,* University of Pennsylvania

[6] Yang, Yiming; Zhang, Jian; Carbonell, Jaime; Jin, Chun: „*Topic-Conditioned Novelty Detection“*, Carnegie Mellon University, 2002.

[7] Wayne, Charles L.: „*Multilingual Topic Detection And Tracking: Successful Research Enabled By Corpora And Evaluation“,* Department of Defense

[8] Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich: „*Introduction To Information Retrieval“*, Cambridge University Press, 2008.

[9] www.jdom.org, 6.4.2010.

[10] http://math.nist.gov/javanumerics/jama, 6.4.2010.

[11] Bekavac, Božo: „*Rječnik Korpusne Lingvistike“*, Filozofski fakultet, Zagreb

[12] Šnajder, Jan; Dalbelo Bašić, Bojana; Tadić, Marko: „*Automatic Acquisition Of Inflectional Lexica For Morphological Normalisation“,* Information Processing and Management, vol.44, no. 5, pp 1720-1731, 2008.

[13] Dalbelo Bašić, Bojana: „*Dodatak: Nadzirano Učenje – Evaluacija“,* Fakultet elektrotehnike i računarstva, Zagreb, 2010.

[14] Ljubešić, Nikola: „*Pronalaženje događaja u višestrukim izvorima informacija“*, 2009.

[15] Vural, Ahmet: „*On-Line New Event Detection And Clustering Using The Concepts Of The Cover Coefficient-Based Clustering Methodology“*, Bilkent University, 2002.

[16] De, Indro: *„Experiments In First Story Detection“*, Ursinus College, 2005.

Sažetak

Otkrivanje prve vijesti u slijednim tekstnim podacima

Cilj otkrivanja i praćenja tema je razvitak tehnika koje automatski pretražuju, organiziraju i strukturiraju tekstne materijale iz raznih izvora vijesti. Jedan od pet osnovnih zadataka otkrivanja i praćenja tema je zadatak otkrivanja prve vijesti. Zadatak otkrivanja prve vijesti je prepoznavanje nove teme o kojoj se nije ranije raspravljalo. Ovaj rad pruža detaljniji uvid u zadatak otkrivanja prve vijesti i njegovog programskog ostvarenja. Zadatak je ostvaren pomoću inkrementalnog vektorskog prostornog modela na dva načina, bez grupiranja i s grupiranjem. Svi korišteni podatkovni uzorci su vijesti na hrvatskom jeziku prikupljene sa raznih internet portala. Uspješnost ostvarenja je evaluirana i uspoređena sa prvotnim ostvarenjem inicijative za otkrivanje i praćenje tema.

Ključne riječi: otkrivanje i praćenje tema, otkrivanje prve vijesti, pretraživanje informacija, vektorski prostorni model, tf-idf, grupiranje dokumenata

Abstract

First story detection in a text data stream

The objective of topic detection and tracking is to develop technologies that automatically search, organize and structure textual materials from a variety of broadcast news media. One of the five basic topic detection and tracking tasks is the first story detection task. The goal of the first story detection task is to recognize when a news topic appears that had not been discussed earlier. This paper provides a detailed insight into the first story detection task and its implementation. The task was implemented using an incremental vector space model with two different approaches, the clustering approach and the approach without clustering. All used data samples are Croatian language news collected from various internet portals. The performance of the implementation was evaluated and compared with the first implementation of the topic detection and tracking initiative.

Keywords: topic detection and tracking, first story detection, information retrieval, vector space model, tf-idf, document clustering