SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
IVANA LUČIĆA 3

Nikola Ljubešić

PRONALAŽENJE DOGAĐAJA U VIŠESTRUKIM IZVORIMA INFORMACIJA

Doktorski rad

Zagreb, 2009.

SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
IVANA LUČIĆA 3

Nikola Ljubešić

PRONALAŽENJE DOGAĐAJA U VIŠESTRUKIM IZVORIMA INFORMACIJA

Doktorski rad

mentor: prof. dr. sc. Damir Boras

Zagreb, 2009.

Poglavlje 1

Uvod

Informacijska revolucija koja se dogodila drugom polovicom prošlog stoljeća suvremenu je civilizaciju postavila pred novi izazov - polinomijalni rast količine pohranjenih informacija [Lyman and Varian, 2003]. Kako bi naša civilizacija mogla djelotvorno koristiti pohranjene informacije, potrebne su sve moćnije metode njihove automatske obrade.

Većina je informacija zapisana, odnosno kodirana prirodnim jezikom, tj. jezikom koji koriste ljudi za svakodnevnu komunikaciju. Za očekivati je da će tehnologije automatske obrade podataka zapisanih prirodnim jezikom u skoroj budućnosti biti među vodećim tehnologijama.

Potrebu za automatskom obradom takve vrste podataka znanstvenici su primijetili na samom početku ere automatske obrade podataka, i to prvenstveno kroz problem strojnog prevođenja [Weaver, 1955, Bar-Hillel, 1960].

Područje koje se danas bavi problemom obrade podataka kodiranih prirodnim jezikom naziva se obrada prirodnog jezika (engl. *natural language processing*), odnosno kraće OPJ (engl. *NLP*). Glavni izazov u obradi prirodnog jezika jest njegova višeznačnost koja ujedno predstavlja i osnovnu razliku prirodnih prema formalnim jezicima. Formalne jezike ljudi oblikuju kako bi jednoznačno komunicirali s računalima pa je osnovni napor područja obrade prirodnog jezika formalizacija prirodnog jezika bez gubitka informacije da bi on bio strojno obradiv.

Vrlo blisko područje obradi prirodnog jezika, više usmjereno na lingvistiku

nego na računalstvo, naziva se računalna lingvistika (engl. *computational linguistics*). Zanimljiva slika ovog područja prikazana u [Tadić, 2003] jest ona dviju sigurnih znanstvenih obala - lingvistike i računalstva te računalne lingvistike koja plovi točno između tih dviju obala.

Interdisciplinarno područje koje uključuje i problem obrade prirodnog jezika jest informacijska znanost (engl. *information science*) čiji je osnovni fenomen proučavanja informacija. Važno područje unutar informacijske znanosti jest područje upravljanja znanjem (engl. *knowledge management*) koje se redovito oslanja na metode obrade prirodnog jezika.

Ova disertacija istražuje mogućnost organiziranja znanja višestruko komuniciranog preko mrežnih izvora poput novinskih portala. Zbog prevelikog broja mogućih izvora informacija i nemogućnosti pojedinca da sve te izvore prati, mogućnost organiziranja danih informacija u strukture koje omogućuju jednostavnu percepciju svih diseminiranih informacija vrlo je privlačna.

Područje koje se unutar obrade prirodnog jezika bavi problemom uspješnog praćenja i organiziranja višestrukih izvora informacija se naziva pronalaženje i praćenje teme (engl. *topic detection and tracking*).

Nazvano je prema najvećoj zajedničkoj inicijativi, a to je TDT, tj. *Topic Detection and Tracking* projekt koji je trajao od 1998. do 2004. kao zajednički projekt više američkih sveučilišta i tvrtki. Glavne su zadaće tog projekta, a time i novo određenog područja sljedeće [Allan, 2002]:

1. segmentacija novosti (engl. *story segmentation*)
2. pronalaženje prve novosti (engl. *first story detection*
3. pronalaženje grozdova (engl. *cluster detection*)
4. praćenje (engl. *tracking*)
5. pronalaženje veza između novosti (engl. *story link detection*)

O navedenom će projektu biti više govora u poglavlju 1.3 koje se bavi prethodnim istraživanjima u području od interesa za ovaj rad.

U okviru ovog doktorskog rada razrađuje se problem organizacije znanja komuniciranog novinskim portalima, i to kroz pronalaženje pojedinih

događaja, odnosno grupiranje dokumenata prema događaju o kojem izvještavaju. O samom pojmu događaja više će biti govora u poglavlju 1.1, dok će u poglavlju 1.2 pobliže biti opisana problematika pronalaženja događaja. Završno će u uvodnom poglavlju poglavlja 1.3 biti govora o dosadašnjim rezultatima istraživanja u području pronalaženja događaja.

1.1 Što je to događaj

Bez skupa pravila koja bi odredila sadržaj pojma događaj (engl. *event*), pojedinci bi zasigurno imali različito *a priori* shvaćanje tog pojma. Sljedeća definicija je preuzeta iz pravila TDT inicijative, prvog i jedinog organiziranog znanstvenog napora u istraživanju pronalaženja događaja u višestrukim izvorima informacija. Prema njoj, događaj je "nešto posebno što se dogodilo u određeno vrijeme" [Allan et al., 1998a]. U ovoj definiciji razlikujemo dva važna svojstva događaja:

- svojstvo posebnosti - događaj je nešto posebno, nešto što je moguće razlikovati od ostaloga
- svojstvo vremena - događaj je vezan uz određeni trenutak

Postoje druge definicije događaja poput Popperove [Popper, 1968] koja predlaže da se "pad zrakoplova" smatra događajem, dok bi "pad zrakoplova US Air 427" bila pojava (engl. *occurence*). Popperova je definicija bliska onoj znanstvenika i filozofa koji tvrde da do događaja dolazi kad se pojavljuje sukob između fizičkih objekata. Filozofi često zaključuju da su dva događaja identična ako imaju istu prostorno-vremensku povijest te ako imaju iste uzroke i posljedice. U [Lombard, 1986] se raspravlja zašto ta svojstva nisu dovoljni uvjeti za pronalaženje događaja. Predstavlja se model za događaje koji uključuje aspekt promjene definirane kao "dodatak, odnosno gubitak svojstva".

Prvi ozbiljniji napori rasprave o fenomenu događaja i njemu bliskim pojmovima sa stanovišta višestrukih izvora informacija i automatske obrade podataka učinjeni su u disertaciji Rona Papke [Papka, 1999] koja općenito

do velike mjere opisuje rezultate TDT2 inicijative. On osnovno razlikuje događaj i temu (engl. *topic*) te je po njemu svojstvo vremena ono što razlikuje ta dva pojma. Primjer događaja koji navodi jest "pronađen računalni virus u British Telecomu 3. ožujka 1993." dok se temom kojoj taj događaj pripada po njemu može smatrati tema "računalni virusi". Dalje navodi da se u TDT2 inicijativi ta originalna podjela na događaj i temu dodatno proširuje te se razlikuju sljedeće definicije [Doddington, 1999]:

- tema - skup direktno povezanih utjecajnih događaja ili aktivnosti
- događaj - nešto što se dogodilo u određeno vrijeme na određenom mjestu (nesreće, zločini, prirodne katastrofe)
- aktivnost (engl. *activity*) - povezani skup radnji koje imaju zajednički fokus ili svrhu (istrage, suđenja, akcije spašavanja poslije prirodnih katastrofa)

Događaj i aktivnost razlikuje upravo isto ono što razlikuje događaj i temu - stupanj vremenske određenosti. Dok događaj i dalje ostaje vremenski znatno određen, aktivnost tu određenost djelomično napušta. Pitanje koje se potom prirodno postavlja jest razlika između teme i aktivnosti. Razlika leži u tome što je aktivnost i dalje atomarna, odnosno ima svojstvo posebnosti, dok to svojstvo tema kao skup bliskih događaja nema. Uvođenjem pojma aktivnosti u TDT2 inicijativi priznaje se kompleksnost pojma događaja te problematičnost svojstva trenutnosti, redovitog problema u određivanju granica pojedinog događaja.

U posljednjoj verziji TDT inicijative - TDT5, definicije su bitno elaboriranije [TDT, 2004]. Događajem se smatra "nešto posebno što se dogodi u određeno vrijeme na određenom mjestu, zajedno sa svim preduvjetima i nezaobilaznim posljedicama". Aktivnost se smatra "povezanim skupom događaja koji imaju zajednički fokus ili cilj te se događaju u određeno vrijeme na određenom mjestu". Temom se smatra "neki događaj ili aktivnost zajedno sa svim direktno povezanim događajima i aktivnostima". U ovim je definicijama moguće primijetiti da se uz svojstvo posebnosti i vremena uvodi

i svojstvo prostora. Dakle, u TDT5 inicijativi važna svojstva događaja su sljedeća:

- svojstvo posebnosti - događaj je nešto posebno, nešto što je moguće razlikovati od ostaloga
- svojstvo vremena - događaj je vezan uz određeni trenutak
- svojstvo prostora - događaj je vezan uz određeni prostor

Isto je tako moguće primijetiti proširenje definicije događaja preduvjetima i posljedicama tog događaja. Primjer koji se navodi u [TDT, 2004] je kidanje žice gondole u skijaškom odmaralištu Cavelese u Italiji od strane američkog vojnog zrakoplova. Događajem se, dakle, može smatrati taj događaj te njegova posljedica - 20 mrtvih. Proširenom definicijom događaja, dakle, taj je događaj podijeljen na sljedeće elemente:

- preduvjet - niski let američkog vojnog zrakoplova
- svojstvo posebnosti - kidanje žice gondole od strane vojnog zrakoplova
- svojstvo vremena - 3. veljače 1998.
- svojstvo prostora - skijaško odmaralište Cavelese, Italija
- posljedica - 20 mrtvih

Kako bi se od događaja došlo do teme, potrebno je pronaći druge događaje koji su direktno vezani na ovaj. To bi u ovom slučaju bilo spašavanje preživjelih, pogrebi žrtava, izjave američkih marinaca o pravilnicima treniranja u nastanjenim područjima te istraga koja je slijedila.

Pitanje koje se postavlja jest gdje je potrebno povući granicu teme. Taj je zrakoplov, primjerice bio u sklopu kontingenta koji je redovito patrolirao nad Bosnom i Hercegovinom. Je li moguće povezati ovaj događaj s ratom u Bosni? Odgovor je negativan iz razloga što nesreća prouzročena vojnim avionom i rat u Bosni u kojem je taj avion sudjelovao nisu direktno povezani.

Od inicijative TDT3 razlikuje se i 13 različitih tipova događaja [TDT, 2004]. To su:

1. izbori
2. skandali, saslušanja
3. pravni slučajevi
4. prirodne katastrofe
5. nesreće
6. sukobi i ratovi
7. znanost i otkrića
8. financijske novosti
9. novi zakoni
10. sportske novosti
11. politički i diplomatski susreti
12. poznate ličnosti
13. razno

U idućem se poglavlju definira osnovni problem kojim se ovaj doktorski rad bavi - pronalaženje događaja.

1.2 Pronalaženje događaja

Pronalaženje događaja u literaturi opisuje dva srodna, no različita zadatka:

1. analiza dokumenata kako se objavljuju te pokušaj identificiranja dokumenta koji govori o novom događaju
2. analiza svih objavljenih dokumenata te njihovo grupiranje ovisno o tome koji događaj opisuju.

Može se zaključiti da i u jednom i u drugom slučaju postupak otkriva, odnosno identificira događaje uz razliku što se u drugom zadatku događaj otkriva po kriteriju sličnosti između dokumenata dok se u prvom pristupu događaj otkriva po kriteriju različitosti trenutno objavljenog dokumenta u odnosu na sve do tada objavljene dokumente. Ovaj rad istražuje drugi pristup. Oba će pristupa biti detaljnije prikazana u poglavlju 1.3 koje govori o prethodnim istraživanjima u ovom području.

Pronalaženje događaja, dakle, kao zadatak ima skup dokumenata organizirati u grupe na takav način da dokumenti u jednoj grupi opisuju jedan, isti događaj.

Kako bi se moglo uspoređivati dokumente, odnosno događaje koje oni opisuju potrebno je oblikovati model pojedinog događaja. Papka u [Papka, 1999] razmatra mogućnost definiranja takvog modela događaja u višestrukim izvorima informacija koji bi trebao omogućiti usporedbu dvaju događaja te dokazivanje njihove identičnosti.

S novinarskog stajališta novost će u sebi sadržavati sljedeće elemente [Mayeux, 1996]:

1. kada se događaj dogodio
2. tko je bio umiješan
3. gdje se dogodio
4. kako se dogodio
5. utjecaj, važnost ili posljedice događaja za ciljanu publiku

To su informacije koje čitatelji očekuju od novinara. Neka od ovih svojstava, primjerice odgovor na pitanje kada?, kako?, gdje? moguće je izvući iz teksta raznim metodama obrade prirodnog jezika, no uglavnom je problem što odgovori na ta pitanja redovito nisu eksplicitno navedeni u novostima. Tako, primjerice, u jednoj novosti koja govori o potresu u Japanu 1995. godine ni jednom se riječju ne precizira da se radi o potresu već se kroz dokument referira na "najgoru katastrofu u povijesti Japana".

Isto tako, kako se neki događaj, odnosno aktivnost razvija za očekivati je da će se odgovori na gore postavljena pitanja također mijenjati. Tako je, primjerice, u slučaju bombaškog napada u Oklahoma Cityju u 1995. prvo bilo izvještavano o spašavanju na mjestu događaja. Potom je fokus novinara prešao na pretpostavku da se radi o napadu islamističkih ekstremista. Kako su počinitelji uhvaćeni, započeta je rasprava o radikalnim desnim milicijama u Sjedinjenim američkim državama. Pitanje je bi li se ovaj događaj mogao smatrati jedinstvenim događajem. Cijela rasprava oko uzroka ovakvog napada kao i izvještavanje o suđenju počiniteljima zasigurno bi oblikovali temu, no barem neki od gore navedenih elemenata morali bi se smatrati događajem koji zbog izrazite važnosti i nepredvidivosti cijele situacije u više navrata doživljava sadržajni zaokret. Iz tog se razloga navedeni događaj zasigurno ne može opisati odgovorima na prije navedena pitanja.

Papka zaključuje da će općenito isti događaj u raznim novostima biti opisan na drugačiji način te da jednostavno uspoređivanje riječi koje se pojavljuju u dokumentima neće za to biti dovoljno. Isto tako zaključuje da postoje neka svojstva koja se mogu jednostavno modelirati, primjerice vrijeme.

U ovom su poglavlju spontano uvedena tri termina koji zahtijevaju definiciju kako bi bili jednoznačno shvaćeni. To su događaj, novost i dokument. Njihove su definicije sljedeće:

- događaj je promjena u stanju stvarnog svijeta
- novost je jezični iskaz koji opisuje događaj
- dokument je fizički nosilac jezičnog iskaza, odnosno novosti

U ovom se radu postavlja jednostavan model mrežne novosti objavljene na novinskom portalu koji će biti ishodište za daljnji rad na zadatku pronalaženja događaja. Papka je u prethodno opisanoj raspravi dokazao da zapravo niti čovjek nije u mogućnosti novost, odnosno događaj koji ona opisuje, strukturirati na novinarski način, te da je usto i usporedba tako strukturiranih novosti vrlo problematična. Valja suprotno tome primijetiti što u pravilu svaka mrežna novost posjeduje:

- naslov novosti
- tekst novosti
- vrijeme objave dokumenta
- mjesto objave dokumenta

Prikazana struktura ujedno je i osnovni model dokumenta kakav se koristi u ovom radu.

Tekst i naslov pojedine novost valja razlikovati zato što je za očekivati da naslov predstavlja sažetu informaciju o sadržaju cijelog članka. Naravno, najbolji dokaz da to nije uvijek istina je žuto novinarstvo.

Vrijeme objave dokumenta je također vrlo bitno zato što je, kako je u prošlom poglavlju rečeno, jedno od bitnih svojstava događaja i njegovo vrijeme zbivanja. Za pretpostaviti je da će novosti koje opisuju neki događaj, ovisno o relevantnosti događaja po zajednicu, biti objavljene u slično vrijeme.

Na kraju, mjesto objave neke novosti, misleći pritom na novinski portal koji je objavio novost, može također biti vrijedan podatak iz razloga što je intuitivno da isti novinski portal neće izvještavati o istom događaju više puta. I ovdje se može pretpostaviti da će o nekom važnom te kompleksnom događaju neki novinski portal izvještavati više od jedanput.

Dva posljednja elementa osnovnog modela dokumenta, dakle vrijeme i mjesto objave dokumenta, ujedno su i poticaj na dvije potencijalne heuristike koje se među ostalima istražuju u ovom radu. To su:

1. dokumenti koji izvještavaju o istom događaju bit će objavljeni u slično vrijeme
2. jedan novinski portal neće izvještavati više puta o istom događaju

Daljnja rasprava o ovim heuristikama kao i njihova analiza i testiranje se vrše u poglavlju 3.1.4.

Do sada se pretpostavljalo da je funkcija preslikavanja iz skupa događaja na skup novosti te na skup dokumenata bijekcija. Je li to uvijek istina? Svaki će pojedinac poslije kraćeg razmišljanja moći ustvrditi da je odgovor

na pitanje negativan. Naime, događaji se isprepliću, jedni su uzroci drugima i slično te je najprirodnije za očekivati da će i u izvještavanju o tim događajima u pojedinim novostima, odnosno dokumentima biti govora o više povezanih događaja. Isto se tako ne može očekivati da će u svakom dokumentu biti izviješteno o cijelom događaju. Ukratko, odnos elemenata skupova događaja, novosti i dokumenata nije moguće direktno opisati funkcijom već težinskim grafom gdje svaka veza prikazuje vezu i težinu te veze određenog elementa jednog skupa s elementom drugog skupa.

U ovom će se doktorskom radu zbog potrebe za formalizacijom problema događaja i njegovog fizičkog opisa u obliku dokumenta pretpostaviti da se jednim dokumentom bilježi jedna novost koja opisuje jedan događaj. To će sa sobom, naravno, povući i određene probleme. Više o problemima koji se pojavljuju pri pokušaju organizacije dokumenata u grupe gdje svaka grupa sadrži dokumente koji opisuju određeni događaj bit će rečeno u poglavlju 3.1.2.

Osnovni je zadatak ovog rada osmisliti način kako izvesti dijeljenje novosti na grupe tako da u svakoj grupi budu novosti koje izvještavaju o nekom događaju. Taj problem zapravo predstavlja problem klasifikacije dokumenata u nepoznati broj klasa.

U računalnoj znanosti, odnosno strojnom učenju, klasifikacija može biti nadzirana ili nenadzirana. Razlika između ta dva pristupa jest u tome ima li algoritam za klasifikaciju primjere na kojima može naučiti razlikovati elemente klasa ili ne.

U slučaju pronalaženja događaja zasigurno ne postoje primjeri svakog događaja iz kojega bi algoritam mogao naučiti kako prepoznati novost koja govori o tom događaju. Dakle, potrebno je raditi nenadziranu klasifikaciju poznatu pod imenom grožđenje (engl. *clustering*). Grupe koje nastaju metodom koja ne zahtijeva primjere za učenje nazivaju se grozdovi (engl. *clusters*).

Algoritmi za grožđenje dijele se na hijerarhijske i parcijalne algoritme. Glavna razlika među njima jest je li unaprijed poznat broj grupa, odnosno grozdova. Kako se u ovom zadatku radi o dijeljenju skupa dokumenta na nepoznati broj grupa, a to uključuje da nama naime nije *a priori* poznat broj

događaja opisan u skupu dokumenata, za ovaj je zadatak potrebno koristiti hijerarhijske algoritme za grožđenje.

Algoritmi za grožđenje kao ulaz primaju tzv. matricu sličnosti (engl. *similarity matrix*) u kojoj je navedena sličnost svaka dva elementa, odnosno dokumenta.

Kako bi se izračunala matrica sličnosti potrebno je moći izračunati sličnost dviju točaka, odnosno dvaju dokumenata. Kako bi se dokumenti mogli uspoređivati potrebno je formalizirati njihov prikaz, odnosno naš trenutni model dokumenta pretvoriti u stvarni zapis. U praksi se takvi zapisi najčešće prikazuju kao vektori svojstava.

Pri formalizaciji dokumenata redovito se koristi više metoda iz područja obrade prirodnog jezika, odnosno formalizira se informacija zapisana na određenoj razini lingvističke apstrakcije. Tako je, primjerice, jedan od najjednostavnijih prikaza dokumenata prikaz skupa riječi koje se pojavljuju u tekstu. Nešto kompleksniji je onaj koji uzima u obzir nizove riječi, posebno one čije je supojavljivanje statistički značajno. Nadalje, u formalizirani zapis se mogu uključiti morfosintaktičke informacije poput vrste riječi i drugih morfosintaktičkih kategorija. Mogu se zapisivati i sintaktičke, semantičke, pragmatičke, diskursne i druge informacije.

Nakon što su dokumenti formalno prikazani, moguće je preko odabrane funkcije sličnosti (engl. *similarity function*) izračunati sličnost svaka dva dokumenta.

O grožđenju će općenito više riječi biti u poglavlju 2, dok će o algoritmima grožđenja za pronalaženje događaja više govora biti u poglavlju 2.4.

1.3 Prethodna istraživanja

Većina istraživanja provedena do danas u području pronalaženja događaja je zasigurno učinjena u sklopu TDT inicijative [Allan et al., 1998a, Allan et al., 1998b, Yang et al., 1998, Papka and Allan, 1998, Papka, 1999, Doddington, 1999, Allan, 2002, TDT, 2004]. Ta inicijativa kao osnovni zadatak ima organiziranje novosti objavljivanih iz višestrukih izvora na osnovi pojma događaja. U njenom provođenju je sudjelovalo više od deset sveučilišta

i tvrtki financiranih prvenstveno od vlade Sjedinjenih američkih država.

Izvori informacija koje TDT koristi su televizija, radio i Internet. Kako su izvori djelomično tekstualni, a djelomično auditivni, inicijativa koristi i sustave za pretvaranje govora u tekst (engl. *speech-to-text systems*). Isto tako inicijativa pretpostavlja da izvori informacija budu višejezični čime se javlja potreba za strojnim prevođenjem. U kasnijim fazama inicijative u istraživanja su uvedeni arapski i kineski tekstovi te je korišten sustav SYSTRAN [Systran, 2009] za njihovo prevođenje na engleski. Generalno je cilj TDT sustava podatkovni tok (engl. *data stream*) iz izvora rastaviti na pojedine novosti, prepoznavati novosti koje opisuju nove događaje te organizirati novosti u grupe koje predstavljaju pojedini događaj, odnosno grupu događaja.

Glavne su zadaće te inicijative sljedeće [Allan, 2002]:

1. segmentacija novosti (engl. *story segmentation*) - problem dijeljenja transkripta radijskih ili televizijskih novosti na pojedine novosti

2. pronalaženje prve novosti (engl. *first story detection* - problem pronalaženja novosti koja izvještava o novom događaju

3. pronalaženje grozdova (engl. *cluster detection*) - problem grupiranja svih novosti koje izvještavaju o nekom događaju, odnosno temi

4. praćenje (engl. *tracking*) - zahtijeva praćenje toka informacija kako bi se pronašle novosti koje govore o određenom događaju, odnosno temi

5. pronalaženje veza između novosti (engl. *story link detection*) - donošenje odluke izvještavaju li dvije novosti o istom događaju, odnosno temi

Moguće je primijetiti kako ove zadaće nikako nisu međusobno isključujuće te da često riješavaju isti ili sličan problem s drugog kuta gledišta.

U pronalaženju grozdova koje je najzanimljivije za ovo istraživanje osnovna razlika u TDT projektu u odnosu na klasični zadatak grožđenja je evaluacija. U grožđenju se algoritmi ne kažnjavaju drastično ako izostave

podatkovnu točku prvo uvedenu u uzorak. Vremenska komponenta podatkovnih točaka često uopće ne postoji. U prepoznavanju nove teme prvi je događaj iznimno bitan za potpunost te teme. Kod grožđenja događaja evaluacija je sličnija klasičnom zadatku grožđenja. U TDT projektu provodi se tvrdo grožđenje što znači da svaka novost može pripadati samo jednom grozdu unatoč činjenici da postoje slučajevi kada se u nekoj novosti izvještava o više događaja, odnosno tema.

Nadalje je specifično za TDT da radi grožđenje na vezi (engl. *on-line clustering*), a ne retrospektivno (engl. *retrospective clustering*) što znači da se podatkovni tok sluša cijelo vrijeme te kako se pojavi neka novost, ona se dodaje u već postojeću strukturu grozdova.

U kontekstu izbora algoritma za grožđenje, istraživanja u TDT-u redovito najbolje rezultate postižu algoritmom grožđenja na vezi, pojedinačne veze (engl. *single-link*) jednim prolaskom (engl. *single-pass*) koji koristi vremensku komponentu [Papka, 1999]. Vremenska se komponenta uključuje u klasifikatore kao parametrirana komponenta praga pridruživanja nepoznatog dokumenta grozdu te redovito pokazuje bolje rezultate od klasifikatora s parametrom 0.

U skopu TDT inicijative primjenjeni su razni pristupi prikazani od uspješnijih prema onima manje uspješnima:

- tvrtka BBN koristi inkrementalni k-sredina algoritam te eksperimentira s probabilističkim i vektorskim mjerama sličnosti [Walls et al., 1999]
- grupa na Sveučilištu u Massachusettsu koristi algoritam jednim prolaskom koji iskorištava i vremensku komponentu [Papka, 1999]
- tvrtka Dragon Systems koristi betabinomijalni model miješanja koji daje skoro identične rezultate kao multinomijalni [Lowe, 1999]
- tvrtka IBM koristi vektorski prostor za prikaz dokumenata te eksperimentira s POS označavanjem, korjenovanjem i određivanjem svojstava unigrama i digrama imenica, također koristeći TF-IDF mjeru [Dharanipragada et al., 1999]

- grupa na Sveučilištu Pensilvanije koristi aglomerativni algoritam pojedinačne veze i TF-IDF mjeru te ne koristeći pri tome korjenovanje za razliku od ostalih pristupa [Schultz and Liberman, 1999]

- grupa na Sveučilištu Carnegie Mellon koristi inkrementalni aglomerativni algoritam i TF-IDF mjeru uspoređujući samo dokumente koji se pojavljuju u određenom vremenskom prostoru čime postižu malo poboljšanje rezultata [Carbonell et al., 1999]

Osim istraživanja u sklopu TDT inicijative već više desetljeća postoje mnoga istraživanja u području grožđenja, odnosno grožđenja dokumenata, posebno u području pretraživanja informacija [van Rijsbergen, 1979, Voorhees, 1986, Willett, 1988, Salton, 1988]. Najdetaljnije istraživanje u području grožđenja dokumenata s naglaskom na metode formalizacije dokumenata metodama obrade prirodnog jezika je doktorska disertacija Richarda Forstera "Document Clustering in Large German Corpora Using Natural Language Processing" [Forster, 2006]. On uspoređuje dva pristupa - jedan kojim OPJ metodama pojednostavljuje prikaz dokumenta te drugi kojim ga proširuje. Cijeli niz metoda evaluira na pet različitih skupova dokumenata čime ova disertacija zaslužuje epitet najsustavnijeg eksperimenta u području grožđenja dokumenata.

U pristupu u kojem pojednostavljuje prikaz dokumenta postavlja kao osnovni pristup (engl. *baseline*) prikaz dokumenta korjenovanjem (engl. *stemming*) i uklanjanjem stop riječi (engl. *stop-word removal, stopping*). Tako obrađene pojavnice prikazuje kao vreću riječi (engl. *bag of words*). On u doktorskom radu razlikuje lingvističke od statističkih metoda koseći se pritom s naslovom iz razloga što obje grupe metoda pripadaju OPJ metodama. Usto, za raspravu je koliko primjerice korjenovanje ili uklanjanje stop-riječi nisu lingvističke metode.

Prva metoda koju uspoređuje sa svojim osnovnim pristupom je lematizacija, odnosno POS označavanje čiji rezultat prikazuje kao vreću lema (engl. *bag of lemmata*). Njegovi rezultati ukazuju na to da je vreća lema prikladniji način prikaza dokumenata za zadatak grožđenja dokumenata od same vreće riječi, no da u odnosu na korjenovani tekst ne predstavlja znat-

no poboljšanje. Zato zaključuje da jednostavnija metoda korjenovanja kao predobrade lematizaciju čini suvišnom. Uz to, lematizacija je pojednostavila prostor svojstava za 15-23 posto, dok je korjenovanje to učinilo za 25-30 posto. Kroz to pojednostavljenje prostora svojstava moguće je pravdati i trošak predobrade dokumenata. Na kraju autor zaključuje da lematizacija sama po sebi ne daje bolje rezultate, no da se ne smije zanemariti kao mogući korak u kompleksnijim metodama koje bi na taj način mogle biti poboljšane.

Podrezivanje (engl. *pruning*) kao statistička metoda pokazala se prihvatljivom metodom u slučaju da se primjenjuje vrlo ograničeno te time pojednostavi model za otprilike 25 posto. Daljnje podrezivanje znatno šteti rezultatima. Isto je tako primijećeno da se ovisno o skupu podataka podrezivanje može provoditi različito drastično bez narušavanja rezultata.

Uklanjanje stop riječi uspješno smanjuje matricu, u pravilu bez narušavanja rezultata. Statističke metode pronalaženja stop riječi pokazuju u primjeni bolje rezutate od onih lingvističkih. Moguće je primijetiti da u nekim slučajevima uklanjanje stop riječi narušava rezultat te iz tog razloga treba oprezno pristupiti primjeni ove metode.

Eksperimentiranje s odabirom svojstava po vrsti riječi Forstera dovodi do sljedećih zaključaka:

- imenice su najbrojnija i najvažnija svojstva
- pridjevi su po važnosti drugi po redu
- osobna imena su po važnosti sljedeća, no ponekad ne pokazuju napredak
- glagoli, apozicije i brojevi nisu od pomoći te ponekad i kvare rezultate
- glagoli su unatoč tome što su otvorena klasa vrlo općeniti te u njima nije direktno kodirano značenje kao u imenicama i pridjevima te stoga čak i glagoli s uklonjenim modalnim glagolima ne pokazuju napredak

Zaključno, pokazano je da odabir imenica i pridjeva ili imenica, pridjeva i osobnih imena najviše popravlja rezultat uz znatnije reduciranje matrice nego što se to čini uklanjanjem stop riječi.

Eksperimentirano je i s težinskim faktorima te je pokazano da je umjesto odabira svojstava preko vrste riječi bolje povećati težinski faktor imenicama i pridjevima. Smanjivanje težinskog faktora stop riječi nije pokazalo napredak.

Morfološke metode obogaćivanja prikaza dokumenta poput rastavljanja složenica, vrlo čestih u njemačkome, pokazuju poboljšanje rezultata, no ne veće od lematizacije ili korjenovanja.

Jednostavnije sintaktičke metode kao što su uključivanje svojstava poput digrama, složenih osobnih imena i imeničnih fraza u pravilu ne popravljaju rezultate više nego što to čine jednostavnije metode uz redovitu iznimku skupa dokumenata koji su dulji i redovito kompleksniji.

Eksperimenti sa semantičkim mrežama pokazuju da takve intervencije skoro uvijek kvare rezultate što navodi na zaključak da su općenito leksičkosemantički izvori nedovoljno jasni i određeni te da se od takvih resursa bez kritičnijeg oblikovanja resursa kao i njihove primjene ne mogu očekivati poboljšanja u zadacima OPJ-a.

Generalno Forster zaključuje da jednostavnije metode redovito donose jednako ili veće poboljšanje u odnosu na one kompleksnije. Za pretpostaviti je da je uzrok toga nedorađenost i neupoznatost s fenomenima koje ti pristupi pokušavaju modelirati.

Za razliku od Forsterovog istraživanja grožđenja dokumenata, takvo istraživanje za problem pronalaženja događaja ne postoji. Najbliži pokušaj tome predstavlja doktorski rad Rona Papke [Papka, 1999], no njegov je doktorski rad ipak pretežno kompilacija rezultata niza istraživanja nad različitim uzorcima podataka. Ovaj doktorski rad pokušava provesti sistematično istraživanje niza varijabli na identičnom uzorku.

Zanimljiv pristup pronalaženju događaja prikazan je u [Li et al., 2005] gdje se istraživanje vrši nad povijesnim korpusom novinskih tekstova, odnosno vrši se retrospektivno pronalaženje još neotkrivenih događaja (engl. *RED - retrospective event detection*) kroz sustav nazvan HISCOVERY.

Klasična tržišna primjena metoda pronalaženja događaja je prikazana u [Wei and Lee, 2004] gdje se pronalaženje događaja koristi za pregledavanje okoline (engl. *environmental scanning*), odnosno prikupljanje informacija o okolini nekog poslovnog subjekta na temelju kojih se razrađuju taktike i

strategije daljnjih poteza poslovnog subjekta.

Zaključno se može reći kako je TDT inicijativa najdetaljniji pristup istraživanju problema pronalaženja događaja, no da ono još nije doživjelo dovoljno sistematičnih istraživanja koja bi ponavljanjem rezultata na različitim uzorcima pružila jasniji uvid u problem. Za razliku od pronalaženja događaja, grožđenje dokumenata bolje je, no i dalje nedovoljno istraženo. Opći je zaključak da, kako u grožđenju dokumenata, tako i u pronalaženju događaja, jednostavnije nelingvističke metode postižu bolje rezultate od onih kompleksnijih i lingvističkih.

U idućem će poglavlju detaljnije biti prikazan proces grožđenja - metoda koja je središnja u širem zadatku pronalaženja događaja.

Poglavlje 2

Grožđenje

Grožđenje (engl. *clustering*) je pristup nenadzirane klasifikacije entiteta prikazanih podatkovnim točkama (engl. *data points*), odnosno vektorima svojstava (engl. *feature vectors*)) u grupe zvane grozdovi (engl. *clusters*). Kao metoda se primjenjuje u različitim stručnim i znanstvenim područjima. Osnovna mu je zadaća, kao i kod većine metoda analize podataka, dvojaka:

1. istražujuća - služi prikazu podataka koji istraživaču daje određeni uvid u podatke te mu omogućuje oblikovanje heuristika

2. potvrđujuća - rezultat metode grožđenja se koristi pri odlučivanju, primjerice kako organizirati neki skup dokumenata.

Metoda grožđenja grupira dakle entitete najčešće prikazane kao vektore u višedimenzionalnom prostoru u grupe zvane grozdovi. Elementi nekog grozda međusobno pokazuju veći stupanj sličnosti od onog s elementima drugih grozdova [Jain et al., 1999].

Neke potrebne formalne definicije pretežno preuzete iz [Jain et al., 1999] su sljedeće:

- entitet x je pojedinačna podatkovna točka u procesu grožđenja te se sastoji od vektora koji sadrži d vrijednosti, $x = (x_1, x_2, ...x_d)$

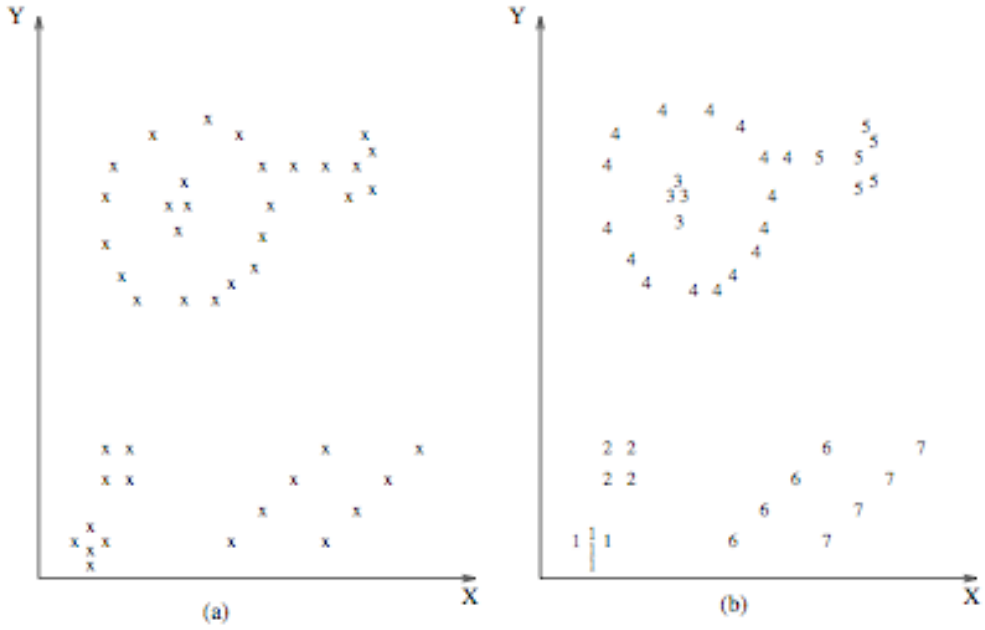
- svaka od d vrijednosti se naziva svojstvo

- d zapravo označava broj dimenzija vektorskog prostora
- taj se vektorski prostor naziva prostorom svojstava (engl. *feature space*)
- skup entiteta se prikazuje kao $E = \{x_1, ...x_n\}$, dok se i-ti element prikazuje kao $x_i = (x_{i1}, ...x_{id})$
- klasa može biti definirana kao izvor entiteta čije su razdiobe svojstava ovisne o pripadnosti toj pojedinoj klasi, grožđenjem se analizom svojstava entiteta pokušava rekonstruirati ta pripadnost entiteta nekoj klasi
- tvrdo grožđenje (engl. *hard clustering*) svakom entitetu x dodijeljuje jednu oznaku klase o_i, skup klasa za skup E je $K = \{k_1, ...k_n\}$ s $k_i \in \{1, ...m\}$ gdje je m broj različitih klasa
- meko grožđenje (engl. *soft clustering*) također zvano smušeno grožđenje (engl. *fuzzy clustering*) svakom entitetu x_i dodjeljuju stupanj pripadnosti p_{ij} za svaku klasu j
- mjera udaljenosti je mjera udaljenosti dvaju entiteta u prostoru svojstava

Velika je razlika između grožđenja kao nenadzirane metode i nadzirane klasifikacije. Naime, u nadziranoj klasifikaciji algoritmu je dan skup označenih podataka, takozvani skup za učenje (engl. *training set*) na kojem algoritam vrši mjerenja svojstava te njihovu ovisnost o oznaci. Oznaka označava pripadnosti grupi, odnosno kategoriji. Pomoću izmjerenih vrijednosti algoritam je u mogućnosti predvidjeti pripadnost kategoriji elemenata čiju kategoriju ne poznaje [Alpaydin, 2004]. Dva su najveća problema nadziranog učenja

1. tzv. usko grlo usvajanja (engl. *acquisition bottleneck*), odnosno velika količina označenih podataka potrebna za mjerenje potrebnih vrijednosti
2. neprimjenjivost metode na zadatke gdje *a priori* kategorije nisu poznate, kao na primjer u pronalaženju događaja grožđenjem (grupe će biti događaji koji se često još nisu niti dogodili).

Slika 2.1: Primjer moguće organizacije podatkovnih točaka u grozdove



Nadzirane metode često u problemima s ograničenim brojem kategorija uz razumnu količinu označenih podataka daju zadovoljavajuće rezultate. Pravi izazov leži, slažu se istraživači, u nenadziranim metodama gdje algoritam na raspolaganju nema označeni skup podataka već on analizom neoznačenih podataka prepoznaje pravilnosti u podacima te preko njih, odnosno iz impliciranih različitosti grupira entitete.

Primjer rezultata postupka grožđenja vidljiv je na slici 2.1 gdje su u grafikonu (a) prikazane podatkovne točke, a u grafionu (b) te podatkovne točke grupirane u moguće grozdove.

2.1 Koraci u grožđenju

U pravilu se postupak grožđenja sastoji od sljedećih koraka [Jain and Dubes, 1988]:

1. prikaz entiteta, odnosno podatkovnih točaka (uključuje mogući odabir i određivanje svojstava (engl. *feature selection and extraction*)

2. izračun matrice udaljenosti funkcijom udaljenosti (računa se udaljenost između svake dvije podatkovne točke)

3. grožđenje podataka

4. apstrakcija podataka (ukoliko je potrebna)

5. procjena rezultata (ukoliko je potrebna).

Od ovih pet navedenih koraka prva se tri susreću redovito u literaturi te se tako u kontekstu grožđenja dokumenata spominju sljedeća tri koraka [Forster, 2006]:

1. formalizacija dokumenata

2. izračun matrice udaljenosti

3. grožđenje dokumenata

U nastavku poglavlja 2.1 koraci u grožđenju bit će prikazani kroz naslove prikaza entiteta, izračuna matrice udaljenosti i procesa grožđenja.

2.1.1 Prikaz entiteta

Entiteti koje se želi grupirati grožđenjem u početku se mogu nalaziti u stvarnom, odnosno podatkovnom obliku (stvarni bicikl ili pak neke vrijednosti koje opisuju taj bicikl). Sami podaci o entitetu se često nalaze u različitim vrstama zapisa - u numeričkom, tekstualnom, zvučnom, grafičkom, video, multimedijskom, i to u računalno obradivom, odnosno računalno neobradivom zapisu. U prvom se koraku entitete prikazuje nizom vrijednosti. Taj se niz vrijednosti najčešće naziva vektor. Svaka vrijednost iz vektora opisuje neko svojstvo pojedinog entiteta. Vrijednosti svojstava mogu prema [Gowda and Diday, 1991] pripadati jednoj od sljedećih vrsta:

1. kvantitativne vrijednosti

   (a) kontinuirane vrijednosti (npr. težina)

(b) diskretne vrijednosti (npr. broj elemenata)

(c) intervalne vrijednosti (npr. trajanje nekog vremenskog intervala)

2. kvalitativne vrijednosti

(a) nominalne vrijednosti (npr. boja)

(b) ordinalne vrijednosti (npr. vrsta entiteta u popisu mogućih vrsta)

Tako se, primjerice bicikl, može među ostalima opisati i svojstvima težine (kontinuirana), broja brzina (diskretna), starosti (intervalna), boje (nominalna) te tipa (ordinalna) itd.

Entiteti se kroz te vrijednosti žele prikazati na način da se naglase najvažnija svojstva pojedinog entiteta. Primjerice, ako se podatkovne točke međusobno ne razlikuju po nekom svojstvu, odnosno ako ono nema diskriminativnu vrijednost, vjerojatno je da to svojstvo nećemo htjeti iskoristiti za opis entiteta.

Proces odabira svojstava kojima se prikazuje entitet se naziva odabir svojstava (engl. *feature selection*), a proces transformiranja osnovnih svojstava u nova svojstva se naziva određivanje svojstava (engl. *feature extraction*) [Jain et al., 1999]. Primjer odabira svojstava u području obrade prirodnog jezika je odabir onih pojavnica koje imaju semantički sadržaj (ne odabiru se veznici, pomoćni glagoli i slični oblici koji sami po sebi nemaju semantički sadržaj) [Manning and Schütze, 1999a]. Primjer određivanja svojstava u području obrade prirodnog jezika je morfosintaktička obrada dokumenta te prikaz dokumenta kroz morfosintaktičke kategorije koje posjeduje [Manning and Schütze, 1999b].

Proces prikaza entiteta, odnosno formalizacije dokumenata je detaljnije razložen u poglavlju 2.1.1.

2.1.2 Izračun matrice udaljenosti

Drugi korak u grožđenju ima za zadatak izračunati simetričnu matricu udaljenosti (engl. *distance matrix*) $U = (u_{ij})_{nxn}$. Iz definicije matrice U je

vidljivo da matrica ima broj dimenzija jednak broju entiteta x_i čije su udaljenosti njome opisane. Matrica udaljenosti se u literaturi povremeno naziva i matricom sličnosti (engl. *similarity matrix*) ovisno o tome jesu li u njoj pohranjene mjere sličnosti, odnosno različitosti entiteta. Za izračun matrice udaljenosti odabire se funckija udaljenosti (engl. *distance function*), odnosno sličnosti (engl. *similarity function*).

Pojedine funkcije udaljenosti opisuje poglavlje 2.3.

2.1.3 Proces grožđenja

Sam proces grožđenja kao ulaz prima matricu udaljenosti te u pravilu samo na temelju nje vrši grupiranje, odnosno grožđenje entiteta. Poneki algoritmi pri grožđenju koriste dodatno znanje o entitetima, odnosno postavljaju određena ograničenja na postupak grožđenja. Primjer takvog ograničenja je grožđenje nekih entiteta s ograničenjem da se u jednom grozdu smije nalaziti samo jedan entitet iz neke unaprijed poznate kategorije.

Algoritmi za grožđenje se najčešće dijele na hijerarhijske i particijske, odnosno nehijerarhijske algoritme [Jain et al., 1999]. Osnovne su značajke hijerarhijskih, odnosno particijskih algoritama sljedeće [Manning and Schütze, 1999c]:

Hijerarhijski algoritmi:

- poželjni su za detaljnu analizu podataka
- pružaju više informacija od particijskog grožđenja
- nije potrebno *a priori* definirati broj krajnjih grozdova, već stupanj grožđenja na kojemu se daljnje grožđenje prekida
- računalno su zahtjevniji od particijskih algoritama

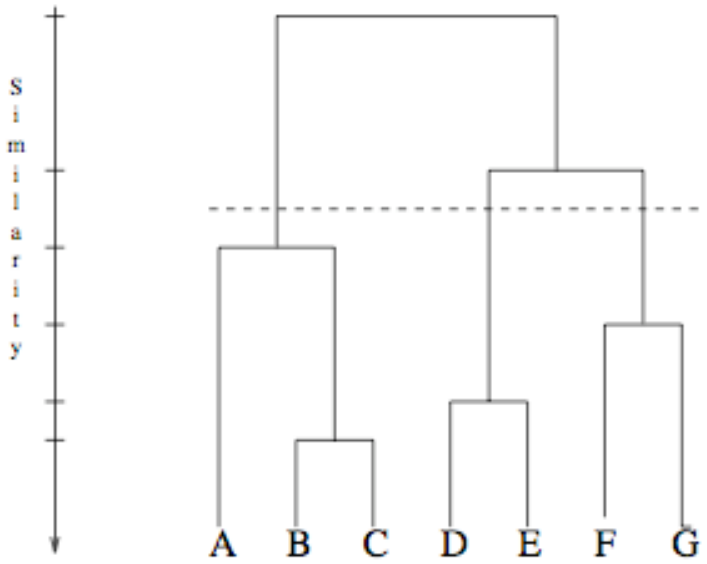
Particijski algoritmi:

- poželjni su za grožđenje veće količine podataka
- pružaju manje podataka o odnosima između entiteta od hijerarhijskog grožđenja
- potrebno je *a priori* definirati broj grozdova
- računalno su jednostavniji od hijerarhijskih algoritama

Osnovna je značajka hijerarhijskih algoritama da grožđenje provodi u koracima. Dvije su osnovne podgrupe hijerarhijskih algoritama aglomerativni i divizivni [Wikipedia, 2009a]. Aglomerativni algoritmi započinju sa svakim entitetom u vlastitom grozdu te spajaju grozdove do trenutka kad se svi entiteti nalaze u jedinom grozdu, odnosno kreću se od dna prema gore (engl. *bottom-up*). Suprotno tome, divizivni algoritmi započinju sa svim entitetima u jednom grozdu koji dijele do trenutka kada je svaki entitet u vlastitom grozdu, odnosno kreću se od vrha prema dolje (engl. *top-down*).

Cijeli se proces grožđenja, često prikazan dendogramom poput onoga na slici 2.2, u istražujućim zadacima koristi za proučavanje međuodnosa entiteta. No čest je slučaj da se, kako u istražujućim tako i u potvrđujućim zadacima, kao rezultat želi dobiti entitete grupirane u grozdove. Kako se primjerice hijerarhijsko aglomerativno grožđenje provodi od trenutka kad je definirano n grozdova nad n entiteta do trenutka kad je preostao samo jedan, potrebno je odrediti trenutak u kojem se želi prekinuti daljnje grožđenje. Na slici 2.2 je vodoravna iscrtkana linija primjer takvoga trenutka koji se često definira kao vrijednost p nazvana prag (engl. *threshold*). Drugi često korišten termin za prag je i točka rezanja (engl. *cut-off point*) [Manning and Schütze, 1999c].

Slika 2.2: Primjer prikaza rezultata grožđenja dendogramom

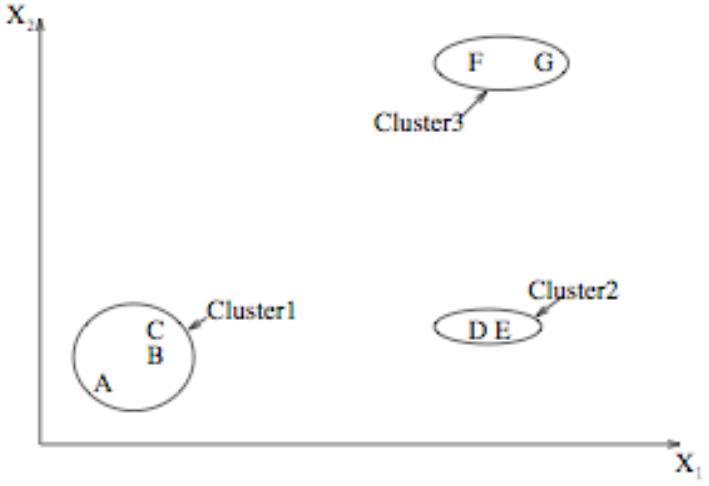


U praksi se prag često definira kao minimalna udaljenost između dva grozda potrebna da se ta dva grozda spoje u jedan. Rezultat primjene praga na slici 2.2 je organizacija entiteta u grozdove kakva je prikazana na slici 2.3.

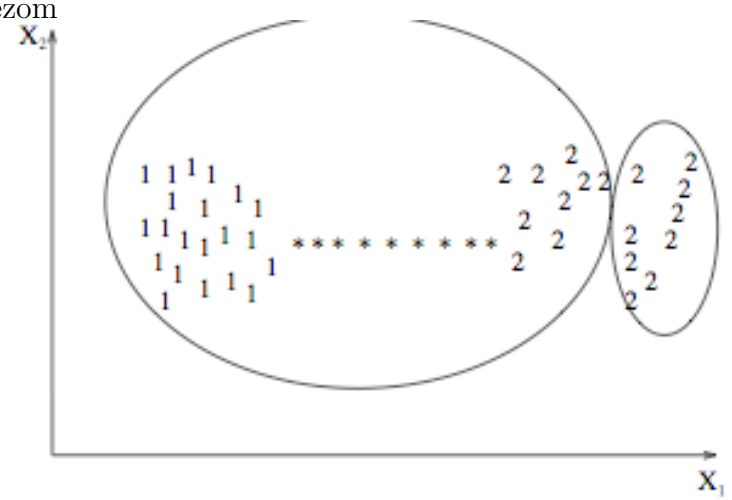
Najprimjenjivaniji hijerarhijski algoritmi za grožđenje su aglomerativni algoritam pojedinačom vezom (engl. *single-link agglomerative clustering*), aglomerativni algoritam potpunom vezom (engl. *complete-link agglomerative clustering*) te algomerativni algoritam prosječnom vezom (engl. *average-link agglomerative clustering*). Vrsta veze ukazuje na to kako se računa udaljenost između dva grozda - kao udaljenost između dva najbliža entiteta (pojedinačna veza), kao udaljenost dvaju najudaljenijih entiteta (potpuna veza) ili pak kao prosječna udaljenost svih entiteta (prosječna veza). Razlika u rezultatu ovih algoritama jest u tome da algoritmi pojedinačnom vezom pokazuju tendenciju izduženih grozdova. Naime, oni ne uzimaju u obzir udaljenost potencijalne točke i svih točaka u grozdu, odsnono one najudaljenije već samo nje i najbliže točke u grozdu. Jedni, odnosno drugi algoritmi pokazuju bolje rezultate na različitim zadacima. Primjer grožđenja pojedinačnom vezom prikazan je na slici 2.4, a analogni primjer potpunom vezom na slici 2.5.

Za razliku od hijerarhijskih algoritama, particijski algoritmi grožđenje

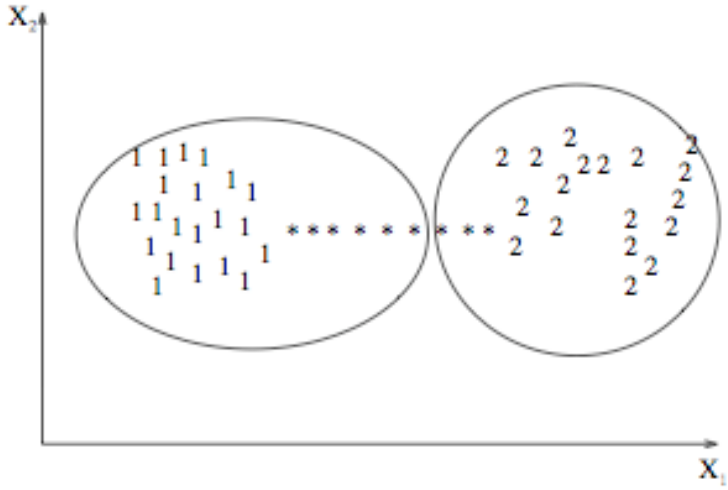
Slika 2.3: Rezultat grožđenja na temelju kojega je napravljen dendogram sa slike 2.2



Slika 2.4: Primjer rezultata algoritma aglomerativnog grožđenja pojedinačnom vezom



Slika 2.5: Primjer rezultata algoritma aglomerativnog grožđenja potpunom vezom



vrše direktno, u jednom koraku. Većina particijskih algoritama se zasniva na optimiziranju neke funkcije gubitka (engl. *loss function*) koja je definirana lokalno (na podskupu entiteta) ili globalno (na cijelom skupu entiteta E). Kombinatorno pretraživanje skupa mogućih pripadnosti entiteta klasama za optimalnu vrijednost funkcije gubitka je računalno iznimno zahtjevno. Zato se u praksi takvi algoritmi najčešće pokreću određeni broj puta nad različitim početnim stanjima te se od mogućih rješenja bira ono koje daje optimalnu vrijednost funkcije gubitka. Jedna od popularnijih funkcija gubitka je ona vrijednosti kvadrata pogreške (engl. *squared error*):

$$e^2(E,K) = \sum_{j=1}^{m} \sum_{i=1}^{n_j} ||x_i^j - c_j||^2 \tag{2.1}$$

gdje je E skup entiteta, K skup klasa, odnosno grozdova, x_i i-ti entitet iz E koji pripada klasi, odnosno grozdu j, a c_j centroid (predstavnik) klase, odnosno grozda j. Cilj je particijskog algoritma pronaći optimalnu, tj. minimalnu vrijednost ove funkcije. Očito je da će ova vrijednost biti minimalna za najkompaktnije grozdove te će svoj minimum uvijek dostići u situaciji gdje je svaki entitet u vlastitom grozdu. Kako takvo rješenje nije informativno,

particijski algoritmi zahtijevaju *a priori* znanje o broju grozdova, odnosno klasa što je usput i najveće ograničenje particijskih algoritama. Za uzvrat, particijski su algoritmi za razliku od hijerarhijskih u pravilu računalno manje zahtjevni te ne zahtijevaju definiranje praga.

Jedan od najkorištenijih particijskih algoritama koji koristi metodu minimizacije kvadrata pogreške definiranu u izrazu 2.1 je K-sredina (engl. *K-means*) [Macqueen, 1967]. Taj algoritam se izvršava u određenom broju iteracija na način da se slučajnim odabirom odabere po jedan predstavnik za svaki od K grozdova te se svi preostali entiteti pridruže onom grozdu kojem su najbliži. Broj se iteracija najčešće određuje dinamički uvjetom maksimalne dozvoljene vrijednosti funkcije gubitka ili minimalnog smanjenja vrijednosti funkcije gubitka od prošle iteracije [Jain et al., 1999].

Drugi česti particijski algoritmi su MST algoritam [Zahn, 1971] koji pripada algoritmima iz teorije grafova, te EM algoritam [Dempster et al., 1977] koji pripada mješovitim modelima (engl. *mixture models*).

Algoritmi za grožđenje korišteni u ovom doktorskom radu su detaljnije opisani u poglavlju 2.4.

2.2 Prikaz dokumenata za grožđenje dokumenata

U poglavlju 2.1 grožđenje je opisano u tri koraka:

1. prikaz entiteta
2. izračun matrice udaljenosti
3. proces grožđenja

Govoreći o grožđenju dokumenata, prva točka - prikaz entiteta, odnosno dokumenta predstavlja njegovu formalizaciju kako bi u drugoj točki bilo moguće izračunati udaljenost između svih entiteta, odnosno dokumenata.

Kako se kod svakog dokumenta radi zapravo o nizu znakova prikazanih kodnim vrijednostima, potrebno je primijeniti metode koje su u mogućnosti

prodrijeti do određene razine u značenjski sadržaj dokumenta. Upravo se tim problemom, među ostalima, bavi područje obrade prirodnog jezika.

Entiteti se najčešće, kako je rečeno u poglavlju 2.1.1 prikazuju kao niz svojstava koja opisuju taj entitet. Formalni zapis tih svojstava je najčešće vektor u kojem svaka dimenzija, odnosno vrijednost prikazuje neko svojstvo. Najjednostavniji su vektori takozvani binarni vektori gdje je vrijednost zapisana u dimenziji 0 ili 1 ovisno o tome posjeduje li entitet to svojstvo ili ne. Kompleksniji vektori sadrže diskretne, odnosno kontinuirane vrijednosti koje ukazuju na stupanj relevantnosti svojstva za određeni entitet.

Jedan od najjednostavnijih prikaza dokumenata je onaj gdje se dokument smatra skupom pojavnica, odnosno riječi kakve se pojavljuju u tekstu. U tom prikazu različnice - različite riječi kakve se pojavljuju u tekstu - iz cijelog korpusa, odnosno skupa dokumenata, čine popis svojstava.

Prethodno opisani postupak je primjer odabira svojstava nekog entiteta, odnosno dokumenta. Dodatni korak u formalizaciji dokumenta je pridruživanje kvantitativne vrijednosti pojedinom svojstvu koje ukazuje na važnost tog svojstva za taj entitet. Za pridruživanje te vrijednosti u obradi prirodnog jezika se najčešće koriste mjere težine svojstava koje mjere relevantnost nekog svojstva za neki entitet, odnosno dokument. Tako, primjerice, prethodno opisani slučaj prikaza dokumenta svojstvima različnica možemo upotpuniti pridruživanjem binarne vrijednosti svakom određenom svojstvu na temelju kojega znamo je li se to svojstvo, odnosno ta različnica, pojavila u dokumentu ili nije.

Ovime dolazimo do zaključka da se postupak formalizacije dokumenta sastoji od dva zadatka:

1. određivanja svojstava nekog skupa dokumenata
2. mjerenja povezanosti pojedinog svojstva za neki dokument

Kako znamo da čestota pojavljivanja neke pojavnice govori nešto o važnosti te pojavnice u nekom dokumentu, tako možemo pretpostaviti da bi bilježenje čestote neke pojavnice kao mjere težine svojstva za dokument bio vjerojatno bolji način prikaza relevantnosti svojstva za neki dokument.

Nadalje, svijesni smo da su najčešće pojavnice upravo funkcijske riječi, odnosno one koje nemaju semantički sadržaj. Sa statističkog stajališta one nisu pretjerano relevantne za neki dokument upravo zato što im je distinktivna vrijednost između dokumenata iznimno niska. Iz tog se razloga redovito kao mjere težine svojstava koriste one mjere koje upjevaju prikazati koliko je neko svojstvo posebno za neki entitet, odnosno koliko ono uspijeva razlikovati taj entitet od drugih entiteta.

U sljedeća dva poglavlja 2.2.1 i 2.2.2 dan je prikaz lingvističkih disciplina i njihovih mogućnosti odabira, odnosno određivanja svojstava te mjera težina tih odabranih i određenih svojstava kojima se prikazuje neki dokument.

2.2.1 Razine u obradi prirodnog jezika

Govoreći o razinama u obradi prirodnog jezika, odnosno o mogućnosti analize nekog jezičnog uzorka na određenoj jezičnoj razini, najčešće se pribjegava klasičnoj podjeli gramatike na

- fonetiku i fonologiju
- morfologiju
- sintaksu
- semantiku
- pragmatiku

U daljnjem je tekstu kratki osvrt na mogućnosti primjene određene jezične razine na formalizaciju dokumenata pri grožđenju.

Fonetika i fonologija

Fonetika i fonologija se bave glasovnim aspektom jezika, odnosno govora. Unutar obrade prirodnog jezika njihova je velika važnost u obradi zvučnog signala govora koja se najčešće bavi dvama zadacima:

1. analizom govora, odnosno pretvaranjem govora u tekst

2. sintezom govora, odnosno pretvaranjem teksta u govor

U kontekstu grožđenja dokumenata od koristi može biti sustav za analizu govora u slučaju da dokumenti na raspolaganju nisu samo tekstualni, već i zvučni. Ovakvom se predobradom podataka za grožđenje i druge obrade upravo bavila TDT inicijativa [TDT, 2004].

U slučaju da su izvorni podaci prikazani tekstualno nema potrebe za primjenom metoda ove jezične razine.

Morfologija

Morfologija se bavi dvama fenomenima - tvorbom riječi te promjenom riječi. Ukratko rečeno, u centru interesa su riječ i njezine promjene. Flektivna morfologija, tj. onaj dio koji se bavi promjenom nekog leksema unutar njegove paradigme je područje koje se vrlo često upotrebljava u prikazu dokumenata za grožđenje. Najčešće metode koje se primjenjuju se korjenovanje, odnosno svođenje neke pojavnice na njen korjen te lematizacija, odnosno svođenje neke pojavnice na njen osnovni oblik, odnosno lemu. U nekim se jezicima koji su tvorbeno vrlo aktivni, poput njemačkoga, vrlo često koriste i tvorbene obrade poput rastavljanja složenica na sastavne lekseme [Forster, 2006].

Morfološki podaci, točnije rečeno morfosintaktički podaci poput vrste riječi, pripadajuće leme ili morfosintaktičke kategorije često se također upotrebljavaju kao svojstva, odnosno kao kriteriji odabira neke pojavnice u prikazu dokumenta.

Metode ove jezične razine češće se primjenjuju kod morfološki kompleksnijih jezika kao što je hrvatski. Važan razlog česte primjene morfološke obrade jest činjenica da često pospješuje rezultate, dok je postupak dalje relativno jednostavan u usporedbi s postupcima na višim jezičnim razinama.

Sintaksa

Sintaksa se bavi ustrojem rečenica, odnosno pravilima nizanja riječi u rečenici.

U obradi prirodnog jezika sintaksa se, kao i u lingvistici, nastavlja na morfološku obradu jezika. Veza morfologije i sintakse vidljiva je i u terminu

morfosintakse koji ukazuje da je nemoguće odrediti točnu primjenjenu morfološku kategoriju neke pojavnice bez njenog konteksta, odnosno sintaktičkih podataka.

Prvi sintaktički zadatak koji se riješava obradom prirodnog jezika je tzv. razdjeljivanje (engl. *chunking*), odnosno pronalaženje osnovnih gradivnih elemenata rečenice, a to su imenične, prijedložne i glagolske sveze (engl. *noun, prepositional and verb phrases*). Od tih se elemenata hijerarhijskim povezivanjem ustanovljava struktura cijele rečenice. Postupak ustanovljavanja strukture rečenice naziva se raščlamba (engl. *parsing*) i redovito se oslanja na prethodno izvršeno razdjeljivanje.

Jednostavnija alternativa razdjeljivanju, postupak koji ne zahtijeva nikakvo jezično znanje već se oslanja na statističku analizu supojavljivanja pojavnica naziva se pronalaženje kolokacija (engl. *collocation detection*) te se u praksi često koristi kao jednostavnija alternativa razdjeljivanju za pronalaženja višečlanih izraza u tekstu.

Postoje sintaktički problemi koji prelaze granice rečenice, to je, primjerice razrješavanje anafore (engl. *anaphora resolution*). Anaforom se smatra referiranje nekog jezičnog elementa na neki drugi element. Zamjenice po svojoj prirodi pretpostavljaju upotrebu anafore. To referiranje često prelazi granice rečenice. Kako je razrješavanje morfološke kategorije zahtijevalo i sintaktičko znanje, tako i razrješavanje anafore zahtijeva semantičko. Ovakvo preklapanje jezičnih razina redovit je slučaj.

U kontekstu formalizacije prikaza dokumenata vršeni su mnogi pokušaji uklapanja sintaktičkih podataka u vektore. To je činjeno i u slučaju daljnjeg grožđenja tih dokumenata [Forster, 2006]. Opći trend tih pokušaja jest da iznimno otežava pretprocesiranje i znatno povećava prikaz dokumenta, a da je pritom pomak u rezultatima zanemariv ili čak negativan.

Semantika

Semantika se bavi značenjem te se razlikuje semantika riječi i semantika rečenice.

Semantika riječi se bavi značenjem pojedinog leksema, odnosno

značenjskim međuodnosom više leksema. Najpoznatiji računalni model, odnosno alat leksičke semantike jest Wordnet [Fellbaum, 1998].

Semantika rečenice bavi se značenjem cijelih jezičnih iskaza. Čest način formalnog prikaza značenja je logika prvog reda, odnosno predikatni račun. Takav prikaz jezičnih iskaza do određene mjere omogućuje izvođenje novih zaključaka.

U rješavanju zadataka pronalaženja događaja, odnosno grožđenja dokumenata postoje pokušaji uključivanja leksičkosemantičkih podataka najčešće u obliku uključivanja istoznačnica leksema [Forster, 2006]. Takvi pokušaji redovito pokazuju negativan trend u krajnjem rezultatu. Razlog tome leži u kompleksnosti odnosa značenja pojedinih leksema i njihovoj višeznačnosti o čemu se pri takvim pokušajima ne uspijeva voditi računa.

Pragmatika

Pragmatika se bavi načinom kako kontekst sudjeluje u značenju. Ona se unatoč svojoj kompleksnosti također pokušava formalizirati i učiniti računalno obradivom. Praktične primjene pragmatike u obradi prirodnog jezika do danas su ipak rijetke.

U rješavanju problema pronalaženja događaja pragmatičko modeliranje nije još uključivano.

2.2.2 Mjere težine svojstava

Već je u poglavlju 2.2 objašnjeno kako je moguće koristiti različite mjere relevantnosti nekog svojstva za neki entitet, odnosno dokument. Argumentirano je zašto je korisno upotrebljavati komleksnije mjere relevantnosti od onih binarnih koje govore o prisustvu, odnosno neprisustvu nekog svojstva u nekom entitetu.

Područja iz kojih dolaze mjere težina svojstava su najčešće

1. vjerojatnost
2. informacijska teorija

3. statistički testovi

Najčešće upotrebljavane mjere težine svojstava su sljedeće [Manning and Schütze, 1999d, Curran, 2004, Forster, 2006, Jurafsky and Martin, 2008]:

1. vjerojatnost
2. uvjetna vjerojatnost
3. TF-IDF
4. pojedinačna međusobna informacija
5. t-test

Vjerojatnost, uvjetna vjerojatnost i TF-IDF pripadaju grupi mjera iz područja vjerojatnosti, pojedinačna međusobna informacija grupi mjera iz područja informacijske teorije, a t-test grupi mjera iz područja statističkih testova.

Vjerojatnost

Vjerojatnost slučajne varijable X se općenito računa procjenom najveće vjerojatnosti kao

$$P(X = x) = \frac{broj(X = x)}{broj(X)} \tag{2.2}$$

Vjerojatnost nekog svojstva u nekom dokumentu jednaka je omjeru broja pojava tog svojstva i broja pojava svih svojstava. Ako je S slučajna varijabla svojstava, najjednostavniju mjeru težine svojstva korištenu u ovom doktorskom radu je moguće prikazati ovako:

$$tez_{vj} = \frac{broj(S = s)}{broj(S)} \tag{2.3}$$

Ova mjera očito ne koristi podatke iz drugih dokumenata već samo govori kolika je vjerojatnost svojstva unutar jednog dokumenta.

Uvjetna vjerojatnost

Mjera vjerojatnosti koja koristi informacije i iz drugih dokumenata je uvjetna vjerojatnost dokumenta ako je poznato svojstvo, odnosno $P(D = d|S = s)$. Ona se procjenom najveće vjerojatnosti može zapisati kao

$$P(D = d|S = s) = \frac{broj_s(D = d, S = s)}{broj(S = s)} \tag{2.4}$$

Kako je odnos uvjetne i zajedničke vjerojatnosti

$$P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X) \tag{2.5}$$

zajedničku je vjerojatnost $P(D = d, S = s)$ moguće zapisati kao

$$P(D = d, S = s) = \frac{broj_s(D = d, S = s)}{broj(S = s)} \frac{broj(S = s)}{broj(S)} = \frac{broj_s(D = d, S = s)}{broj(S)} \tag{2.6}$$

Na kraju, uvjetnu vjerojatnost $P(D = d|S = s)$ moguće je prikazati kao

$$P(D = d|S = s) = \frac{\frac{broj_s(D=d,S=s)}{broj(S)}}{\frac{broj(S=s)}{broj(S)}} = \frac{broj_s(D = d, S = s)}{broj(S = s)} \tag{2.7}$$

Mjera težine svojstva se krajnje može zapisati kao

$$tez_{uvvj}(D = d, S = s) = \frac{broj_s(D = d, S = s)}{broj(S = s)} \tag{2.8}$$

TF-IDF

TF-IDF je danas najupotrebljavanija mjera težine nekog svojstva u obradi prirodnog jezika. Svoju je popularnost stekla u području pretraživanja in-

formacija [Manning et al., 2008a]. TF-IDF se zapravo sastoji od dvije mjere - TF - mjere frekvencije termina identične našoj mjeri težine vjerojatnosti - te IDF - opće mjere važnosti nekog termina.

TF se računa kao i izraz 2.3, odnosno

$$TF(S = s) = \frac{broj(S = s)}{broj(S)} \tag{2.9}$$

te ne uzima u obzir dokaze iz drugih dokumenata. Kako bi mjera činila to, TF se množi s mjerom IDF koja je logaritam faktora broja dokumenata i broja dokumenata u kojima se neki termin pojavljuje. Sam faktor, dakle, poprima vrijednosti od 1 na više gdje 1 dobivaju ona svojstva koja su prisutna u svakom dokumentu, a veće brojeve ona svojstva koja su prisutna u manjem broju dokumenata, odnosno ona koja su općenito specifičnija za pojedine dokumente. Logaritam te vrijednosti donju granicu IDF mjere postavlja na 0. To znači da će svojstvo koje se pojavljuje u svakom dokumentu, nevezano uz njegovu frekvenciju u trenutnom dokumentu, imati TF-IDF vrijednost jednaku 0.

Izraz za IDF je

$$IDF(S = s) = \log \frac{broj(D)}{broj_d(D, S = s)} \tag{2.10}$$

Cijeli izraz odnosno mjera težine svojstva TF-IDF se može zapisati kao

$$tez_{TF-IDF}(D = d, S = s) = \frac{broj(S = s)}{broj(S)} \log \frac{broj(D)}{broj_d(D, S = s)} \tag{2.11}$$

Pojedinačna međusobna informacija

Međusobna informacija je informacijska mjera koja govori koliko su dvije slučajne varijable međusobno povezane, odnosno koliko znamo o jednoj ako

poznajemo drugu. Izraz koji opisuje međusobnu informaciju je

$$I(X,Y) = \sum_x \sum_y P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{2.12}$$

Za razliku od međusobne informacije, pojedinačna međusobna informacija je mjera između dvije vrijednosti slučajnih varijabli koja govori koliko se često te dvije vrijednosti supojavljuju s obzirom na očekivanje da su te dvije vrijednosti nezavisne. Izraz za računanje pojedinačne međusobne informacije je

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{2.13}$$

Za računanje pojedinačne međusobne informacije u vektorskom prostoru kao mjere povezanosti leksema i svojstva se koristi sljedeći izraz

$$pmi(d,s) = \log_2 \frac{P(d,s)}{P(d)P(s)} = \log_2 \frac{\frac{broj(D=d,S=s)}{broj(S)}}{\frac{broj(D=d)}{broj(D)} \frac{broj(S=s)}{broj(S)}} = \tag{2.14}$$

$$= \log_2 \frac{broj(D=d,S=s)broj(D)}{broj(D=d)broj(S=s)} \tag{2.15}$$

Kako je izraz $broj(D)$ konstantan, moguće ga je ukloniti iz izračuna isto kao i izraz $broj(D=d)$ iz razloga što je uvijek jednak 1.

Krajnji izraz za računanje pojedinačne međusobne informacije je sljedeći:

$$tez_{pmi}(d,s) = \log_2 \frac{broj(D=d,S=s)}{broj(S=s)} \tag{2.16}$$

Ovdje je moguće primijetiti da je izraz za računanje pojedinačne međusobne informacije jednak logaritmu izraza za računanje uvjetne vjerojatnosti $P(D|S)$.

T-test

T-test je najpoznatiji statistički test za testiranje hipoteza. U [Curran, 2004] taj je test primijenjen kao mjera težine svojstva testirajući nezavisnost svojstva i entiteta za koji se težina svojstva računa. T-test je zapisan na sljedeći način:

$$t(D = d, S = s) = \frac{P(D = d, S = s) - P(D = d)P(S = s)}{\sqrt{P(D = d, S = s)}} \tag{2.17}$$

Korištenjem procjene najveće vjerojatnosti taj je izraz moguće zapisati kao

$$t(D = d, S = s) = \frac{\frac{broj(D=d,S=s)}{broj(S=s)} - \frac{broj(D=d)}{broj(D)}\frac{broj(S=s)}{broj(S)}}{\sqrt{\frac{broj(D=d,S=s)}{broj(S=s)}}} \tag{2.18}$$

2.3 Funkcije sličnosti za grožđenje dokumenata

U grožđenju i sličnim zadacima gdje je potrebna funkcija sličnosti koja omogućuje izračun matrice sličnosti između podatkovnih točaka razvijen je veliki broj različitih funkcija. Te se funkcije u pravilu mogu s obzirom na izvor podijeliti u sljedeće grupe:

1. funkcije iz područja geometrije
2. funkcije iz područja trigonometrije
3. funkcije iz područja teorije skupova
4. funkcije iz područja teorije informacija

U ovom će radu biti korištene one funkcije koje se u praksi najčešće upotrebljavaju. Svaka od prethodno navedenih grupa će imati barem jednog predstavnika. Korištene će biti sljedeće mjere sličnost:

1. *Manhattan* udaljenost (L1 norma) i Euklidova udaljenost (L2 norma)
2. Kosinus
3. *Jaccard*ov i *Dice* koeficijent
4. Jensonovo i Shannonovo odstupanje

2.3.1 *Manhattan* i Euklidova udaljenost

Manhattan i Euklidova udaljenost pripadaju skupini geometrijskih udaljenosti. One mjere geometrijsku udaljenosti između dva vektora u n-dimenzionalnom prostoru.

Manhattan udaljenost je

$$udalj_{manhattan}(\vec{x}, \vec{y}) = \sum_{i=1}^{N} |x_i - x_j| \tag{2.19}$$

Manhattan udaljenost mjeri zapravo udaljenost dvaju vektora s obzirom na sve dimenzije. Alternativni naziv joj je *Taxicab* udaljenost [Wikipedia, 2009b]. Oba naziva kao analogiju korsite kretanje automobila, odnosno taksija po planski građenim četvrtima gdje su sve ulice pod pravim kutom.

Euklidova je udaljenost

$$udalj_{euklid}(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{N} (x_i - x_j)^2} \tag{2.20}$$

Za razliku od *Manhattan* udaljenosti koja mjeri udaljenost između svih dimenzija zasebno, Euklidova udaljenost mjeri stvarnu, geometrijsku udaljenost između dva vektora.

Manhattan udaljenost se često naziva L_1 normom, dok se Euklidova udaljenost naziva L_2 normom. Skup tih mjera se naziva *Minkowski* udaljenostima [Curran, 2004] te tvore sve mjere od L_1 do L_∞ gdje mjera L_∞ tendira maksimalnoj udaljenosti dvaju vektora.

Obje vrijednosti ne mjere sličnost, već udaljenost te će u ovoj disertaciji one u matrice sličnosti ulaziti s negativnim predznakom, odnosno vrijedit će sljedeće:

$$slc_{manhattan} = -udalj_{manhattan}, slc_{euklid} = -udalj_{euklid} \tag{2.21}$$

Manhattan i Euklidova udaljenost kao geometrijske mjere pružaju intuitivno rješenje problema udaljenosti vektora, no njihov je problem što su neotporne na stršeće podatke [Lee, 1999].

Kako bi se izbjegli stršeći podaci, vektore je potrebno normalizirati. Upravo taj problem rješava mjera iz područja trigonometrije.

2.3.2 Kosinus

Kosinus je kao mjera sličnosti popularizirana u pretraživanju informacija. Glavna joj je prednost prema *Manhattan* i Euklidovoj udaljenosti činjenica da je otporna na stršeće podatke te je identična skalarnom produktu dvaju normaliziranih vektora. Izraz za računanje kosinusne mjere dvaju nenormaliziranih vektora je

$$slc_{kosinus}(\vec{x}, \vec{y}) = \frac{\vec{x}\vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{N} x_i \times y_i}{\sqrt{\sum_{i=1}^{N} x_i^2}\sqrt{\sum_{i=1}^{N} y_i^2}} \tag{2.22}$$

Kosinus je danas u području obrade prirodnog jezika najzastupljenija mjera sličnosti, među ostalim zato što mu sličnost pada u rasponu 0.0 i 1.0 opisujući sličnost identičnih vektora s 1.0, a sličnost vektora pod pravim kutem s 0.0. Čest je problem L_1 i L_2 mjera što daju vrijednosti koje je potrebno normalizirati.

2.3.3 *Jaccard* i *Dice* koeficijenti

Jaccard i *Dice* koeficijenti korijene imaju u teoriji skupova te predstavljaju mjeru sličnosti dvaju skupova. Originalna formalna definicija tih dvaju mjera

je [van Rijsbergen, 1979]

$$slc_{jaccard\_bin}(s_1, s_2) = \frac{|s_1 \cap s_2|}{|s_1 \cup s_2|} \tag{2.23}$$

$$slc_{dice\_bin}(s_1, s_2) = \frac{2 * |s_1 \cap s_2|}{|s_1| + |s_2|} \tag{2.24}$$

Jaccard koeficijent, ponekad nazivan Tanimoto mjerom [Tanimoto, 1958] jednostavno dijeli veličinu skupa presjeka s veličinom skupa unije dvaju skupova. Time taj koeficijent vraća vrijednost 0.0 ako je brojnik jednak 0, odnosno ako dva skupa nemaju presjek, dok vraća 1.0 ako su skupovi identični, odnosno presjek i unija daju isti rezultat. *Dice* koeficijent [Dice, 1945] je vrlo sličan *Jaccard* koeficijentu. On dijeli presjek skupova sa sumom veličine skupova te rezultat normalizira množeći brojnik s dva.

Kao mjere koje se primjenjuju nad skupovima prilagođene su binarnim vektorima, no obje imaju i inačice za kontinuirane vrijednosti. Tako se *Jaccard* koeficijent za kontinuirane vrijednosti često definira kao [Grefenstette, 1994]

$$slc_{jaccard}(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{N} min(x_i, y_i)}{\sum_{i=1}^{N} max(x_i, y_i)} \tag{2.25}$$

dok se *Dice* koeficijent definira kao [Curran, 2004]

$$slc_{dice}(\vec{x}, \vec{y}) = \frac{2 * \sum_{i=1}^{N} min(x_i, y_i)}{\sum_{i=1}^{N} x_i + \sum_{i=1}^{N} y_i} \tag{2.26}$$

Time se presjek skupova smatra sumom manjih vrijednosti dimenzija, dok se presjek smatra sumom većih vrijednosti u pojedinim dimenzijama.

2.3.4 *Jensen-Shannon* odstupanje

Jensen-Shannonovo odstupanje pripada grupi mjera teorije informacije, odnosno distribucijskih mjera koje mjere koliko jedna distribucija odstupa od druge, odnosno koliko je dodatne informacije potrebno da se jednom distribucijom objasni druga [Curran, 2004]. Takvu metodu uspoređivanja vektora u obradu prirodnog jezika uvodi [Pereira et al., 1993]. Najjednostavnija mjera iz grupe mjera teorije informacije je *Kullback-Leibler* odstupanje poznatije kao relativna entropija [Cover and Thomas, 1991]. Njegov je formalni zapis

$$D(P||Q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{2.27}$$

Relativna entropija zapravo mjeri gubitak informacije modelirajući slučajnu varijablu P distribucijom slučajne varijable Q. U slučaju da se jedna distribucija može zamijeniti drugom bez velikog gubitka informacije, može se pretpostaviti da su ta dva fenomena koji su njima opisani slični.

Problem s relativnom entropijom je taj što ona nema rješenje gdje je $q(x) = 0$ i $p(x) \neq 0$ što je u slučaju formalizacije jezičnih ostvarenja vektorima vrlo čest slučaj. Nadalje, nije simetriča, odnosno, u slučaju njene primjene vrijedilo bi $slc(a, b) \neq slc(b, a)$ što nije poželjno zato što se koncepti poput sličnosti redovito modeliraju kao simetrični fenomeni.

Postoje dva moguća rješenja ovog problema [Curran, 2004]. Prvi bi bilo izravnjavanje distribucija $P(x)$ i $Q(x)$ kako za nijedan x ne bi bile jednake nuli što zahtijeva mnogo dodatnog računanja. Drugi bi pristup bio računanje odstupanja od aritmetičke sredine dviju distribucija. Ono se naziva *Jensen-Shannon* odstupanje te se sastoji od računanja odstupanja obje distribucije od aritmetičke sredine obiju distribucija. Formalno se definira kao

$$JS(P, Q) = D(P||\frac{P+Q}{2}) + D(Q||\frac{P+Q}{2}) \tag{2.28}$$

Ovim je izrazom ostvarena simetričnost mjere. Isto tako vrijednosti padaju na skali od 0.0 do 1.0 gdje je samo za identične distribucije mjera odstu-

panja jednaka 0.0. Kako ova mjera zapravo mjeri udaljenost, a ne sličnost, u praksi će se oduzimati od 1, odnosno funkcija sličnosti koja koristi *Jensen-Shannon* odstupanje glasit će

$$slc_{js}(A, B) = 1 - JS(A, B) \tag{2.29}$$

2.4 Algoritmi grožđenja za pronalaženje događaja

Za zadatak pronalaženja događaja u pravilu u obzir dolaze samo hijerarhijski algoritmi, a ne parcijalni iz razloga što broj grozdova u tom zadatku nije poznat *a priori*. Postoje inačice parcijalnog grožđenja koje broj grozdova doživljavaju kao slobodni parametar koji također pokušavaju optimizirati u odnosu na neku funkciju troška [Wikipedia, 2009a].

U ovom doktorskom radu naglasak nije na različitim algoritmima za grožđenje, već na sveukupnom zadatku pronalaženja događaja grožđenjem što uključuje i velik broj slobodnih parametara, odnosno varijabli u koraku prikaza dokumenta. Iz tog će razloga biti istražena uspješnost algoritma za hijerarhijsko algomerativno grožđenje te za grožđenje jednim prolaskom. Oba algoritma imaju svojstvo da započinju sa svakim dokumentom u vlastitom grozdu te završavaju kad se svi dokumenti nalaze u jednom grozdu.

Kako rezultat u kojem su svi dokumenti u jednom grozdu nema smisla, potrebno je odrediti takozvanu točku rezanja, odnosno trenutak u grožđenju kad se grozdovi prestaju dalje spajati. Najčešći pristup određivanju točke rezanja je slobodni parametar praga koji određuje minimalnu sličnost između grozdova koja je potrebna da se dva grozda spoje u jedan. grožđenje staje kad nema više grozdova koji su sličniji od određenog praga. Vrijednost tog slobodnog parametra se redovito određuje eksperimentalno.

Hijerarhijsko grožđenje se, kao što je rečeno u poglavlju 2.1.3, može izvoditi u dva smjera. Prvi je smjer taj da su na početku sve točke u jednom grozdu koji se potom dijeli na više grozdova prema nekom kriteriju sve dok

se svaka točka ne nalazi u vlastitom grozdu. Takvo se hijerarhijsko grožđenje naziva divizivno. Kako se postupak hijerarhijskog grožđenja grafički najčešće prikazuje dendogramom, odnosno vrstom stabla, taj se pristup često naziva i od vrha prema dolje. U slučaju da se hijerarhijsko grožđenje izvršava od dna prema gore, odnosno da je svaka točka na početku u svom grozdu koji se potom spajaju određenim redoslijedom, taj se algoritam naziva aglomerativnim [Wikipedia, 2009a].

Postoje tri osnovna načina uspoređivanja grozdova. To su uspoređivanje

- pojedinačnom vezom - dva su grozda onoliko udaljena koliko su udaljena dva najbliža elementa ta dva grozda
- potpunom vezom - dva su grozda onoliko udaljena koliko su udaljena dva najudaljenija elementa
- prosječnom vezom - dva su grozda onoliko udaljena koliko su prosječnu udaljeni element obaju grozdova

Treba primijetiti da će, u slučaju da su funkcija dijeljenja i funkcija sličnosti identične, rezultat aglomerativnog i divizivnog grožđenja biti identičan. Aglomerativni su algoritmi za grožđenje u pravilu popularniji te će oni biti korišteni u ovom doktorskom radu. Korištenjem triju različitih načina uspoređivanja grozdova razlikuje se tri aglomerativna algoritma za grožđenje:

1. aglomerativni algoritam grožđenja pojedinačnom vezom
2. aglomerativni algoritam grožđenja potpunom vezom
3. aglomerativni algoritam grožđenja prosječnom vezom

Aglomerativni je algoritam implementiran na sljedeći način [Manning and Schütze, 1999c]:

- ulaz je skup entiteta
- neka svaki entitet pripada jednom grozdu
- neka postoji funkcija slic() koja vraća sličnost para grozdova

- dok je broj grozdova veći od 1 ili postoji par grozdova sličniji od praga:
  - pronađi najsličniji par grozdova
  - spoji te grozdove u novi grozd
  - izbriši originalne grozdove

Za pronalaženje događaja općenito se upotrebljavaju hijerarhijski algoritmi iz razloga što broj klasa nije *a priori* poznat, a parcijalni algoritmi postaju vrlo skupi kada se primjenjuju na problem nepoznatog broja grozdova.

Vremenski vrlo efikasni algoritmi su oni jednim prolaskom. Njihova je kompleksnost, naime, linearna, $O(n)$, a isto tako ne zahtjevaju *a priori* poznat broj grozdova. Za razliku od njih, hijerarhijskim algoritmima je vremenska kompleksnost polinomijalna, odnosno minimalno $O(n^2)$ [Wikipedia, 2009a].

Algoritam jednim prolaskom implementiran je na sljedeći način:

- ulaz je skup entiteta
- neka svaki entitet pripada jednom grozdu
- izračunaj sličnost svaka dva entiteta te je sortiraj padajuće
- za svaki par entiteta u popisu:
  - ako je sličnosti entiteta manja od praga, završi
  - spoji grozdove u kojima se nalaze entiteti

Pri implementaciji ovog algoritma treba paziti na praćenje puta određenog entiteta iz razloga što se udaljenost između entiteta računa jednokratno. Ukratko, treba postojati indeks entiteta koji pokazuje na grozd u kojemu se taj entitet trenutno nalazi.

2.5 Evaluacija algoritama za grožđenje

Metode evaluacije rezultata grožđenja korištene u ovom radu pretežno se preuzete iz [Manning et al., 2008b] te [Jain et al., 1999]. Cilj je je svake optimizacije postizanje optimalnog rješenja objektivne funkcije koja govori koliko je neko rješenje dobro. Objektivna funkcija u zadatku grožđenja formalizira cilj dobivanja grozdova s maksimalnom sličnosti unutar grozdova i minimalnom sličnosti između grozdova. Takav se kriterij naziva unutarnjim kriterijem kvalitete grožđenja. Činjenica je da taj unutarnji kriterij ne mora uvijek biti relevantan za neku primjenu grožđenja.

Alternativna metoda evaluacije je direktna evaluacija zadatka koji se grožđenjem pokušava riješiti. U primjeni grožđenja na pretraživanje informacija moguće je mjeriti vrijeme potrebno da pojedini korisnici pronađu traženu informaciju ovisno o tome koji je algoritam grožđenja primjenjen. Ovakva evaluacija u objektivnije mjeri uspješnost nekog algoritma u rješavanju danog problema, no skupa je i kompliciranа.

Iz tog se razloga najčešće pribjegava upotrebi surogata direktne evaluacije - zlatnog standarda. Zlatni je standard uzorak podataka na kojemu je problem koji algoritam treba riješiti riješen od strane najčešće više osoba, tzv. ljudskih označitelja. Više osoba označava isti uzorak kako bi bilo moguće izmjeriti dvije vrijednosti - pogrešku pri označavanju te dogovor između označitelja [Agirre and Edmonds, 2007].

Prva mjera - pogreška - govori o tome koliko je vjerojatnost ljudske pogreške pri ručnom označavanju. Tako je, primjerice, u zadatku označavanja vrste riječi točnost zlatnog standarda oko 97 posto što za posljedicu ima da nije moguće tvrditi da neki algoritam taj zadatak rješava bolje od tog postotka iz razloga što nije moguće znati pogađa li algoritam točno ili pogrešno rješenje koje prelazi stupanj točnosti zlatnog standarda. Taj problem je još dublji u slučaju nadziranog učenja iz razloga što algoritam i uči na podacima koji sadrže tu razinu pogreške. Ova mjera predstavlja jednu vrstu stropa (engl. *ceiling*), mjere koja pretpostavlja maksimalno moguće rješenje nekog problema.

Druga mjera ukazuje na kompleksnost zadatka, odnosno mogućnost da

se taj zadatak jednoznačno riješi od strane čovjeka. Tako je, primjerice, u zadatku razrješavanja leksičke višeznačnosti vrlo često dogovor između označitelja samo 60 posto [Agirre and Edmonds, 2007]. Ta vrijednost također predstavlja vrstu stropa određenog zadatka iz razloga što se uspoređujući rezultat algoritma sa zlatnim standardom za određeni postotak oznaka u zlatnom standardu ne može biti siguran da neki drugi označitelj ne bi taj zadatak riješio na neki drugi način, moguće i onakav na koji je to riješio algoritam.

Stropovi općenito predstavljaju gornje granice rješenja (engl. *upper bound*). Mjere koje predstavljaju doljnju granicu rješenja (engl. *lower bound*) nazivaju se osnovnim pristupom (engl. *baseline*). Kao osnovni pristup često se koristi industrijski standard te se uvodeći novu metodu istražuje daje li ta metoda bolji rezultat od one koja je trenutno u upotrebi. Čest je slučaj da nove, kompleksnije metode, zapravo i ne popravljaju rezultat u odnosu na one ustaljene, jednostavnije.

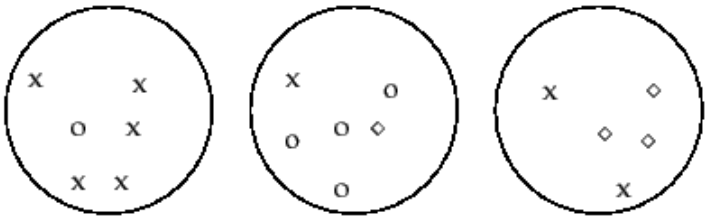
Kada industrijski standard ne postoji, kao osnovni pristup često se koriste i slučajna rješenja ili pak najvjerojatnija tj. najčešća rješenja te se promatra uspijeva li algoritam riješiti zadatak bolje nego što je to vjerojatno čistim slučajem, odnosno odabirom najčešćeg rješenja.

Direktnom evaluacijom na nekom zadatku ili zlatnim standardom određuje se vanjski kriterij kvalitete grožđenja. Najčešće mjere vanjskog kriterija grožđenja su

- čistoća
- normalizirana međusobna informacija
- rand indeks
- F mjera

Čistoća je najjednostavnija i najintuivnija evaluacijska mjera. Normalizirana međusobna informacija ima pozadinu iz teorije informacije. Rand indeks ima tu prednost što kažnjava lažno pozitivne i lažno negativne slučajeve.

Slika 2.6: Primjer rezultata grožđenja za prikaz evaluacijskih mjera



F mjera dodatno podržava težinsko faktoriranje između te dvije vrste pogrešaka. U nastavku su detaljnije prikane značajke svake pojedine mjere. Usto su prikazani njihovi izračuni na primjeru prikazanom na slici 2.6 iz [Manning et al., 2008b].

2.5.1 Čistoća

Čistoća je najintuivnija evaluacijska mjera za grozdove koja mjeri koliko su grozdovi "čisti" s obzirom na klase elemenata kojima elementi u pojedinim grozdovima pripadaju. Točnije, svakom se grozdu dodjeljuje broj najzastupljenijih pripadnika neke klase. Čistoća se računa kao broj tih najzastupljenijih primjeraka neke klase u pojedinom grozdu podijeljen s brojem elemenata u grozdovima.

Formalnije, za skup grozdova $G = \{g_1, ...g_k\}$ i skup klasa $K = \{k_1, ...k_m\}$ čistoća je

$$cistoca(G, K) = \frac{1}{N} \sum_{k} \max_{m} |g_k \cap k_m| \tag{2.30}$$

gdje je N broj entiteta u skupu entiteta E. Opisno, za svaki grozd iz G računamo presjek s klasama iz K te uzimamo veličinu onog rezultata koji je najveći s obzirom na klase te računamo sumu tih rezultata. Sumu najvećih presjeka svakog grozda s klasama dijelimo s brojem entiteta u grozdovima.

Očita loša strana ove mjere, unatoč njenoj intuitivnosti, jest ta da ne kažnjava prekomjerno stvaranje grozdova. Taj problem je aktualan u zadacima u kojima broj grozdova nije poznat *a priori*, poput onoga kojim se bavi

ova disertacija. Ova mjera vraća vrijednosti između 1 i blizu 0. Problem je što će čistoća uvijek vratiti 1 kada je svaki entitet u vlastitom grozdu.

2.5.2 Normalizirana međusobna informacija

Glavna mana prethodno opisane evaluacijske mjere čistoće je nekažnjavanje stvaranja prevelikog broja grozdova što u ekstremnom slučaju vodi do činjenice da će grožđenje biti smatrano savršenim ako se svaki entitet nalazi u vlastitom grozdu.

Kako bi se tom problemu doskočilo, potrebno je uvesti mjeru koja čini kompromis između čistoće grožđenja i broja grozdova. Takva je mjera normalizirana međusobna informacija. Formalni zapis te mjere je sljedeći:

$$NMI(G,K) = \frac{I(G,K)}{[H(G)+H(K)]/2} \tag{2.31}$$

I je međusobna informacija koja mjeri količinu informacije koju nosi poredak entiteta u G s obzirom na pripadnost klasama u K, odnosno koliko saznajemo o pripadnosti entiteta klasama iz rasporeda u grozdovima. Njen je formalni zapis sljedeći:

$$I(G,K) = \sum_k \sum_j P(g_k \cap k_m) \log \frac{P(g_k \cap k_m)}{P(g_k)P(k_m)} \tag{2.32}$$

$$= \sum_k \sum_j \frac{|g_k \cap k_m|}{N} \log \frac{N|g_k \cap k_m|}{|g_k||k_m|} \tag{2.33}$$

gdje je $P(g_k)$ vjerojatnost da dokument pripada grozdu k, $P(k_m)$ vjerojatnost da dokument pripada klasi m te $P(g_k \cap k_m)$ vjerojatnost da dokument pripada grozdu k i klasi m. Iz formule 2.32 slijedi 2.33 ako se vjerojatnost računa pomoću procjene najveće vjerojatnosti.

H je entropija, odnosno mjera nesigurnosti čiji je formalni zapis sljedeći:

$$H(G) = -\sum_k P(g_k) \log P(g_k) \tag{2.34}$$

$$= -\sum_k \frac{|g_k|}{N} \log \frac{|g_k|}{N} \tag{2.35}$$

gdje iz 2.34 ponovno slijedi 2.35 ako se vjerojatnost računa pomoću procjene najveće vjerojatnosti.

Kao što je već rečeno, mjera međusobne informacije mjeri količinu informacije koju dobivamo znanjem o rasporedu entiteta u grozdovima s obzirom na stvarni raspored po klasama. Minimalna vrijednosti međusobne informacije je 0 u slučaju da je raspored u grozdovima slučajan s obzirom na pripadnost klasama. U tom slučaju, informacija o tome u kojem se grozdu nalazi neki entitet ne daje nam nikakvu informaciju o tome kojoj klasi taj entitet pripada. Maksimalna mjera međusobne informacije se postiže u slučaju kada raspored u grozdovima točno odgovara pripadnosti klasama, no to uključuje i slučaj kada su, primjerice, entiteti iste klase razbijeni u više grozdova. Tako, u slučaju da je $|K| = N$, odnosno broj grozdova jednak broju entiteta, postiže se maksimalna međusobna informacija. Naime, informacija o tome kojem grozdu pripada entitet govori sve o njegovoj pripadnosti klasi. Kod međusobne informacije se, dakle, ponovno susreće problem koji je glavni problem u mjeri čistoće - nemogućnost kažnjavanja stvaranja prekomjernog broja grozdova.

Iz tog se razloga upotrebljava normalizirana međusobna informacija koja se računa kao međusobna informacija normalizirana entropijom. Ta normalizacija rješava problem zato što entropija kao mjera nereda raste s brojem grozdova. Na primjer, $H(K)$ dostiže maksimalni $\log N$ za $|K| = N$ što osigurava niski NMI u slučaju da je $|K| = N$, odnosno da se svaki entitet nalazi u vlastitom grozdu. Izraz za normalizaciju je $[H(G) + H(K)]/2$ iz razloga što on predstavlja usku gornju granicu na izraz $I(G, K)$ što omogućuje da rezultat mjere normalizirane međusobne informacije uvijek bude između 0 i 1.

2.5.3 Rand indeks

Alternativa prethodno opisanoj mjeri iz područja teorije informacije jest da se grožđenje interpretira kao niz odluka, po jedna za $N(N-1)/2$ parova entiteta. Cilj grožđenja u tom slučaju je svrstavanje dvaju entiteta u isti grozd ako i samo ako su slični.

Istinito pozitivno (IP) rješenje dodjeljuje dva slična entiteta u isti grozd, dok istinito negativno (IN) rješenje dodjeljuje dva neslična entiteta u različite grozdove. Uz ove točne odluke moguće je ostvariti dvije vrste pogrešaka - lažno pozitivno (LP) rješenje pri kojem se dva neslična entiteta dodijele istom grozdu te lažno negativno (LN) rješenje kada se dva slična entiteta dodijele dvama grozdovima.

Rand indeks mjeri postotak odluka koje su istinite. Time je Rand indeks mjera identična klasičnoj mjeri točnosti u području obrade prirodnog jezika. Formalno ona mjeri

$$RI = \frac{IP + IN}{IP + IN + LP + LN} \tag{2.36}$$

Kako bi izračunali rand indeks za primjer prikazan na slici 2.6 potrebno je izračunati broj parova koji u danim grozdovima tvore parove entiteta određene istinosne vrijednosti. To je problem koji rješavaju permutacije s ponavljanjem, odnosno za računanje parova entiteta istinitih pozitivnih rješenja u sva tri grozda potrebno je izračunati

$$IP = \binom{5}{2} + \binom{4}{2} + \binom{3}{2} + \binom{2}{2} = 20 \tag{2.37}$$

iz razloga što je u grozdovima nađeno $5 + 4 + 3 + 2$ slučajeva više od jednog entiteta iste klase.

Kako bi izračunali broj preostalih parova moramo izračunati IP+NP, odnosno općeniti broj parova u grozdovima:

$$IP + LP = \binom{6}{2} + \binom{6}{2} + \binom{5}{2} = 40 \tag{2.38}$$

Iz 2.37 i 2.38 slijedi

$$LP = IP + LP - IP = 40 - 20 = 20 \tag{2.39}$$

Izraz $IN + LN$, odnosno broj negativnih parova je izračunljiv preko broja kombinacija entiteta iz različitih grozdova, odnosno izrazom

$$IN + LN = 6 * 6 + 6 * 5 + 6 * 5 = 96 \tag{2.40}$$

dok je broj lažnih negativnih izračunljiv preko broja kombinacija entiteta u pogrešnim grozdovima sa svim preostalim entitetima te klase, odnosno izrazom

$$LN = 5 * 1 + 5 * 2 + 1 * 2 + 1 * 4 + 1 * 3 = 24 \tag{2.41}$$

Iz 2.40 i 2.41 jednostavno je izračunati broj lažnih pozitivnih izrazom

$$IN = IN + LN - LN = 96 - 24 = 72 \tag{2.42}$$

Na temelju dobivenih podataka moguće je oblikovati tablicu slučajeva kakva je prikazana tablicom 2.1

Točnost, odnosno rand indeks je prema izrazu 2.36 u tom slučaju

$$RI = \frac{20 + 72}{20 + 24 + 20 + 72} = 0.676 \tag{2.43}$$

Tablica 2.1: Tablica slučajeva primjera grožđenja za prikaz evaluacijskih mjera

	isti grozd	različiti grozd
ista klasa	IP=20	LN=24
različita klasa	LP=20	IN=72

2.5.4 Preciznost, potpunost i F mjera

Mogući nedostatak rand indeksa jest da jednako težinski faktorira lažne pozitivne i lažne negativne, dok je, ovisno o namjeni rezultata grožđenja, povremeno važnije imati čišće, a povremeno potpunije grozdove. Iz tog se razloga često primjenjuje F mjera - klasična mjera u području obrade prirodnog jezika koja omogućuje težinsko faktoriranje te dvije vrste pogrešaka. Osnovne vrijednosti koje ulaze u F mjeru su preciznost i potpunost, odnosno vrijednosti koje mjere koliko je precizno neki zadatak obavljen, odnosno koliko potpuno. Izrazi za računanje preciznosti i potpunosti su

$$PR = \frac{IP}{IP + LP} \tag{2.44}$$

$$PO = \frac{IP}{IP + LN} \tag{2.45}$$

F mjera koristi faktor β koji u slučaju da je $\beta > 1$ više naglašava potpunost dok, u slučaju da je $\beta < 1$ više naglašava preciznost. Izraz za F mjeru je

$$F_\beta = \frac{(\beta^2 + 1) * PR * PO}{\beta^2 * PR + PO} \tag{2.46}$$

Vrijednosti mjera za evaluaciju grožđenja za naš primjer sa slike 2.6 moguće je vidjeti u tablici 2.2.

Tablica 2.2: Vrijednosti evaluacijskih mjera primjera grožđenja za prikaz evaluacijskih mjera

mjera	vrijednost
čistoća	0.706
NMI	0.365
RI	0.676
preciznost	0.5
potpunost	0.455
F_5	0.456

Poglavlje 3

Nacrt istraživanja

U ovom poglavlju bit će dan nacrt provedenog istraživanja. Prvi će dio poglavlja govoriti o oblikovanju uzorka za istraživanje te o nekim *in vitro* mjerenjima nad uzorkom. U drugom će dijelu poglavlja biti prikazan niz varijabli i slobodnih parametara čijim se vrijednostima eksperimentira te način evaluacije tih postupaka kao i gornja i donja granica istraživanja.

3.1 Jezični uzorak

Najčešća metoda evaluacije algoritama za obradu prirodnog jezika jest ona usporedbe rezultata sa zlatnim standardom - ručno označenim uzorkom. U slučaju pronalaženja događaja taj se zlatni standard sastoji od skupa skupova identifikatora dokumenata u kojima svaki skup sadrži dokumente koji opisuju isti događaj.

Uzorak dokumenata koji se koristi u ovom doktorskom radu ustupljen je od strane Zavoda za poslovna istraživanja [ZAPI, 2009]. Razdoblje u kojemu su dokumenti objavljivani jest od 4. do 6. svibnja. 2008. godine. Radi se sveukupno o 2,486 dokumenata prikupljenih sa 17 portala. Razdioba dokumenata prema portalu objave prikazana je u tablici 3.1.

Odabrani su dokumenti objavljeni u tri uzastopna dana iz sljedećih razloga:

- ideja je stvoriti tri uzorka te vršiti eksperimente na jednom ili više

Tablica 3.1: Čestotna razdioba dokumenata s obzirom na portal na kojemu su dokumenti objavljeni

portal	broj dokumenata
javno.com	389
totalportal.hr	311
business.hr	225
dnevnik.hr	207
jutarnji.hr	175
index.hr	174
tportal.hr	146
vecernji.hr	134
mojportal.hr	120
net.hr	90
glas-slavonije.hr	89
poslovni.hr	85
seebiz.eu	77
rtl.hr	74
nacional.hr	66
sutra.hr	64
liderpress.hr	60
sveukupno	2486

njih ovisno o tome koliko je varijabla nad kojom se eksperimentira zahtjevna, odnosno koliko je razlika u evaluacijskim mjerama s obzirom na vrijednost istraživane varijable jednoznačna

- činjenica da se uzorci nastavljaju jedan na drugi omogućuje *in vitro* i *in vivo* evaluaciju prve heuristike prikazane u poglavlju 1.2, odnosno evaluaciju na većem, odnosno manjem uzorku uzastopno objavljivanih vijesti

U nastavku će biti opisano označavanje uzorka, osnovna statistička analiza označenog uzorka te *in vitro* istraživanje dviju pretpostavljenih heuristika.

3.1.1 Označavanje uzorka

Označavanje odabranog uzorka je potrebno provesti kako bi se nad tim podacima mogli vršiti eksperimenti te njihovi rezultati evaluacijskim mjerama usporediti upravo s tim uzorkom označenim od strane ljudskog označitelja. Algoritam pri eksperimentu, naravno, ne korisiti oznake koje ukazuju na to koji dokumenti opisuju isti događaj.

Kako bi se označio uzorak korišten je softver razvijen u sklopu diplomskog rada studenta Nikole Bakarića [Bakarić, 2009]. Aplikacija je razvijena u programskom jeziku Python te posjeduje i grafičko sučelje razvijeno pomoću alata TkInter. Ideja koja stoji iza aplikacije jest da bi ovakvo označavanje uzorka trebalo biti što jednostavnije i preglednije. Naime, pri označavanju ovakvog uzorka potrebno je u velikom popisu dokumenata (u ovom slučaju popisu od 2,486 dokumenata) pronaći sve dokumente koji govore o istom događaju te ih označiti istom oznakom. Ovakav bi zadatak rješavajući ga pregledavajući samo tekstualne datoteke bio skoro nemoguć.

Iz tog je razloga izvršena predobrada podataka te se softveru za označavanje osim samog skupa dokumenata daje i popis međusobnih sličnosti dokumenata. Softver, dakle, kao ulaz očekuje dvije datoteke

- prvu koja sadržava popis dokumenata (id, naslov članka, tekst članka)
- te drugu koja sadrži predložene veze među dokumentima, odnosno udaljenost između svaka dva dokumenta ne manju od 0.1.

Te udaljenosti, odnosno predložene veze služe kao smjernice ljudskim označiteljima kako ne bi za svaki dokument morali pregledavati n-1 dokumenata, odnosno izvršiti n*(n-1) usporedbi (u ovom slučaju 6,177,710 usporedbi).

Mjere sličnosti dokumenata računate su tako da su dokumenti opojavničeni, pretvoreni u mala slova te su od njih oblikovani vektori svojstava. Kao svojstva su uzimane sve pojavnice koje počinju nekim slovom. Mjera težine određenog svojstva za određeni vektor računata je pomoću TF-IDF mjere. Sličnost između dokumenata računata je kosinusnom mjerom. Kao što je već napomenuto, u veze su uključeni svi parovi vektora, odnosno dokumenata čija sličnost nije manja od 0.1.

Problem ovakvog pristupa, odnosno ovakvog računanja preliminarnih rezultata i prikaza podatka označitelju jest sljedeći - ako postoji funkcija $u(d_1, d_2)$ koja vraća udaljenost prvog i drugog dokumenta, te ako vrijede izrazi $u(d_1, d_2) = 0.2$, $u(d_1, d_3) = 0.5$ i $u(d_2, d_3) = 0.08$ te ako pretpostavimo da sva tri dokumenta d_1, d_2 i d_3 govore o istom događaju pa ako se prvotno ponude dokumenti slični dokumentu d_2, u popis neće biti moguće dodati d_3 iz razloga što je ispod praga sličnosti dokumenata. Ovaj će problem potencijalno postati očit označitelju tek kad pregleda dokumente slične d_1, odnosno d_3.

Djelomično moguće rješenje ovog problema jest prikazati dokumente redoslijedom suprotnim broju dokumenata koji su sličiniji od 0.1. Time se vjerojatnost neuključivanja nekog dokumenta u grupu smanjuje, no ne i eliminira.

Potpuno moguće rješenje ovog problema bilo bi izvršiti ne samo računanje mjera udaljenosti nego izvršiti i neki oblik grožđenja te označitelju ustupiti već grupirane podatke. Loša strana ovog pristupa je ta što se kosi sa željom da se označitelju daju što siroviji podaci kako označitelj ne bi bio pod utjecajem rezultata algoritama koji se tim istim uzorkom kane provjeravati.

Ovaj je problem primijećen tijekom označavanja te je rješavan na način da su dokumenti koji slučajno nisu odmah dodani u grozd dodani naknadno.

Iz tog su razloga operacije koje su kroz softver omogućene označitelju sljedeće:

- pregledavanje određenog dokumenta
- pregledavanje dokumenata redoslijedom sličnosti uz prikaz stupnja sličnosti
- oblikovanje novog grozda spajanjem prvog sličnog dokumenta s trenutnim dokumentom
- dodavanje daljnjih dokumenata u grozd
- uklanjanje nekog dokumenta iz grozda
- uklanjanje grozda uklanjanjem posljednje veze između trenutnog dokumenta i njemu sličnih dokumenata
- uklanjanje dokumenta iz popisa dokumenata ako je identičan nekom drugom dokumentu (u pravilu sličnost iznad 0.95) ili pak nije potpun (pogreška pri prikupljanju ili objavi dokumenta)

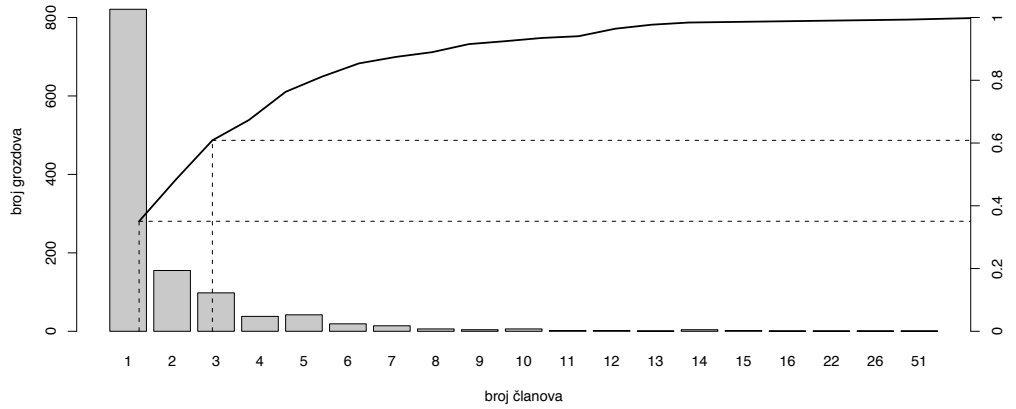
Mogućnosti koje su navedene omogućene su za svaki dokument, bio on već dodan u neki grozd ili ne. Posljednja mogućnost, uklanjanje dokumenta iz popisa dokumenata omogućena je zato što se u popisu dokumenata često nalaze identični dokumenti prikupljeni na alijasima njihovih putanja (primjerice "http://www.seebiz.eu", odnosno "http://seebiz.eu") te zato što ponekad nije sačuvan cijeli dokument. Tom su metodom izbačena 83 dokumenta te ih je u uzorku preostalo 2,403.

3.1.2 Analiza označenog uzorka

U ovom se potpoglavlju vrši osnovna statistička analiza označenog uzorka. Za očekivati je da će takva statistička analiza dati uvid u fenomen opisa događaja u dokumentima objavljenima na internetu te da će dati smjernice za oblikovanje algoritma za pronalaženje događaja, tj. istraživanje koje slijedi.

Uzorak se poslije označavanja sastoji od 2,403 dokumenta koji su raspoređeni u 1,218 grozdova, odnosno u 2,403 dokumenta, kojima se opisuje 1,218 različitih događaja. Čestotna razdioba dokumenata prema broju dokumenata u grozdovima prikazana je u tablici 3.2. Iz tih je podataka vidljivo

Slika 3.1: Odnos broja članova u grozdovima i broja grozdova te kumulativna funkcija dokumenata s obzirom na broj članova u grozdovima u kojima se nalaze



kako je većina dokumenata sama u grozdu te da je odnos ranga grozdova s obzirom na broj dokumenata koji sadrže zipfijanski. Broj dokumenata u grozdu je moguće poistovjetiti s brojem pojavljivanja pojavnica, dok se rang grozda s obzirom na broj dokumenata može poistovjetiti s rangom pojavnice s obzirom na njenu čestotu. Dakle, većina grozdova sastoji se od jednog dokumenta kao što se većina pojavnica pojavljuje vrlo rijetko, a samo rijetki grozdovi imaju velik broj dokumenata kao što se mali dio pojavnica pojavljuje vrlo često.

U kontekstu pronalaženja događaja ovaj rezultat navodi na princip kako u spajanju dokumenata u grozd treba biti vrlo konzervativan iz razloga što se 34 posto dokumenata nalazi samo u grozdu. Grozdovi do veličine tri tako sadržavaju 45 posto dokumenata.

Odnos broja članova u grozdovima i broja takvih grozdova prikazan je na slici 3.1. Krivuljom je prikazana kumulativna funkcija dokumenata s obzirom na broj članova u grozdovima u kojima se ti dokumenti nalaze. Iz te krivulje je vidljivo da se u grozdovima od jednog člana nalazi oko 35% dokumenata dok se u grozdovima do tri člana nalazi više od 60% dokumenata.

Pitanje koje se logički nameće jest je li čestota objavljivanja novosti o događaju koje ne prate drugi mediji zavisna o portalu na kojemu se ta novost

Tablica 3.2: Čestotna razdioba grozdova prema broju dokumenata u grozdovima

broj dokumenata	broj grozdova
1	821
2	155
3	98
4	38
5	42
6	19
7	14
8	6
9	4
10	6
11	2
12	2
13	1
14	4
15	2
16	1
22	1
26	1
51	1

Tablica 3.3: Čestotna razdioba dokumenata koji opisuju specifičan događaj s obzirom na portal na kojemu su objavljeni

portal	postotak
sutra.hr	0.766
glas-slavonije.hr	0.605
net.hr	0.518
jutarnji.hr	0.44
javno.com	0.412
poslovni.hr	0.393
seebiz.eu	0.351
totalportal.hr	0.302
index.hr	0.287
dnevnik.hr	0.278
vecernji.hr	0.269
liderpress.hr	0.25
business.hr	0.25
tportal.hr	0.247
rtl.hr	0.231
nacional.hr	0.227
mojportal.hr	0.225
hrt.hr	0.143

objavljuje, odnosno je li fenomen objavljivanja specifičnih novosti uniformno razdjeljen. Iz tog je razloga u tablici 3.3 prikazana čestotna razdioba takvih događaja ovisno o portalu na kojemu su objavljeni.

Iz podataka se može zaključiti kako očita pravilnost ne postoji. Popularni internetski portali kao i novinski portali slučajno su raspoređeni u ovom popisu te se ne može izvesti nikakav jasan zaključak. Mogući problem ovakvog prikaza podataka jest što je ovdje prikazano koliko neki portal sudjeluje u masi dokumenata koji opisuju specifične događaje, a da nije u obzir uzeta veličina tog portala. No, poznavajući zastupljenost portala među dokumentima u uzorku, odnosno podatke iz tablice 3.1, moguće je zaključiti kako zavisnost varijable portala i količine objavljenih dokumenata koji opisuju specifičan događaj ne postoji.

3.1.3 Neki problemi, odnosno odluke donesene pri označavanju

Pri ručnom označavanju uzorka primijećeni su određeni problemi. Osnovni problem redovito jest odrediti granicu nekog događaja. Primjerice, u slučaju paljenja hrvatske zastave od strane slovenskih vojnika u dokumentima se govori o sljedećim poddogađajima:

- dokumenti koji govore o paljenju zastave od strane slovenskih vojnika
- dokumenti koji prenose očitovanje slovenskog Ministarstva obrane da slovenski vojnici nisu palili zastavu
- dokumenti koji govore o tome da zastavu nije zapalio slovenski vojnik, već njegovi prijatelji
- dokument koji govori o prigovorima slovenskih turista na prvomajske praznike provedene u Hrvatskoj te spominjanje incidenta sa zastavom u posljednjem odlomku

Kako je prihvaćena pretpostavka da pojedini dokument opisuje jedan događaj, odnosno da se u ovom doktorskom radu dokument smatra jediničnom mjerom informacije, nemoguće je ove poddogađaje proglasiti događajima.

Redoviti je problem također situacija kada se u dokumentu d_1 izvještava o događajima do_1 i do_2, dok se u dokumentu d_2 izvještava o događaju do_1, a u dokumentu d_3 o događaju do_2. U svim se slučajevima poput toga ljudski označitelji vode pravilom čestote, odnosno pokušavaju dokumente podijeliti na događaje na način da što češće bude zadovoljen uvjet da je u jednom grozdu maksimalno puno opisa jediničnog događaja. Tako će se u slučaju prethodnog primjera, ako se bitno češće o događajima do_1 i do_2 izvještava u istom dokumentu, ti događaji smatrati jednim jediničnim događajem.

Stvarni primjer neoptimalnog preklapanja događaja i dokumenata kakav je prije opisan jest i slučaj sastanka Ive Sanadera, poslodavaca i sindikata. U nizu dokumenata koji opisuju taj događaj izvještava se o sljedećim poddogađajima:

- socijalno partnerstvo
- konkurentnost gospodarstva
- devalvacija kune
- fiskalna politika
- politika zapošljavanja
- određivanje minimalca
- cijena benzina
- saborske mirovine

Isto se tako, primjerice, podatak da je Slaven Bilić produžio ugovor i objavio popis reprezentacije te čuđenje zbog poziva Sharbiniju smatraju istim događajem s reakcijom Ladića, Sharbinija, Leke i drugih.

Samo se u rijetkim situacijama, kada, primjerice samo jedan dokument spaja dva događaja, ti događaji smatraju odvojenima te se dotični dokument priklanja onom događaju o kojemu se više govori u dokumentu.

3.1.4 Testiranje heuristika na zlatnom standardu

U nastavku se istražuju heuristike uvedene u poglavlju 1.2, i to *in vitro*, dakle nad podacima, a ne *in vivo*, odnosno u nekoj stvarnoj situaciji. Dvije pretpostavljene heuristike su sljedeće:

1. podaci o jednom događaju se pretežno objavljuju u istom danu
2. isti događaj se ne objavljuje dva puta na istom portalu

Prva heuristika

Prva heuristika pretpostavlja da su članci koji opisuju isti događaj objavljeni istog dana. Prvo mjerenje koje je učinjeno na uzorku jest računanje postotka dokumenata koji nisu sami u grozdu te su objavljeni istog dana kao i većina dokumenata u tom grozdu.

Osnovni problem ovakve analize je taj što se mjerenje vrši nad onime što jest u uzorku, dok se ne zna što u uzorak zbog njegovog vremenskog ograničenja nije ušlo. Naime, mjereći koji je najčešći dan objave u nekom grozdu, izostavlja se sve dane koji nisu u uzorku, odnosno zanemaruje se dokumente koji su objavljeni prije 4., odnosno poslije 6. svibnja. Iz tog se razloga nude različiti pristupi provjere prve heuristike:

1. mjerenje se može vršiti nad svim grozdovima s više od jednog dokumenta, tako je za očekivati najpovoljniji rezultat s obzirom na prvu heuristiku, naime, zanemaruje se činjenica da se time povećava vjerojatnost da će biti istraživani događaji o kojima se objavljuje i izvan vremenskog prozora od 4. do 6. svibnja

2. mjerenje se može vršiti nad grozdovima čija je većina dokumenata objavljena 5. svibnja, ovakvim se pristupom može očekivati srednji rezultat, no on predstavlja najbolju logičku procjenu, naime, tako postoji najveća vjerojatnost da su uzeti događaji zaista vezani uz 5. svibnja

3. mjerenje se može vršiti i nad grozdovima koji sadrže barem jedan članak objavljen 5. svibnja, ovaj će pristup generirati najnepovoljniji rezultat, izvedbeno je to najbolja procjena iz razloga što se u praksi tako odabiru dokumenti

Rezultati mjerenja ovim trima pristupima prikazani su u tablici 3.4.

Što se broja grozdova tiče, rezultati su kao i očekivani - najviše se grozdova istražuje ako se uzmu u obzir svi višečlani grozdovi iz sva tri dana. Najmanje grozdova obuhvaća pristup gdje se odabiru samo grozdovi koji sadrže većinu dokumenata objavljenih 5. svibnja. Treći pristup sadrži srednji broj grozdova, on, naime, uključuje sve višečlane grozdove koji posjeduju dokument objavljen 5. svibnja.

Broj dokumenata također odgovara pretpostavkama. On zapravo prati prijeopisani broj grozdova.

Što se tiče postotka zadovoljavanja prve heuristike, ovdje su rezultati isto kao i očekivani. Kao najpovoljniji pristup za dokazivanje heuristike pokazuje

Tablica 3.4: Mjerenja nad uzorkom vezana uz prvu heuristiku izvršena trima definiranim načinima

	1	2	3
broj grozdova	397	167	210
broj članaka objavljenih isti dan	1,358	544	742
sveukupni broj članaka	1,582	650	957
postotak	0.858	0.837	0.775

se prvi iz razloga što on odabire sve višečlane grozdove. Njegov se rezultat može smatrati preoptimističnim iz razloga što zahvaća zasigurno mnoge događaje koji imaju članove objavljene prije, odnosno poslije vremenskog prozora uzorka te to rezultat čini pretjerano optimističnim. Drugi pristup daje srednji rezultat te se može, što se tiče prirode problema, smatrati najrealnijim. Tom su metodom odabrani grozdovi, odnosno događaji za koje se može pretpostaviti da su zaista povezani s dotičnim datumom. Treći pristup daje najslabiji rezultat, no za pretpostaviti je da je taj pristup najrealniji što se tiče buduće metode rješavanja problema. Naime, primjenom prve heuristike pronalaženje događaja vršit će se nad svim dokumentima koji su objavljeni u određenom vremenskom prozoru.

Generalno je moguće zaključiti kako je drugi pristup najrealniji što se samog fenomena tiče te da on dokazuje ispravnost heuristike u 84 posto slučajeva.

Nadalje je istraženo postoji li zavisnost varijable veličine grozda od činjenice u kojem postotku zadovoljava prvu heuristiku. Za očekivati je da će veći grozdovi imati više dokumenata objavljenih drugih dana. Kao što je prije odlučeno, primijenjena je druga metoda odabira grozdova, odnosno odabrani su samo oni grozdovi čiji su dokumenti većinom objavljeni 5. svibnja. Rezultati ovog eksperimenta prikazani su u tablici 3.5.

Iz podataka prikazanih u tablici nije moguće uočiti povezanost tih dviju varijabli. Numerički dokaz nepovezanosti je mjera korelacije varijabli veličine grozda (VG) i postotka pozitivnih dokumenata (PP), tj. dokumenata objavljenih istog datuma. Korelacija, naime, iznosi 0.02, dakle vrlo je niska.

Tablica 3.5: Zavisnost veličine grozdova od postotka zadovoljenja druge heuristike kroz veličinu grozda (VG), broj pozitivnih primjera (BP), broj primjera (B) i postotak pozitivnih primjera (PP)

VG	BP	B	PP
2	121	152	0.796
3	84	93	0.903
4	57	72	0.792
5	62	70	0.886
6	35	42	0.833
7	38	42	0.905
8	16	16	1.0
9	14	18	0.778
10	30	30	1.0
11	20	22	0.909
12	11	12	0.917
14	23	28	0.821
15	7	15	0.467
16	13	16	0.813
22	13	22	0.591

Tablica 3.6: Deset događaja s najmanjim postotkom novosti objavljenih u istom danu (VG - veličina grozda, PP - postotak pozitivnih)

događaj	VG	PP
Slovenci zapalili hrvatsku zastavu	15	0.4667
stanje novčanih fondova	4	0.5
pijanom britanskom paru oduzeta djeca u Portugalu	4	0.5
ruski zakon o ograničavanju stranih ulaganja	6	0.5
početak suđenja dr. Šimiću	22	0.591
Tereza Kesovija ne nastupa na HRF-u	5	0.6
započeta obrana bosanskih Hrvata	6	0.667
Mesić planira sazvati sjednicu Vlade	9	0.667
zabrana pušenja u domovima umirovljenika	7	0.715
bijeli štrajk u T-HT-u	14	0.714

Moguća argumentacija ovih podataka je ta da veći grozdovi zapravo predstavljaju opis događaja koji su očito važniji. Za manje se grozdove može pretpostaviti da opisuju manje važne događaje. Kako se može očekivati da će veći grozdovi biti raspršeniji tako se može i pretpostaviti da su manji grozdovi, tj. manje zanimljivi događaji manje aktualni te da njihovo objavljivanje nije toliko gusto.

Kako bi se dalje istražio razlog neobjavljivanja istog događaja u istom danu ispisano je 10 događaja s najmanjim postotkom objavljivanja u istom danu uz uvjet da grozd što sadrži dokumente koji opisuju događaj sadrži četiri ili više dokumenata. Rezultati tog mjerenja prikazani su u tablici 3.6.

Iz rezultata je vidljivo kako se kod manjih grozdova radi o sporednim novostima koje nikako nisu dovoljno aktualne da budu obavezno objavljene što ranije. Kod većih grozdova se radi zapravo o slučajevima poput paljenja hrvatske zastave od strane Slovenaca te početka suđenja dr. Šimiću - događajima koji se protežu na nekoliko dana i vrlo je teško do nemoguće, pa i pogrešno, te događaje podijeliti na više zasebnih događaja.

Nadalje je istražena gustoća objavljivanja novosti o nekom događaju. Mjereno je prosječno vrijeme između dvije novosti u cijelom uzorku. Ono iznosi 200.95 minuta, odnosno nešto više od 3 sata sa standardnom devijaci-

Tablica 3.7: Odnos veličine grozdova (VG) i prosječne udaljenosti (PU) između dvije objavljene novosti

VK	PU
2	409.947
3	203.127
4	273.625
5	149.551
6	201.854
7	112.703
8	73.925
9	164.832
10	40.553
11	99.32
12	147.759
14	134.235
15	149.875
16	140.398
22	140.993

jom od 328.18.

Što se tiče aritmetičke udaljenosti prve i posljednje novosti u grozdu, ona iznosi 434.59 minute, odnosno nešto više od 7 sati sa standardnom devijacijom od 430.22.

Nadalje je istražena zavisnost veličine grozdova i gustoće objavljivanja novosti. Rezultati su prikazani u tablici 3.7.

Moguće je primijetiti kako je prosječna udaljenost prilično konstantna uz iznimku malih grozdova gdje prosječna udaljenost raste. Ovi podaci su još jedan dokaz o tome kako je tendencija da manji grozdovi opisuju očito manje aktualne događaje čime prosječna udaljenost objavljivanja između dokumenata raste.

Nadalje je prikazana zavisnost veličine grozda i udaljenosti prve i posljednje novosti u grozdu. Ti su podaci prikazani u tablici 3.8.

Iz ovih rezultata može se zaključiti da je prosječna udaljenost između prve i posljednje novosti konstanta što je, uzevši u obzir broj dokumenata u tim

Tablica 3.8: Odnos veličine grozdova (VG) i prosječne udaljenosti (PU) između prve i posljednje objavljene novosti

VG	PU
2	409.947
3	359.802
4	500.874
5	495.348
6	597.843
7	676.219
8	517.475
9	598.658
10	364.978
11	273.2
12	185.35
14	305.058
15	658.25
16	665.967
22	80.85

Tablica 3.9: Rezultati analize uzorka s obzirom na drugu pretpostavljenu heuristiku

broj grozdova	397
sveukupni broj članaka	1,582
broj članaka s jedinstvenog portala u grozdu	1,358
postotak članaka s jedinstvenog portala u grozdu	0.858

grozdovima, još jedan dokaz kako su manji grozdovi širi, a veći uži.

Zaključno se o prvoj heuristici može reći da ona analizom uzorka podataka prilično čvrsto stoji, tj. da članci pokazuju tendenciju da budu objavljeni unutar istog dana. Prva heuristika vrijedi u slučaju više od 80 posto višečlanih grozdova. Što se raspršenosti dokumenata unutar grozdova tiče, jaka pravilnost je da su manji grozdovi u pravilu raspršeniji od velikih grozdova. To navodi na zaključak kako manji grozdovi opisuju manje bitne događaje pa se time o događaju u različitim izvorima ne izvještava toliko brzo, dok se kod većih grozdova radi o važnijim događajima pa je i njihovo objavljivanje vremenski gušće.

Druga heuristika

Druga heuristika pretpostavlja da se o jednom događaju ne izvještava na više od jednog portala. Rezultati analize te heuristike na uzorku višečlanih grozdova prikazani su u tablici 3.9. Od 1,218 grozdova u uzorku, 821 sadrži samo jedan dokument te nije uključen u ovu analizu. Tako preostaje 397 grozdova. Od početnog broja dokumenata od 2,403 dokumenta ostaje ih 1,582. U tom uzorku broj članaka što je s jedinstvenog portala u grozdu jest 1,358. To znači da je otprilike 86 posto dokumenata u uzorku, zanemarivši dokumente koji se sami nalaze u grozdu, jedini dokument u grozdu s određenog portala, odnosno da u tom postotku slučajeva druga heuristika vrijedi.

Krajnji je zaključak vezan uz analizu dviju heuristika, odnosno njihovu *in vitro* evaluaciju da obje heuristike pokazuju vrlo jasne afirmacijske trendove te da su sigurno vrijedne daljnjeg empirijskog istraživanja na zadatku pronalaženja događaja, odnosno *in vivo* evaluacije.

3.2 Varijable koje se istražuju

Varijable koje se istražuju u ovom doktorskom radu su sljedeće:

- algoritam grožđenja (AG)
- mjera udaljenosti (MU)
- pretpostavljene heuristike (PH)
- mjera težine svojstava (MTS)
- metode određivanja i odabira svojstava
  - utjecaj interpunkcija (UI)
  - utjecaj veličina slova (UVS)
  - važnost naslova (VN)
  - isključivanje hapax legomena (IHL)
  - isključivanje funkcijskih riječi (IFR)
  - korjenovanje (K)
  - morfosintaktičko označavanje (MO)
  - prepoznavanje osobnih imena (POI)
  - statistički značajni digrami (SZD)
- utjecaj veličine korpusa na IDF mjeru (VIDF)

Većina ovih varijabli je nominalna. Primjerice, varijabla algoritma grožđenja sastoji se od četiri moguće vrijednosti - hijerarhijskog algoritma potpunom vezom, hijerarhijskog algoritma prosječnom vezom, hijerarhijskog algoritma pojedinačnom vezom te algoritma jednim prolaskom pojedinačnom vezom.

Određeni se dio varijabli zapravo sastoji i od podvarijabli, odnosno parametara. Tako se, uspoređujući različite algoritme grožđenja, odnosno vrijednosti varijable AG za svaku vrijednost treba optimizirati vrijednost parametra praga p.

Neke su varijable i binarne. Primjerice, varijabla pretpostavljenih heuristika se zapravo sastoji od dvije binarne podvarijable - primjene heuristike H_1 i primjene heuristike H_2. Naime, te dvije varijable nose vrijednost istinitosti ili neistinitosti.

Kako je broj varijabli, podvarijabli i parametara velik te kako bi kartezijev produkt kombinacije vrijednosti varijabli i parametara kao posljedicu imao vrlo velik broj potrebnih eksperimenata, u pojedinom će se eksperimentu raditi s jednom ili jednim dijelom varijabli, dok će druge biti zaključane na određenoj vrijednosti. Moguća procjena potrebnog broja eksperimenata prema mogućim vrijednostima varijabli prema prethodnoj podjeli bez procjene parametara je ugrubo $4 * 6 * 2 * 2 * 5 * 2 * 3 * 5 * 2 * 3 * 3 * 2 * 2 * 2$ što bi odgovaralo 2.073.600 eksperimenata. Odluku o tome koje će se varijable zajedno istraživati treba temeljiti na prirodi međusobnog utjecaja varijabli, odnosno njihove nezavisnosti. Dokazivanje nezavisnosti varijabli izlazi iz područja interesa ovog doktorskog rada te se iz tog razloga ta odluka temelji na intuitivnom poznavanju prirode problema. Redoslijed, odnosno odluka o zajedničkom, tj. odvojenom eksperimentiranju varijablama donosit će se u tijeku eksperimenta ovisno o dotadašnjim rezultatima.

Redoslijed eksperimentiranja varijablama je određen intuicijom važnosti optimalne vrijednosti varijable na cijeli zadatak. Iz tog se razloga prvo eksperimentira algoritmom grožđenja što je zapravo i najvažniji korak u rješavanju cjelokupnog zadatka.

Već je rečeno kako će varijable nad kojima se trenutno ne eksperimentira biti zaključane na određenim vrijednostima. U slučaju da je s nekom varijablom već eksperimentirano te je pronađena njena optimalna vrijednost, ta će varijabla do daljnjega imati upravo tu vrijednost. Do tada će ona imati vrijednost koja se može pretpostaviti kao optimalna ili koja pojednostavljuje istraživanje.

U nastavku je dan kratak opis svake varijable s pretpostavljenom vrijednosti.

3.2.1 Varijable procesa grožđenja

U nastavku će biti prikazane varijable procesa grožđenja. To su

- varijabla algoritma grožđenja (AG)
- varijabla mjere udaljenosti između vektora (MU)
- varijabla primjene heuristika u procesu grožđenja (PH)
- varijabla mjere težine svojstava u formalnom prikazu dokumenta (MTS)

U nastavku će svaka od ovih varijabli biti pobliže opisana.

Varijabla algoritma grožđenja (AG)

Bit će uspoređena četiri algoritma grožđenja:

- algoritam jednim prolaskom pojedinačnom vezom
- hijerarhijski algoritam potpunom vezom
- hijerarhijski algoritam prosječnom vezom
- hijerarhijski algoritam pojedinačnom vezom

Više o ovim algoritmima je rečeno u poglavlju 2.4.
Ova se varijabla istražuje kao prva iz tri razloga:

- postupak grožđenja je centralni postupak u rješavanju problema pronalaženja događaja te se na njega zato obraća posebna pažnja
- ova varijabla sa sobom nosi dodatni parametar praga koji uvelike uvećava broj potrebnih eksperimenata te se zato taj parametar pokušava ustaliti što je prije moguće

Ta će varijabla, jednom istražena, do kraja istraživanja biti zaključana na svojoj optimalnoj vrijendosti. Za primijetiti je da parametar p neće moći biti zaključan iz razloga što će, primjerice, različite mjere udaljenosti vektora davati vrijednosti u različitim intervalima te se neće moći pretpostaviti da je vrijednost parametra p nezavisna o vrijednosti varijable mjere udaljenosti (MU).

Dok će se ova varijabla istraživati, sve će druge varijable biti zaključane na svojim pretpostavljenim vrijednostima. Te će vrijednosti biti navedene u nastavku uz kratki opis samih varijabli.

Varijabla mjere udaljenosti (MU)

Nominalna varijabla mjere udaljenosti u ovom israživanju ima sljedeće moguće vrijednosti:

- *Manhattan* udaljenost
- Euklidova udaljenost
- kosinusna mjera
- *Jaccard* koeficijent
- *Dice* koeficijent
- *Jensen-Shannon* odstupanje

Ove mjere udaljenosti su detaljnije opisane u poglavlju 2.3, a pretpostavljena vrijednost ove varijable je kosinusna mjera.

Varijabla primjene heuristika (PH)

Dvije su pretpostavljene heuristike u poglavlju 3.1.4:

1. H_1 - podaci o jednom događaju se pretežno objavljuju u istom danu
2. H_2 - na nekom se portalu o istom događaju ne izvještava dva puta

U poglavlju 3.1.4 te su varijable i testirane na podacima, odnosno *in vitro*. U sklopu ovih eksperimenata bit će provedena i provjera na zadatku, odnosno *in vivo*.

Binarne podvarijable varijable PH su PH_1 - varijabla primjene prve heuristike i PH_2 - varijabla primjene druge heuristike.

Pretpostavljena vrijednost varijable PH_1 je pozitivna, dok je pretpostavljena vrijednosti PH_2 negativna. Razlog vrijednosti varijable PH_1 je čisto pragmatičke prirode - omogućuje da se svi dotadašnji eksperimenti provode samo nad uzorkom pojedinog dana što ograničava broj dokumenata, odnosno broj podatkovnih točaka te ne vodi do kombinatorne eksplozije. Naime, broj kombinacija koje algoritam za grožđenje mora usporediti, i to ponekad i više puta, jest $n * (n-1)/2$. Pretpostavljena vrijednost druge varijable je negativna iz dva razloga - prirodnije je ne primjenjivati nedokazanu heuristiku te je usto to i algoritamski jednostavnije.

Varijabla mjere težine svojstava (MTS)

Bit će eksperimentirano s idućim mogućim vrijednostima nominalne varijable mjere težine svojstava (MTS):

- vjerojatnost
- uvjetna vjerojatnost
- pojedinačna međusobna informacija
- TF-IDF mjera
- t-test mjera

Više je o tim mjerama rečeno u poglavlju 2.2.2. Pretpostavljena vrijednost ove varijable je najčešće korištena mjera - TF-IDF.

3.2.2 Varijable određivanja svojstava na razini pojavnica

U nastavku će biti prikazane varijable određivanja svojstava na razini pojavnica. Naime, u najjednostavnijem modelu formalnog prikaza dokumenta, dokument se prikazuje kao vektor pojavnica koje dokument sadrži. Postoji mogućnost intervencije u taj popis tako da se, primjerice,

- isključe interpunkcije
- sačuva informacija o veličini slova ili ona pak zanemari
- pojavnice iz naslova ponove određeni broj puta

Varijable koje istražuju ova pitanja pobliže su prikazane u nastavku.

Varijabla utjecaja interpunkcija (UI)

Prva varijabla određivanja svojstava je binarna koja istražuje utjecaj interpunkcija. Želi se istražiti pospješuje li rješavanje zadatka uključivanje interpunkcija kao svojstava. Pretpostavljena vrijednost ove varijable je negativna.

Varijabla utjecaja veličine slova (UVS)

Druga varijabla određivanja svojstava je ona utjecaja veličine slova. Želi se istražiti način na koji je optimalno prikazati pojavnice kao svojstva - uzevši u obzir veličinu slova ili ne. Postoje tri moguće vrijednosti ove nominalne varijable:

- pojavnice se prikazuju kao u originalu
- pojavnice se prikazuju malim slovima
- pojavnice se prikazuju slovima određenima jendostavnim statističkim modelom

Pretpostavljena vrijednost ove varijable je da se pojavnice prikazuju malim slovima.

Varijabla važnosti naslova (VN)

Želi se istražiti koliko su važne pojavnice iz naslova. Primjenjuje se jednostavna metoda - svaka pojavnica iz naslova se ponavlja u formalnom zapisu n puta.

Pretpostavljena vrijednost ove diskretne varijable je 3.

3.2.3 Varijable odabira svojstava na razini pojavnica

Jednom kada su određena svojstva kojima se dokument želi prikazati, moguće je u taj popis intervenirati odabirom samo onih svojstava koja zadovoljavaju neki uvjet. Tako je moguće

- isključiti svojstva koja se pojavljuju samo jednom
- isključiti svojstva koja ne posjeduju statističku razlikovnost

Utjecaj takvih postupaka istražuju varijable prikazane u nastavku.

Varijabla isključivanja hapax legomena (IHL)

Ovom binarnom varijablom se želi istražiti utjecaj isključivanja hapax legomena iz popisa svojstava. Pretpostavljena vrijednost ove varijable je negativna.

Varijabla isključivanja funkcijskih riječi (IFR)

Ovom se varijablom iz popisa svojstava isključuju pojavnice koje nemaju dostatnu razlikovnu vrijednost u razlikovanju jednog dokumenta od drugih. Osnovna mjera koja će se koristiti za razinu razlikovnosti je IDF. Ta se varijabla ustvari sastoji od dvije ordinalne podvarijable prikazujući rang u popisu pojavnica prema pripadajućem IDF-u koje se ne uključuju u popis svojstava. Bit će eksperimentirano s raznim rangovima.

Pretpostavljena vrijednost obje varijable je 0, odnosno *a priori* se funkcijske riječi ne isključuju.

3.2.4 Ostale varijable određivanja i odabira svojstava

Ostale metode određivanja i odabira svojstava koje kao svojstva ne uključuju samo pojavnice su sljedeće:

- korjenovanje
- morfosintaktičko označavanje
- prepoznavanje osobnih imena
- određivanje statistički značajnih digrama

U nastavku će biti prikazane varijable koje istražuju ove mogućnosti.

Varijabla korjenovanja (K)

Nominalnom varijablom korjenovanja se istražuje utjecaj postupka korjenovanja pojavnica na konačni rezultat. Naime, morfološkom unifikacijom postiže se mogućnost da se sva pojavljivanja nekog leksema bez obzira na oblik smatraju istim svojstvom. Svaka unifikacija, naravno, sa sobom nosi i određenu pogrešku. Usto, specifičnost oblika pojavnice isto je oblik informacije koja se ovom unifikacijom gubi. Tri moguće vrijednosti ove nominalne varijable su

- korjenovanje se ne provodi
- provodi se jednostavno korjenovanje nezavisno o jeziku
- provodi se kompleksnije korjenovanje prilagođeno hrvatskom jeziku

Pretpostavljena vrijednost ove varijable jest da se ne provodi nikakvo korjenovanje.

Varijabla morfosintaktičkog označavanja (MO)

Binarna varijabla morfosintaktičkog označavanja istražuje utjecaj prethodnog statističkog morfosintaktičkog označavanja uzorka - naprednije metode morfološke normalizacije od prethodno opisanog korjenovanja. Pri morfosintaktičkom označavanje se, naime, u pravilu određuje točna morfosintaktička kategorija pojavnice kao i njoj pripadajuća lema. To se, naravno, za razliku od korjenovanja, čini pomoću konteksta u kojemu se pojavnica pojavljuje. Usto se koristi i morfološki leksikon koji nudi početni popis mogućih morfoloških tumačenja pojavnice. Statističkim modelom koji je prethodno istreniran na označenom uzorku se vrši razrješavanje morfosintaktičke višeznačnosti.

Morfosintaktičko označavanje korišteno u ovom doktorskom radu plod je rada Zavoda za lingvistiku Filozofskog fakulteta u Zagrebu [Agić and Tadić, 2006]. Pretpostavljena vrijednost ove varijable je negativna.

Varijabla prepoznavanja osobnih imena (POI)

Također binarna varijabla prepoznavanja osobnih imena istražuje utjecaj prethodnog prepoznavanja imena tvrtki i osoba na zadatak pronalaženja događaja. Zadatak prepoznavanja osobnih imena riješen je od strane Zavoda za poslovna istraživanja od kojega je uzorak i dobiven.

Pretpostavljena vrijednost ove varijable je negativna.

Varijabla statistički značajnih digrama (SZD)

Varijabla statistički značajnih digrama istražuje utjecaj višečlanih izraza na dani zadatak. Postoje dva najčešća načina rada s izrazima iznad razine riječi:

- statistički - pronalaze se pojavnice čije supojavljivanje nije statistički slučajno
- lingvistički - nakon morfosintaktičke analize teksta vrši se sintaktička te se pronalaze imenične i glagolske sveze razdjeljivanjem (engl. *chun-*

king), odnosno te se fraze spajaju u sintaktičku rečeničnu strukturu (engl. *sentence parsing*).

Kako su jezični alati za hrvatski jezik na sintaktičkoj razini trenutno tek u nastanku, u ovom se doktorskom radu eksperimentira sa statističkim metodama. Jedna od češćih statističkih metoda za pronalaženje neslučajnog supojavljivanja pojavnica, odnosno kolokacija jest statistički test hi-kvadrat. Taj test uspoređuje očekivanu i promatranu čestotu pojave određenog n-grama te na temelju njega određuje koliko je njihovo supojavljivanje slučajno. Osnovni problem hi-kvadrata je taj što on precjenjuje neslučajnost supojavljivanja pojavnica koje se rijetko pojavljuju. Iz tog se razloga u ovom doktorskom radu računa samo hi-kvadrat digrama čija se oba konstituenta pojavljuju minimalno tri puta. Osnovna formula hi-kvadrata je sljedeća:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{3.1}$$

U slučaju računanja hi-kvadrata na digramima iz nekog korpusa za svaki se digram stvara tablica slučajeva (engl. *contingency table*) te se iz te tablice iznos hi-kvadrata računa sljedećom formulom [Manning and Schütze, 1999e]:

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})} \tag{3.2}$$

Varijabla statistički značajnih digrama se, dakle, istražuje određivanjem dvočlanih svojstava statističkom metodom hi-kvadrata. Eksperimentira se s određivanjem različitog broja dvočlanih izraza, odnosno vrijednosti ordinalne varijable.

Pretpostavljena vrijednost ove ordinalne varijable je 0.

3.2.5 Varijabla utjecaja veličine referentnog korpusa na mjeru težine svojstva

Varijablom utjecaja veličine referentnog korpusa (VRK) se želi istražiti ujecaj količine podataka na kojima se mjeri čestota, odnosno neka druga numerička vrijednost vezana uz određeno svojstvo u referentnom korpusu kako bi se ona usporedila s čestotom, odnosno drugom numeričkom vrijednosti tog svojstva u određenom dokumentu te time izračunala mjera težine tog svojstva (MTS). Mjere težine svojstava koja koriste referentni korpus su

- uvjetna vjerojatnost
- pojedinačna međusobna informacija
- TF-IDF mjera
- t-test mjera

Jedina mjera koja ne koristi referentni korpus je vjerojatnost.

Pretpostavljena vrijednost ove kontinuirane varijable je beskonačna, odnosno koristi se cijeli referentni korpus.

3.3 Evaluacijske mjere korištene u istraživanju

Evaluacijske mjere korištene u ovom istraživanju podrobnije opisane u poglavlju 2.5 su sljedeće:

- čistoća
- normalizirana međusobna informacija
- rand indeks
- preciznost
- potpunost

- F_1 mjera
- $F_{0.5}$ mjera

Odluka o korištenim mjerama se oslanja na čestotu primjene mjera u evaluaciji algoritama za grožđenje, odnosno algoritama za obradu prirodnog jezika.

3.4 Gornja i donja granica istraživanja

Gornja i donja granica istraživanja služe tome da postave moguće rezultate istraživanja u neki okvir - maksimalni mogući rezultat te onaj minimalni. U poglavlju 2.5 je rečeno da se za gornju granicu ili strop često koriste pogreška u zlatnom standardu, odnosno dogovor između označitelja. U ovom se istraživanju koristi dogovor između označitelja.

Kao donju granicu nema smisla koristiti ni slučajno rješenje ni ono najfrekventnije iz razloga što bi oba pristupa zbog velikog permutacijskog potencijala zadatka dale vrijednosti bliske nuli. Kako ovaj problem nije dovoljno istražen te u industriji ne postoji neki standard koji bi se mogao koristiti kao donja granica, donja granica u ovom istraživanju neće biti postavljena.

Dogovor između označitelja (engl. *inter-annotator agreement*) je mjera koja pokazuje u koliko se posto slučajeva dvoje ili više ljudskih označitelja slaže u pridruženim oznakama uzorku koji se ručno označava.

Ta mjera pruža uvid u kompleksnost nekog problema, odnosno mogućnost rješavanja tog problema od strane čovjeka.

Za potrebe računanja te mjere potrebno je imati više označenih uzoraka. Iz tog je razloga uz prvi uzorak - zlatni standard koji je označavan od strane dvoje studenata - još jedan student dobio zadatak označiti isti uzorak uz iste upute. Prvi se uzorak zove Mislav, a drugi Nikola prema imenima studenata što su označavali uzorke. Neki usporedni podaci o tim uzorcima prikazani su u tablici 3.10.

Kako je označavanjem moguće i uklanjati dokumente koji ne odgovaraju kriteriju čistoće dokumenata, za očekivati je različit broj dokumenata u jednom, odnosno drugom uzorku. To je potvrđeno podacima prikazanima

Tablica 3.10: Usporedba broja dokumenata i grozdova u uzorcima za računanje dogovora između označitelja

uzorak	broj dokumenata	broj grozdova
Mislav	2,402	1,217
Nikola	2,414	969

Tablica 3.11: Usporedba broja dokumenata i grozdova u ujednačenim uzorcima za računanje dogovora između označitelja

uzorak	broj dokumenata	broj grozdova	prosječna veličina ($n > 1$)
Mislav	2,398	1,214	3.982
Nikola	2,398	955	4.997

u tablici 3.10. Kako bi se uzorci ujednačili, iz oba su izbačeni dokumenti koji nisu sadržani u oba uzorka. Usporedni podaci o ujednačenim uzorcima prikazani su u tablici 3.11.

Iz podataka je vidljivo kako se uzorak Nikola sastoji od znatno manjeg broja grozdova, odnosno da su, očito, pri oblikovanju tog uzorka primijenjeni liberalniji kriteriji pridruživanja dokumenata u isti grozd. Isto je vidljivo iz podatka prosječne veličine višečlanog grozda.

Kako bi se ova dva uzorka mogla usporediti, oni su prikazani na isti način kao i pri evaluaciji, naime prikazani su u obliku skupa uređenih parova koji se sastoje od dva dokumenta iz istog grozda. Dokument sam sa sobom, naravno, ne ostvaruje vezu, odnosno uređeni parovi redovito sadrže dvije različite vrijednosti.

Kako se na isti se način u istraživanju računaju evaluacijske mjere preciznosti i potpunosti, ovakav način prikaza dvaju uzoraka za računanje dogovora između označitelja je dodatno opravdan.

Pretvaranjem uzoraka u skup veza, broj veza u jednom, odnosno drugom uzorku prikazan je u tablici 3.12. Razlike u broju veza nisu iznenađujuće imajući na umu manji broj grozdova u uzorku Nikola.

Stvoreni uzorci mogu se sada jednostavno usporediti - promatra se pos-

Tablica 3.12: Broj veza u uzorcima nakon pretvaranja uzorka u skup veza

uzorak	broj veza
Mislav	5,100
Nikola	8,458

Tablica 3.13: Rezultati mjera dogovora između označitelja DIO_1 i DIO_2

mjera	rezultat
DIO_1	0.684
DIO_2	0.91

totak identičnih, odnosno neidentičnih veza.

U ovom su radu korištene dvije mjere dogovora između označitelja - DIO_1 i DIO_2 te uzevši u obzir dva skupa veza U_1 i U_2 izgledaju sljedeće:

$$DIO_1 = \frac{2 * |U_1 \cap U_2|}{|U_1| + |U_2|} \tag{3.3}$$

$$DIO_2 = \frac{|U_1 \cap U_2|}{min(|U_1|, |U_2|)} \tag{3.4}$$

Valja primijetiti kako je mjera DIO_1 identična κ koeficijentu - najčešćoj metodi računanja dogovora između označitelja. Rezultati tih mjera su dani u tablici 3.13. Iz njih je vidljivo kako je mjera DIO_1 daje bitno niži rezultat od mjere DIO_2.

Mjera DIO_1 generalno računa objektivniju sličnost dvaju uzoraka. Naime, ona računa stvarni presjek svih članova oba skupa. Druga mjera, DIO_2, pak dijeli duljinu presjeka s duljinom manjeg uzorka, odnosno računa presjek unije skupova. Prva mjera ne daje objektivan rezultat u slučaju velike razlike u broju elemenata prvog, odnosno drugog skupa. To jest slučaj u ovom istraživanju, naime iz podataka u tablici 3.12 vidljivo je kako se u uzorku

Nikola nalazi 65% više veza nego u uzorku Mislav. Iz tog se razloga u ovom skupu podataka objektivnijom mjerom može smatrati DIO_2. Unatoč tome, mjera DIO_1 ukazuje na kompleksnost samog zadatka.

Za gornju granicu, odnosno strop ovog istraživanja, dakle, može se odabrati vrijednost 0.91. Ne smije se zanemariti druga vrijednost - 0.684 - koja ukazuje na punu kompleksnost ovog zadatka i za čovjeka. Donja granica u ovom istraživanju ne postoji.

Poglavlje 4

Rezultati istraživanja

U ovom poglavlju prikazani su rezultati istraživanja. Istražuje se niz varijabli prestavljenih u prošlom poglavlju, i to sljedećim redoslijedom:

- algoritam grožđenja (AG) - bit će uspoređena četiri algoritma grožđenja, tri hijerarhijska i jedan algoritam jednim prolaskom
- mjera udaljenosti (MU) - bit će uspoređeno šest različitih mjera udaljenosti
- pretpostavljene heuristike (PH) - obje će u poglavlju 3.1.4 opisane heuristike biti ispitane na danom zadatku
- mjera težine svojstava (MTS) - bit će ispitano pet mjera težine svojstava
- metode određivanja i odabira svojstava
  - utjecaj interpunkcija (UI) - bit će provjereno ima li smisla uključivati interpunkcije u formalni zapis ili ne
  - utjecaj veličina slova (UVS) - odlučit će se uzimati li u obzir veličinu slova, te ako da, kako
  - važnost naslova (VN) - bit će odlučeno kako vrednovati naslov s obzirom na tekst

- isključivanje hapax legomena (IHL) - bit će ispitano kako na rezultat utječe isključivanje svojstava koja se pojavljuju samo jednom
- isključivanje funkcijskih riječi (IFR) - bit će ispitano kako na rezulat utječe isključivanje svojstava čija je diskriminativna vrijednost najmanja
- korjenovanje (K) - ispitat će se pomaže li zadatku jezično nezavisno, odnosno jezično zavisno korjenovanje
- morfosintaktičko označavanje (MO) - provjerit će se pomaže li uključivanje statistički određenih morfosintaktičkih kategorija i lema u popis svojstava
- utjecaj veličine korpusa na IDF mjeru (VIDF) - bit će istraženo koliko je podataka potrebno za dobru procjenu IDF mjere svojstava s obzirom na zadatak
- prepoznavanje osobnih imena (POI) - bit će ispitano pomaže li prethodno prepoznavanje imena tvrtke ili osobe te njihovo uvrštavanje u svojstva
- statistički značajni digrami (SZD) - provjerit će se pomaže li uvrštavanje statistički značajnih digrama u svojstva

Vrijednosti svih varijabli bit će odabrane metodom maksimizacije određenih evaluacijskih mjera. Evaulacijske mjere koje će također biti kritički sagledane su sljedeće:

- čistoća
- normalizirana međusobna informacija
- rand indeks
- preciznost
- potpunost
- F_1 mjera
- $F_{0.5}$ mjera

4.1 Odabir algoritma grožđenja

Prva istraživana varijabla u nizu je nominalna varijabla algoritama za grožđenje (AG) koja ima sljedeće moguće vrijednosti:

- algoritam jednim prolaskom pojedinačnom vezom - JP
- hijerarhijski algoritam potpunom vezom - H-POTP
- hijerarhijski algoritam prosječnom vezom - H-PROS
- hijerarhijski algoritam pojedinačnom vezom - H-POJ

Za svaku moguću vrijednost varijable, odnosno u svakom algoritmu postoji slobodni parametar praga p koji je potrebno optimizirati. U prvom eksperimentu vrijednosti praga su između 0.05 i 1.0 s korakom od 0.05 što dovodi do $4 * 20 = 80$ eksperimenata.

Kao što je navedeno u nacrtu istraživanja, tijekom potrage za optimalnom vrijednosti jedne varijable, ostale su varijable fiksirane na pretpostavljenoj, odnosno već optimiziranoj vrijednosti. Tako se varijabla AG istražuje dok se kao mjera udaljenosti koristi kosinusna mjera, prva se heuristika primjenjuje, druga se ne primjenjuje, kao mjera težine svojstva se koristi TF-IDF, interpunkcije ne ulaze u popis svojstava, sve se pojavnice prikazuju malim slovima, čestota svojstava iz naslova se množi s tri, hapax legomena i funkcijske riječi su uključene u svojstva, ne vrši se nikakva morfološka normalizacija, za izračun IDF mjere se koristi cijeli korpus, rezultati prepoznavanja osobnih imena se ne koriste te se u svojstva ne uključuju statistički značajni digrami.

Rečeno je da će u eksperimentiranju s prve dvije varijable vrijednost varijable prve heuristike biti pozitivna što odskače od svih drugih pretpostavljenih vrijednosti. Takva je odluka donesena zato što ta heuristika čini problem bitno lakše izračunljivim. Naime, primjenom prve heuristike omogućeno je pronalaženje događaja vršiti nad uzorkom dokumenata objavljenih unutar jednog dana.

Prvi je eksperiment proveden na uzorku 5-SVI, dakle nad dokumentima objavljenima 5. svibnja. Opći rezultati usporedbe ova četiri algoritma dani su u tablici 4.1. Iz podataka u tablici je vidljivo da algoritam grožđenja

Tablica 4.1: Odnos algoritama grožđenja s obzirom na maksimalnu mjeru $F_{0.5}$ ($F_{0.5}$), parametar praga (p) i vrijeme izvršavanja (t)

algoritam	$F_{0.5}$	p	t
JP	0.787	0.65	6.761
H-POTP	0.76	0.8	2995.641
H-PROSJ	0.779	0.7	2846.576
H-POJ	0.787	0.65	2814.0

jednog prolaska pojedinačnom vezom, koji je bitno jednostavniji od hijerarhijskih algoritama, u uspješnosti prati hijerarhijske algoritme te ih čak lagano i nadmašuje.

Moguće objašenjenje ovakvog rezultata leži u tome da je prednost hijerarhijskih algoritama ta što oni u svojoj kompleksnosti hvataju malene razlike između odnosa udaljenosti točaka, odnosno prilagođavaju se do sada donesenim odlukama o pripadnosti grozdu. Za razliku od toga u ovom problemu se svaki dokument prikazuje u visokodimenzionalnom prostoru jakom apstrakcijom od originalnog dokumenta te male razlike u međusobnim udaljenostima vjerojatno nisu bitne. Isto tako, hijerarhijski algoritmi pamte redoslijed spajanja grozdova što za ovaj zadatak nije bitno.

U ovom su istraživanju mjerene sve predložene mjere uspješnosti. U tablici 4.2 prikazane su vrijednosti svih evaluacijskih mjera za p=0.0 do p=1.0 s korakom od 0.05. Prikazani podaci služe istraživanju i usporedbi svojstava svake pojedine evaluacijske mjere na zadatku pronalaženja događaja.

Čistoća je maksimalna kad se u svakom grozdu nalaze samo dokumenti iste klase. Negativna strana ove mjere je ta što će, u slučaju da je svaki dokument u svom grozdu, čistoća biti maksimalna. To je i vidljivo u tablici 4.2, naime, kad je prag 0.0, samo se dokumenti s identičnim vektorima spajaju u grozdove te grožđenje prestaje. Naravno, ovakav se rezultat ne može shvatiti optimalnim te se upravo iz tog razloga čistoća redovito kombinira s nekom mjerom koja kažnjava preveliki broj grozdova. Čistoća je, ipak, i dalje relativno korisna mjera. U tablici 4.2 možemo primijetiti kako ona povećanjem praga postepeno pada, a na vrijednosti praga od ~ 0.7 počinje jače opada-

Tablica 4.2: Odnos parametra p, evaluacijskih mjera (C - čistoća, NMI - normalizirana međusobna informacija, RI - rand indeks, PR - preciznost, POT - potpunost, F_1, $F_{0.5}$) i vremena (t) pri korištenju algoritma JP

p	C	NMI	RI	PR	POT	F_1	$F_{0.5}$	t
0.0	**1.0**	0.93	0.997	**1.0**	0.0	0.0	0.0	7.059
0.05	0.998	0.935	0.997	0.969	0.046	0.089	0.195	7.03
0.1	0.996	0.938	0.997	0.949	0.084	0.154	0.31	6.819
0.15	0.992	0.941	0.997	0.913	0.126	0.221	0.406	6.786
0.2	0.986	0.945	0.998	0.892	0.174	0.291	0.488	6.797
0.25	0.981	0.949	0.998	0.897	0.229	0.365	0.567	6.797
0.3	0.979	0.951	0.998	0.892	0.254	0.395	0.594	6.783
0.35	0.978	0.954	0.998	0.897	0.294	0.443	0.637	6.763
0.4	0.977	0.956	0.998	0.899	0.327	0.479	0.666	6.753
0.45	0.976	0.96	0.998	0.899	0.389	0.543	0.712	6.743
0.5	0.974	0.964	0.998	0.9	0.443	0.594	0.746	6.762
0.55	0.968	0.969	0.998	0.86	0.54	0.664	0.769	6.79
0.6	0.957	0.972	**0.999**	0.828	0.64	0.722	0.782	6.757
0.65	0.944	0.975	**0.999**	0.804	0.724	0.762	**0.787**	6.761
0.7	0.927	**0.976**	**0.999**	0.746	0.804	**0.774**	0.757	6.745
0.75	0.881	0.969	0.997	0.535	0.94	0.682	0.585	6.774
0.8	0.84	0.961	0.996	0.417	0.985	0.586	0.471	6.747
0.85	0.733	0.923	0.984	0.15	0.992	0.261	0.181	6.769
0.9	0.52	0.802	0.901	0.028	0.996	0.054	0.035	6.795
0.95	0.122	0.252	0.221	0.004	1.0	0.007	0.005	6.891
1.0	0.017	0.0	0.003	0.003	**1.0**	0.006	0.004	7.039

ti. Čistoća nam za ovaj primjer govori da je poželjna optimalna vrijednosti parametra p iznad 0.7.

Normalizirana međusobna informacija je mjera međusobne informacije normalizirane entropijom između zlatnog standarda i rezulatata eksperimenta, odnosno NMI kažnjava veći broj grozdova. Tako je sama MI za $p = 0.0$ jednaka 5.974, te opada s rastom parametra p te na $p \sim 0.7$ počinje naglije opadanje. Na $p = 1.0$ MI iznosi 0.0. Sama međusobna informacija naslućuje, kao i čistoća, da je optimalni rezultat negdje ispod 0.7. Entropija zlatnog standarda je konstantnih 5.974, dok se entropija rezultata kreće od 6.874 za $p = 0.0$ te naglije počinje padati oko $p \sim 0.8$ kako bi na $p = 1.0$ bila na 0.0. Normalizacijom MI-a s entropijom rezultata i zlatnog standarda dobiva se mjera koja svoj maksimum postiže na $p = 0.7$ te se na $p \sim 0.8$ brže počinje obrušavati prema 0.0 za $p = 1.0$. Normalizirana međusobna informacija, dakle, postiže plato u području između 0.65 i 0.7.

Rand indeks je identičan mjeri točnosti iz područja obrade prirodnog jezika, odnosno na temelju tablice mogućnosti računa postotak odluka koje su točne. U grožđenju se tablica mogućnosti stvara na način da se sve veze između dva dokumenta iz rezultata proglase istinito pozitivnima, istinito negativnima, lažno pozitivnima i lažno negativnima. Ista se tablica kasnije koristi za preciznost, potpunost te F mjere. Kod RI mjere je odmah moguće primijetiti njene izrazito visoke vrijednosti koje započinju na 0.997 te se tek na $p = 0.8$ s vrijednosti 0.996 počinju obrušavati prema 0. Razlog tomu je visok broj istinito negativnih parova. Taj je broj toliko velik zato što istinito negativne parove čine veze između svih dokumenata koji nisu u istom grozdu. Takvih veza u slučaju većeg broja manjih grozdova ima izrazito puno, odnosno redovito više nego drugih veza. Unatoč visokim vrijednostima, moguće je primijetiti plato RI-a u području između 0.6 i 0.7.

Preciznosti i potpunost se klasično suprostavljaju: preciznost je 1.0 kad je svaki dokument u svom grozdu, odnosno kad je $p = 0.0$ te pada prema nuli kako je sve više dokumenata u istom grozdu, dok se potpunost ponaša suprotno - iznosi 0.0 kad je svaki dokument u svome grozdu te raste na 1.0 kako se dokumenti spajaju u sve manji broj grozdova. Te dvije mjere se redovito ne koriste zasebno, već se promatra njihova težinska harmonijska

sredina poznata kao F_β gdje je β parametar koji važe između preciznosti i potpunosti. Za $\beta = 1$ preciznost i potpunost se smatraju jednako vrijednima, dok se za manji β prednost daje preciznosti, odnosno za veći potpunosti. U tablici 4.2 mjerene su F_1 i $F_{0.5}$. Mjerena je $F_{0.5}$ zbog prirode zadatka. Naime, pri prikazu rezultata pronalaženja događaja neprikazivanje nekog djelomično sličnog dokumenta u grupi dokumenata se može smatrati manjom pogreškom od prikazivanja dokumenta u grupi gdje mu nije mjesto. Informacijska će potreba i u prvom slučaju visoko vjerojatno biti zadovoljena, a rezultati neće biti onečišćeni zalutalim dokumentima. F_1 mjera svoj maksimum postiže na $p = 0.7$, dok $F_{0.5}$ mjera maksimum postiže na $p = 0.65$.

Vrijeme je u algoritmima jednim prolaskom konstantno za razliku od hijerarhijskih gdje vrijeme raste s parametrom p zbog toga što je potrebno više iteracija do trenutka zaustavljanja algoritma, što se zbog kompleksnosti algoritma primjećuje na izmjerenoj vrijednosti.

Iz usporedbe svih dobivenih vrijednosti moguće je zaključiti da je poželjna vrijednost parametra p za algoritam grožđenja jednim prolaskom između 0.6 i 0.7 s prilično jakom tendencijom prema 0.65.

U sklopu ovog skupa eksperimenata želja je provjeriti i koliko je parametar p stabilan, odnosno kolikog su opsega potrebni eksperimenti da bi se on uspješno procijenio. Naime, u istraživanju sljedećih varijabli često će biti potrebno ponovno pronalaženje parametra p. Iz tog će se razloga u sljedećem eksperimentu s parametrom p eksperimentirati na tri uzorka - 4-SVI, 5-SVI i 6-SVI, odnosno dokumentima objavljenima 4., 5. i 6. svibnja. Prva grupa ima bitno manji broj entiteta iz razloga što je 4. svibnja 2008. bila nedjelja, dok je treći uzorak sličan drugome, tj. onome nad kojime je do sada eksperimentirano. Kako se u prethodnim eksperimentima $F_{0.5}$ pokazao kao najinformativnija mjera, u ovoj će grupi eksperimenata biti promatrana stabilnost te mjere s obzirom na parametar p u sva tri uzorka. Parametar p se ispituje između vrijednosti 0.5 i 0.8 s korakom od 0.02. Rezultate te grupe eksperimenata moguće je vidjeti u tablici 4.3. Iz tablice je vidljivo da se maksimalni $F_{0.5}$ postiže oko 0.65 - prethodno pretpostavljene optimalne vrijednosti parametra p.

Tablica 4.3: Odnos $F_{0.5}$ mjere za uzorke 4-SVI, 5-SVI i 6-SVI s obzirom na parametar p

p	4-SVI	5-SVI	6-SVI
0.5	0.728	0.746	0.699
0.52	0.741	0.755	0.71
0.54	0.743	0.76	0.706
0.56	0.776	0.775	0.712
0.58	0.778	0.781	0.709
0.6	0.765	0.782	0.699
0.62	0.745	0.789	**0.71**
0.64	0.748	**0.791**	0.689
0.66	0.755	0.754	0.654
0.68	**0.756**	0.747	0.61
0.7	0.75	0.757	0.552
0.72	0.763	0.745	0.507
0.74	0.733	0.586	0.482
0.76	0.705	0.562	0.449
0.78	0.667	0.489	0.405
0.8	0.664	0.471	0.236

Kako bi se istražila stabilnost vrijednosti parametra p, izračunata je korelacija varijabli mjere $F_{0.5}$ nad opisana tri uzorka. Korelacije iznose 0.903, 0.909 te 0.791 što njihovu aritmetičku sredinu postavlja na 0.868, odnosno u jaku korelaciju. Najslabiji koeficijent korelacije je na samoj granici jake korelacije.

Iz ovog se skupa eksperimenata može zaključiti kako hijerarhijski algoritmi s kvadratnom vremenskom kompleksnosti ne postižu bolje rezultate od bitno jednostavnijeg algoritma jednim prolaskom s linearnom vremenskom kompleksnosti. Mogući razlog tomu je velika razina apstrakcije prikaza dokumenata vektorima te visoka dimenzionalnost tih vektora, kao i činjenica da prednosti hijerarhijskih algoritama - uzimanje u obzir prethodnih odluka u procesu grožđenja kao i njihovo bilježenje - zbog prethodno navedenih razloga nisu važne. Nadalje, istražujući optimalnu vrijednst parametra p za algoritam jednog prolaska zaključeno je kako je parametar stabilan u različitim uzorcima dosežući maksimalnu $F_{0.5}$ mjeru na sličnim vrijednostima i pokazujući jaku korelaciju vrijednosti $F_{0.5}$ mjere među uzorcima. Od mjera evaluacije sve su se pokazale primjenjivima za izvođenje zaključaka, a kao najinformativnija mjera s obzirom na prirodu zadatka se pokazao $F_{0.5}$.

4.2 Odabir mjere udaljenosti

Zahvaljujući zaključcima iz prethodne točke, s varijablom mjere udaljenosti (MU) će se eksperimentirati koristeći algoritam grožđenja jednim prolaskom i maksimizirajući pritom $F_{0.5}$ mjeru.

U ovoj se grupi eksperimenata istražuje optimalna mjera za računanje udaljenosti vektora, odnosno dokumenata uz optimalni parametar p. Sljedeće se mjere udaljenosti istražuju:

- *Manhattan* udaljenost
- Euklidova udaljenost
- kosinusna mjera
- *Jaccard* koeficijent

Tablica 4.4: Vrijednost $F_{0.5}$ s obzirom na vrijednost varijable mjere udaljenosti (MU) s optimalnim parametrom p i vremenom t na uzorku 5-SVI

MU	$F_{0.5}$	p	t
Manhattan	0.701	4.6	310.415
Euklidova	0.433	0.24	327.8
kosinus	0.791	0.64	143.073
Jaccard	0.793	0.85	668.647
Dice	0.801	0.75	607.364
Jensen-Shannon	0.803	0.95	550.742

- *Dice* koeficijent
- *Jensen-Shannon* odstupanje

Kako će za različite mjere udaljenosti za iste vektore rezultati biti drugačiji, bit će potrebno ponovno odrediti vrijednost parametra p. Eksperiment se provodi na uzorku 5-SVI za svaku mjeru udaljenosti s parametrima p od 0.0 do maksimalne udaljenosti u dvadesetijednom koraku. U slučaju da je maksimalna udaljenost među dokumentima stršeći podatak, eksperiment se ponavlja na manjem rasponu kako bi se dobila vrijednost $F_{0.5}$ bliska onoj maksimalnoj. U tablici 4.4 dani su prvi rezultati. Iz njih je vidljivo da je Euklidova udaljenost gubitnik ovih eksperimenata što se slaže s činjenicom da je glavni problem te mjere izrazita neosjetljivost na ekstremne vrijednosti. Euklidova je udaljenost, naime, stvarna udaljenost između vrhova dvaju vektora što je čini izrazito ranjivom u slučaju da neke dimenzije imaju bitno veće vrijednosti od drugih. U formalnom prikazu dokumenata uvijek postoje razlike u dimenzijama prvenstveno iz razloga različite duljine dokumenta koji se prikazuje. Problem stršećih podataka se rješava normalizacijom vektora čime Euklidova udaljenost postaje identična kosinusnoj mjeri. *Manhattan* udaljenost po svojoj prirodi nije toliko osjetljiva na stršeće podatke te daje bolje rezultate od Euklidove udaljenosti, no i dalje bitno slabije od preostale tri mjere. Kosinus, *Jaccard* i *Dice* te *Jensen-Shannon* odstupanje u ovom eksperimentu imaju vrlo sličan $F_{0.5}$. Prednost se unaprijed može dati kosinusnoj mjeri zbog bitno manjeg vremena izračuna od ostalih mjera.

Tablica 4.5: Vrijednost evaluacijskih mjera NMI, RI, F_1 i $F_{0.5}$ s obzirom na varijablu mjere udaljenosti (MU) s optimalnim parametrom p na uzorku 4-SVI

MU	NMI	RI	F_1	$F_{0.5}$
kosinus	0.977	0.997	0.8	0.765
Jaccard	0.974	0.997	0.768	0.721
Dice	0.971	0.997	0.742	0.732
Jensen-Shannon	0.974	0.997	0.779	0.761

Kako bi se dodatno istražile moguće razlike u uspješnosti ovih mjera, provedena je dodatna skupina eksperimenata nad manjim uzorkom 4-SVI pri kojima su bilježene evaluacijske mjere NMI, RI, F_1 i $F_{0.5}$. Rezultati te skupine eksperimenata moguće je vidjeti u tablici 4.5. Iz ovih podataka je vidljivo da je kosinusna mjera redovito lagani pobjednik nad preostale tri mjere udaljenosti. Iz tog razloga od početnih 6 mjera udaljenosti odabire se kosinusna mjera.

Posljednje pitanje vezano uz odabir mjere udaljenosti jest vrijednost parametra p pri kojemu se za kosinusnu mjeru udaljenosti te algoritam grožđenja jednog prolaska postiže optimalni rezultat. Za izvođenje tog zaključka pomoći će rezultati eksperimenata provedeni u prethodnoj točki prikazani u tablici 4.3 kojom je dokazana stabilnost parametra kroz jaku korelaciju mjera $F_{0.5}$ s obzirom na parametar p. Mjera $F_{0.5}$ je korištena iz razloga što se pokazala najinformativnijom s obzirom na prirodu problema koji se rješava ovim postupkom. Iz tablice je vidljivo da se za sva tri uzorka maksimalni $F_{0.5}$ postiže oko 0.65 pa je iz tog razloga ta vrijednost odabrana kao optimalna vrijednost parametra p.

U opisanom je skupu eksperimenata zaključeno kako je za ovaj zadatak kosinusna mjera najbolja mjera udaljenosti dokumenata. *Jaccard* i *Dice* koeficijenti te *Jensen-Shannon* odstupanje su vrlo blizu toj mjeri, no dosljedno pokazuju nešto slabije rezultate. *Manhattan* udaljenost pokazuje bitno slabije rezultate, dok je gubitnik ovih eksperimenata Euklidova udaljenost koja se još jednom potvrdila kao loša mjera u slučaju postojanja stršećih vrijednosti, odnosno velikih razlika u vrijednostima pojedinih dimenzija.

Kao optimalna vrijednost parametra p odabrana je vrijednost 0.65.

Do sada je, dakle, u potpunosti istražen proces grožđenja, odnosno varijable algoritma grožđenja (AG) i mjere udaljenosti (MU) te je odabrana kosinusna mjera udaljenosti i algoritam grožđenja jednim prolaskom s vrijednosti parametra p jednakoj 0.65.

4.3 Testiranje heuristika na zadatku pronalaženja događaja

U ovom se skupu eksperimenata istražuje varijabla pretpostavljenih heuristika (PH), odnosno *in vivo* utjecaj primjene dviju prethodno pretpostavljenih heuristika na uspješnost obavljanja zadatka pronalaženja događaja. Do sada su optimalne vrijednosti odabrane za sljedeće varijable:

- algoritam grožđenja (AG) je algoritam jednim prolaskom s parametrom p=0.65
- mjera udaljenosti (MU) je kosinusna mjera

Ostale varijable imaju svoje inicijalne vrijednosti definirane na početku ovog poglavlja.

Dvije pretpostavljene heuristike su sljedeće:

1. H_1 - podaci o jednom događaju se pretežno objavljuju u istom danu
2. H_2 - na nekom se portalu o istom događaju ne izvještava dva puta

Podvarijable koje se istražuju su

1. PH_1 - binarna varijabla primjene prve heuristike
2. PH_2 - binarna varijabla primjene druge heuristike

U dosadašnjem istraživanju ovih dviju heuristika na označenom uzorku, *in vitro*, došlo se do sljedećih zaključaka:

- prva heuristika na uzorku dokumenata objavljenih u jednom danu i grozdova čija je većina članaka objavljena tog dana vrijedi u 83.7 posto slučajeva, dakle novosti koje izvještavaju o nekom događaju pokazuju tendenciju da budu objavljene unutar istog dana

- druga heuristika na uzorku svih dokumenata vrijedi u slučaju 85.8% članaka

Obje su se heuristike pokazale vrijednima daljnjeg istraživanja te će u ovom skupu eksperimenata biti istražen utjecaj njihove primjene na zadani problem. U dosadašnjim eksperimentima prva je heuristika bila primjenjivana dok druga nije. Naime, prva heuristika omogućuje da se eksperimentira samo na jednom od tri uzorka te se time i sami eksperimenti pojednostavljuju. Broj rezultata mjere udaljenosti koji je preduvjet za algoritam grožđenja jednak je $n * (n-1)/2$ ako je n broj entiteta, odnosno raste polinomijalno s obzirom na broj entiteta n.

Prva će heuristika biti testirana nad dokumentima sva tri dana na sljedeći način:

- kad se heuristika primjenjuje, odvojeno će se vršiti grožđenje za svaki od tri dana, odnosno

- kad se heuristika ne primjenjuje, grožđenje će se vršiti zajedno za dokumente objavljene u sva tri dana.

U oba će se slučaja različitim evaluacijskim mjerama rezultati uspoređivati sa zlatnim standardom.

Druga će heuristika biti testirana na uzorku pojedinog dana, tj. s primijenjenom prvom heuristikom, i to na sljedeći način:

- kad se heuristika primjenjuje, u određeni će se grozd puštati samo dokumenti s različitih portala

- kad se heuristika ne primjenjuje, u grozdove će se puštati dokumenti bez obzira na izvor

Tablica 4.6: Evaluacijske mjere i vrijeme izračuna mjera udaljenosti (t_m) te vrijeme grožđenja (t_g) s obzirom na primjenjenost prve heuristike

PH_1	C	NMI	RI	PR	POT	F_1	$F_{0.5}$	t_m	t_g
da	0.935	0.964	0.999	0.735	0.453	0.561	0.654	320.7	18.5
ne	0.894	0.966	0.999	0.575	0.67	0.622	0.593	875.5	128.5

Kao i u slučaju prve heuristike, oba će rezultata primjene, odnosno neprimjene druge heuristike biti uspoređena evaluacijskim mjerama sa zlatnim standardom.

4.3.1 Prva heuristika

Prva heuristika testira se na način da se u jednoj skupini eksperimenata odvojeno provede grožđenje za uzorak svakog dana te da se rezultat grožđenja spoji, odnosno da se grožđenje provodi nad cijelim uzorkom. Rezultati uspjeha grožđenja za $p = 0.65$ su prikazani u tablici 4.6.

Iz rezultata je vidljivo da se primjenom prve heuristike postiže veća preciznost, dok se njenom neprimjenom postiže veća potpunost. Takvi su rezultati i očekivani iz razloga što se primjenom heuristike postavlja određeno ograničenje koje spriječava pogrešne odluke, odnosno smanjivanje preciznosti. No, isto tako, s obzirom da se rezultat uspoređuje sa zlatnim stadardom koji spaja i dokumente objavljene različitih dana, primjena heuristike smanjuje potpunost iz razloga što nije moguće donijeti odluku o spajanju dokumenata objavljenih različitih dana, dok je to u slučaju neprimjenjivanja heuristike moguće. Kako se do sada mjera $F_{0.5}$ pokazala najboljom procjenom uspješnosti postupka, odnosno kako je preciznost ipak bitnija od potpunosti, prema evaluacijskim mjerama moguće je pretpostaviti da primjena heuristike donosi bolji rezultat.

Stvari postaju dodatno jasnije kad se pogledaju posljednja dva stupca u tablici 4.6, odnosno vrijeme računanja mjere udaljenosti (t_m) te vrijeme računanja grozdova (t_g). Primjena heuristike smanjuje oba vremena, i to višestruko. Vrijeme računanja matrice udaljenosti skraćuje se skoro tri puta, dok se vrijeme računanja grozdova skraćuje više od 12 puta. Razlog sma-

njenju oba vremena leži u polinomijalnom rastu kombinacija dokumenata $(n * (n-1)/2)$ s obzirom na broj dokumenata (n). Dijeljenjem skupa dokumenata u podskupove broj kombinacija dokumenata znatno se smanjuje. Vrijeme grožđenja se primjenom heuristike više smanjuje od vremena izračuna matrice iz razloga što je pri računanju matrice potrebno izračunati kosinusnu mjeru udaljenosti više parova. Računanje kosinusne mjere računalno je jednostavan zadatak. Suprotno tome, pri grožđenju jednim prolaskom se mora sortirati bitno dulja lista te se prolaskom kroz tu bitno dulju liste moraju donositi odluke, odnosno grozdovi spajati u nove grozdove. Te su operacije vremenski zahtjevnije te se zato može primijetiti veće smanjenje vremena u slučaju smanjenja broja parova dokumenata.

Treba primijetiti da bi razlika u ovim vremenima bila još drastičnija da se primjenjuje kompleksniji algoritam grožđenja od ovog jednim prolaskom, primjerice hijerarhijski.

Zaključno se može reći kako primjenom prve heuristike očekivano raste preciznost, a smanjuje se potpunost te da sveukupno mjera $F_{0.5}$ raste. Usto, vrijeme potrebno za računanje matrice udaljenosti, odnosno samo grožeđnje, višestruko se smanjuje što primjenjivanje ove heuristike na zadatku čini vrlo opravdanim.

4.3.2 Druga heuristika

Druga se heuristika testira *in vivo* tako da se u jednom slučaju dokumenti pridružuju grozdu bez obzira na to gdje su objavljeni, a u drugom se slučaju dodaju samo u ako u grozdu ne postoji dokument s tog izvora. Računalno je drugi slučaj djelomično kompliciraniji. Pri testiranju druge heuristike prva se heuristika primjenjuje, odnosno eksperiment se vrši samo na uzorku 5-SVI. Rezultati ovog eksperimenta su prikazani u tablici 4.3.2.

Kao i očekivano, u slučaju primjenjivanja druge heuristike raste preciznost, a potpunost se smanjuje. To se događa iz razloga što se ponovno, kao i u slučaju prve heuristike, uvodi dodatno ograničenje na odlučivanje čime se povremeno onemogućuje pogrešna odluka. No, isto se tako povremeno onemogućuje i dobra odluka zbog čega potpunost pada. Unatoč padu

Tablica 4.7: Evaluacijske mjere i vrijeme grožđenja (t_g) s obzirom na primjenjenost druge heuristike

PH_2	C	NMI	RI	PR	POT	F_1	$F_{0.5}$	t_g
da	0.96	0.973	0.999	0.857	0.613	0.715	0.794	8.214
ne	0.944	0.975	0.999	0.804	0.724	0.762	0.787	8.195

Tablica 4.8: Evaluacijske mjere s obzirom na primjenjenost druge heuristike na uzorcima 4-SVI i 6-SVI

uzorak	PH_2	C	NMI	RI	PR	POT	F_1	$F_{0.5}$
4-SVI	da	0.958	0.971	0.997	0.856	0.685	0.761	0.815
4-SVI	ne	0.945	0.976	0.997	0.735	0.868	0.796	0.758
6-SVI	da	0.956	0.974	0.999	0.817	0.586	0.682	0.757
6-SVI	ne	0.921	0.971	0.998	0.667	0.663	0.665	0.666

potpunosti, iz rezultata je vidljivo da je $F_{0.5}$ viša kada je druga heuristika primijenjena.

Što se tiče vremena potrebnog za grožđenje, kao što se i očekuje, vrijeme je u slučaju primjene heuristike nešto veće. Naime, pri svakom spajanju grozdova potrebno je provjeriti u indeksu dokumenata i njihovih domena je li to spajanje dozvoljeno. No, kako povećanje vremena izvođenja iznosi 19 stotinki unutar sveukupnog vremena od nešto više od 8 sekundi, odnosno od 0.2%, obzirom na korisnost primjene druge heuristike to je povećanje zanemarivo.

Utjecaj primjene ove heuristike provjeren je i na druga dva uzorka, 4-SVI i 6-SVI, te su rezultati prikazani u tablici 4.8.

Primjenom heuristike se redovito postiže viši $F_{0.5}$ s nastavkom trenda da primjena heuristike pozitivno utječe na preciznost, a negativno na potpunost, odnosno da neprimjena heuristike pozitivno utječe na potpunost, a negativno na preciznost.

Kako je izračunljivost grozdova s primjenom druge heuristike skoro identična onoj bez primjene heuristike, zaključeno je da je drugu heuristiku također poželjno primjenjivati.

Zaključno se može reći kako je pokazano da je za ovaj zadatak poželjno

primijenjivati obje heuristike s posebnim naglaskom na prvu heuristiku koja problem čini bitno lakše, a time i brže izračunljivim.

4.4 Odabir mjere težine svojstava

U idućim eksperimentima istražena je varijabla mjera težine svojstava (MTS) sa sljedećim vrijednostima:

- vjerojatnost - V
- uvjetna vjerojatnost - UV
- pojedinačna međusobna informacija - PMI
- TF-IDF mjera - TF-IDF
- t-test mjera - T-TEST

Dosada su odabrane optimalne vrijednosti sljedećih varijabli:

- koristi se algoritam grožđenja (AG) jednim prolaskom
- koristi se kosinusna mjera kao mjera udaljenosti (MU)
- primjenjuju se obje heuristike (PH_1 i PH_2)

Ostale varijable imaju svoje inicijalne vrijednosti kako su definirane na početku ovog poglavlja. Moguće vrijednosti varijable težine svojstava uspoređene su na uzorku 5-SVI. Kako različite mjere težina daju drugačije raspone vrijednosti, ponovno je potrebno eksperimentirati s parametrom p od 0.0 do 1.0 s korakom od 0.05. Za svaku mjeru težine odabran je onaj p pri kome je evaluacijska mjera $F_{0.5}$ maksimalna. Početni su rezultati prikazani u tablici 4.9.

Očiti gubitnik eksperimenta je čista vjerojatnost. To je bilo i za očekivati iz razloga što vjerojatnost jedina ne uzima u obzir razdiobu svojstava u cijelom korpusu. Iduću grupu tvore uvjetna vjerojatnost i pojedinačna

Tablica 4.9: Evaluacijske mjere i optimalni parametar p na uzorku 5-SVI s obzirom na korištenu mjeru težine svojstva

mjera	p	Č	NMI	RI	PR	PO	F_1	$F_{0.5}$
VJ	0.25	0.966	0.955	0.998	0.832	0.342	0.485	0.647
UVVJ	0.95	0.959	0.966	0.998	0.861	0.503	0.635	0.754
PMI	0.6	0.947	0.968	0.998	0.823	0.576	0.678	0.758
TF-IDF	0.7	0.954	0.976	0.999	0.858	0.673	**0.754**	**0.813**
T-TEST	0.9	0.954	0.973	0.999	0.836	0.625	0.715	0.783

Tablica 4.10: Usporedba mjera težine svojstava TF-IDF i t-test na sva tri uzorka s optimalnim parametrom p i evaluacijskim mjerama

	MTS	p	C	NMI	RI	PR	PO	F_1	$F_{0.5}$
4SV	tf-idf	0.67	0.956	0.971	0.997	0.855	0.688	**0.762**	**0.815**
4SV	t-test	0.82	0.976	0.965	0.997	0.918	0.525	0.668	0.798
5SV	tf-idf	0.69	0.957	0.976	0.999	0.86	0.667	**0.752**	**0.813**
5SV	t-test	0.88	0.959	0.971	0.999	0.861	0.572	0.687	0.782
6SV	tf-idf	0.62	0.961	0.972	0.999	0.847	0.547	**0.665**	**0.763**
6SV	t-test	0.88	0.943	0.968	0.998	0.784	0.515	0.621	0.709

međusobna informacija s time da pojedinačna međusobna informacija redovito pobjeđuje uvjetnu vjerojatnost. Kako je u ovom slučaju pojedinačna međusobna informacija zapravo logaritam uvjetne vjerojatnosti, ovaj eksperiment predstavlja još jedan dokaz upotrebljivosti logaritma u obradi prirodnog jezika, odnosno umanjivanja većih vrijednosti u korist onih manjih. Jasni pobjednici ovog eksperimenta su TF-IDF i t-test s time da TF-IDF redovito pobjeđuje t-test.

Kako razlika između t-testa i TF-IDF-a nije značajna, dodatno su istražene te dvije vrijednosti nad uzorcima 4-SVI, 5-SVI i 6-SVI, i to u rasponu od ± 0.1 od vrijednosti parametra p odabrane u prethodnom istraživanju s korakom od 0.01. Opet je odabiran onaj p koji maksimizira mjeru $F_{0.5}$. Rezultati ove grupe eksperimenata prikazani su u tablici 4.10. Rezultati pokazuju kako TF-IDF dosljedno pobjeđuje mjeru t-test.

Zaključno se može reći da je TF-IDF optimalna vrijednost varijable mjere težine svojstava (MTS) zato što redovito daje bolje rezultate od t-testa, druge

po redu optimalne vrijednosti varijable MTS. Vrijeme izračuna pojedinih mjera težina svojstava u ovim eksperimentima nije prikazivano iz razloga što je ono vrlo slično za sve mjere osim mjere vjerojatnosti koja je ovako i onako, upravo zbog svoje pretjerane jednostavnosti, očiti gubitnik ovih eksperimenata.

4.5 Odabir metoda određivanja i odabira svojstava

Predstojeći eksperimenti istražuju različite mogućnosti određivanja i odabira svojstava. Pri tome će biti istražene sljedeće varijable:

- utjecaj interpunkcija (UI) - ima li smisla ostavljati interpunkcije u formalnom zapisu ili ne
- utjecaj veličina slova (UVS) - uzimati li u obzir, i kako veličinu slova
- važnost naslova (VN) - koliko vrednovati naslov s obzirom na tekst
- isključivanje hapax legomena (IHL) - kako na rezultat utječe isključivanje svojstava koja se pojavljuju samo jednom
- isključivanje funkcijskih riječi (IFR) - kako na rezulat utječe isključivanje svojstava čija je diskriminativna vrijednost najmanja
- korjenovanje (K) - pomaže li zadatku jezično nezavisno, odnosno jezično zavisno korjenovanje
- morfosintaktičko označavanje (MO) - pomaže li statističko označavanje morfosintaktičke kategorije te određivanje leme
- utjecaj veličine korpusa na IDF mjeru (VIDF) - koliko je podataka potrebno za dobru procjenu IDF mjere svojstava s obzirom na zadatak
- prepoznavanje osobnih imena (POI) - pomaže li prethodno identificiranje imena tvrtke ili osobe te njeno korištenje kao svojstva

- statistički značajni digrami (SZD) - pomaže li korištenje statistički značajnih digrama kao svojstava

Kao što je vidljivo iz popisa, za određivanje, odnosno odabir svojstava koriste se razne statističke (IHL, SZD...) i lingvističke metode (K, MO...).

Varijable kojima je do sada određena optimalna vrijednost su sljedeće:

- algoritam grožđenja (AG) - koristi se algoritam jednim prolaskom uz optimalni p=0.65
- mjera udaljenosti (MU) - koristi se kosinusna mjera
- pretpostavljene heuristike (PH) - primjenjuju se obje heuristike
- mjera težine svojstava (MTS) - koristi se TF-IDF mjera

Preostale varijable koje se određuju u nastavku imat će inicijalne vrijednosti sve dok im se ne pronađe optimalna.

4.6 Određivanje svojstava na razini pojavnica

U ovom potpoglavlju bit će razmotrene različite metode određivanja svojstava na razini pojavnica. Bit će istražene sljedeće predviđene varijable:

- utjecaj interpunkcija (UI) - ima li smisla ostavljati interpunkcije u formalnom zapisu ili ne
- utjecaj veličina slova (UVS) - uzimati li u obzir, i kako, veličinu slova
- važnost naslova (VN) - koliko vrednovati pojavnice naslova s obzirom na tekst

Tablica 4.11: Utjecaj interpunkcija na evaluacijske mjere F_1 i $F_{0.5}$

	4-SVI		5-SVI	
UI	F_1	$F_{0.5}$	F_1	$F_{0.5}$
da	0.757	0.813	0.713	0.795
ne	**0.759**	**0.814**	**0.714**	**0.796**

4.6.1 Utjecaj interpunkcija

U ovoj grupi eksperimenata istražuje se imaju li interpunkcije kakav učinak, odnosno nose li kakvu informaciju korisnu za zadatak pronalaženja događaja. Nad uzorcima 4-SVI i 5-SVI pokrenut je algoritam nad tekstom opojavničenim uz uključivanje, odnosno isključivanje interpunkcijskih znakova. Rezultati su prikazani u tablici 4.11. Iz rezultata je vidljivo da interpunkcije ne pomažu zadatku, odnosno da predstavljaju dodatni šum u formalnom prikazu dokumenata. Isto je tako moguće, kako su uzorci dokumenata prikupljani s određenim nečistoćama, da je izbacivanjem svega što ne počinje nekim slovom, zapravo, više šuma uklonjeno nego što je informacije u tim interpunkcijama zapisano.

Zaključno je moguće reći da se istraživanjem varijable utjecaja interpunkcija (UI) zaključilo kako uključivanje interpunkcija negativno utječe na rezultat. Iz tog razloga interpunkcije i dalje neće biti uključivane u popis svojstava.

4.6.2 Utjecaj veličine slova

U ovoj grupi eksperimenata istražuje se utjecaj veličine slova na krajnji rezultat zadatka. Do sada su sva slova pretvarana u mala, odnosno zanemarivana je informacija koja je kodirana veličinom slova. U prvom eksperimentu suprostavljena su dva prikaza dokumenta - onaj gdje se sva slova smanjuju te onaj gdje veličina slova ostaje kakva je. Rezultati su prikazani u tablici 4.6.2.

Rezultati eksperimenta su jednoznačni. Zadržavanjem podatka o veličini slova postižu se slabiji rezultati. Ovakav je rezultat i za očekivati iz razloga što se ne postiže maksimalna unifikacija svojstava, odnosno pojavnica na

Tablica 4.12: Utjecaj veličine slova na evaluacijske mjere F_1 i $F_{0.5}$ na uzorcima 4-SVI i 5-SVI

	4-SVI		5-SVI	
UVS	F_1	$F_{0.5}$	F_1	$F_{0.5}$
mala	**0.766**	**0.819**	**0.719**	**0.796**
kakva jesu	0.75	0.809	0.704	0.787

Tablica 4.13: Neke čestote pojavnica u unutrašnjosti rečenice s obzirom na veličinu slova

bih	3853
Bih	2
BIH	188
BiH	11827
hrvatska	3575
Hrvatska	14064

početku rečenice i one u sredini se smatraju različitim pojavnicama. U idućoj grupi eksperimenata pokušat će se statistički intervenirati u veličinu slova te odrediti prema sadržaju unutrašnjosti rečenice treba li prvu riječ u rečenici prikazati malim ili velikim početnim slovom.

Za potrebe takvog statističkog odlučivanja izračunata je razdioba veličine slova pojavnica koje se ne nalaze na početku rečenice. Neke dobivene vrijednosti prikazane su u tablici 4.13. Iz ovih je podataka primjerice vidljivo da je pojavnica "Bih" u odnosu na pojavnicu "bih" vrlo rijetka te da je vjerojatno točnije pojavnicu na početku rečenice koja ima oblik "Bih" pretvoriti u "bih". Isto tako, u slučaju da je pojavnica oblika "BIH" ili "BiH", neće se htjeti intervenirati iz razloga velike čestote takvih oblika. U slučaju da se na početku rečenice nalazi pojavnica "Hrvatska", neće se intervenirati iz razloga što je pojavnica "Hrvatska" bitno češća od one "hrvatska". Ovakav pristup ne uzima u obzir okolinu, već samo osnovne čestotne dokaze iz korpusa, no zasigurno do određene mjere riješava problem velikog slova na početku rečenice. Preostaje jedino istražiti hoće li ovakav pristup pomoći zadatku pronalaženja događaja.

Tablica 4.14: Utjecaj veličine slova na evaluacijske mjere F_1 i $F_{0.5}$ na uzorcima 4-SVI i 5-SVI

	4-SVI		5-SVI	
UVS	F_1	$F_{0.5}$	F_1	$F_{0.5}$
mala	**0.766**	**0.819**	**0.719**	**0.796**
kakva jesu	0.75	0.809	0.704	0.787
statistički	0.76	0.818	0.709	0.791

Izračunata će razdioba, dakle, biti korištena za donošenje odluke kako prikazati pojavnice koje se nalaze na početku rečenice. Uvedeno je jednostavno pravilo - bit će prikazane onako kako su češće pisane u sredini rečenice. Ako pojavnica nije pisana samo početnim velikim slovom, već samo velikim slovima će se, u slučaju da u razdiobi postoji dokaz o više od tri tako pisane pojavnice u korpusu, pojavnica prikazati onako kako je zapisana. U suprotnom će se slučaju prikloniti češćem obliku. U slučaju da je pojavnica pisana miješano velikim i malim slovima, njen oblik neće biti mijenjan. Rezultati usporedbe prikaza uzoraka 4-SVI i 5-SVI malim slovima, slovima kakva jesu te statistički određeno veličinom slova prikazani su u tablici 4.14.

Iz rezultata je vidljivo kako je prikaz malim slovima i dalje najuspješniji, dok je statističko određivanje veličine slova uspješnije od pristupa neintervencije u veličinu slova. To dokazuje smislenost postupka statističkog određivanja veličine slova. Za ovaj je pak zadatak i ovu kvalitetu podataka ovaj pristup i dalje nepovoljniji od onoga kad se zanemaruje veličina slova, no manje nepovoljan od onoga kada se prikazuje veličina kakva je u tekstu. Moguće bi bilo dodatno pospješiti algoritam za predviđanje veličine slova, no za pretpostaviti je da bi teško dostigao razinu onoga gdje se zanemaruje veličina slova. Treba uzeti u obzir kompleksnost takvog izračuna, dok je pretvaranje svih slova u mala izrazito jednostavan postupak.

Dva moguća razloga za uspješnost prikaza malim slovima su sljedeća:

1. pojavnice s različitom veličinom slova često prikazuju iste ili slične koncepte čime se postiže oblik konceptualne unifikacije, a time se više povećava potpunost nego što se ruši preciznost

Tablica 4.15: Utjecaj broja ponavljanja pojavnica iz naslova na evaluacijske mjere F_1 i $F_{0.5}$ na uzorcima 4-SVI, 5-SVI i 6-SVI

	4-SVI		5-SVI		6-SVI	
čestota po naslovu	F_1	$F_{0.5}$	F_1	$F_{0.5}$	F_1	$F_{0.5}$
1	0.74	0.807	0.703	0.795	0.633	0.715
2	0.771	0.825	0.72	0.8	0.661	0.735
3	0.766	0.819	0.719	0.796	0.675	0.747
4	**0.774**	**0.832**	**0.723**	**0.798**	0.68	0.754
5	0.755	0.818	0.716	0.795	**0.681**	**0.766**
6	0.725	0.797	0.702	0.787	0.656	0.751

2. nečistoća izvora se do određene mjere umanjuje zanemarivanjem veličine slova.

Za očekivati je da je prvi razlog dominantniji, no moguće je da i drugi ima određeni utjecaj.

4.6.3 Važnost naslova

Na kraju dijela s rezultatima koji prikazuju uspješnost metoda određivanja svojstava na razini pojavnica želi se istražiti važnost naslova za dokument. Do sada je pojava svojstva u naslovu računata kao tri pojave te pojavnice u tekstu. Rezultati eksperimenata s brojem ponavljanja svojstava iz naslova nalaze se u tablici 4.15.

Rezultati pokazuju da u uzorcima 4-SVI i 5-SVI optimalan rezultat daju 4 ponavljanja svojstava iz naslova, dok je za uzorak 6-SVI optimalni broj ponavljanja 5. Mogući razlog tom pomaku u uzorku 6-SVI je onaj koji je primijećen ručnim pregledavanjem sadržaja uzorka u odnosu na druga dva uzorka. U tom se skupu dokumenata nalaze dokumenti koje je teže podijeliti prema događaju koji opisuju. Poticaj za takav uvid u uzorke dobiven je zbog činjenice što su u dosadašnjem tijeku istraživanja redovito primijećene niže evaluacijske vrijednosti nad tim uzorkom. Uzorak 5-SVI redovito zauzima drugo mjesto, dok se uzorak 4-SVI, što zbog sadržaja dokumenata, što zbog manje količine dokumenata, redovito pokazuje kao uzorak na kojemu je

najjednostavnije riješiti zadani problem.

Moguće objašnjenje činjenice da uzorak 6-SVI preferira češće ponavljanje pojavnica iz naslova zbog svoje veće kompleksnosti jest da su podaci koji se nalaze u naslovu vrlo čisti i ukazuju na bit dokumenta te njihovo često ponavljanje smanjuje utjecaj kompleksnijeg sadržaja dokumenata.

Na temelju ovih rezultata odlučeno je ponavljati svako svojstvo iz naslova četiri puta.

Zaključno je, što se tiče određivanja svojstava na razini pojavnica, moguće reći da uključivanje interpunkcija te podatka o veličini slova ne pomažu zadatku pronalaženja događaja. Bilo bi za pretpostaviti da interpunkcije ne bi trebale nikako utjecati na uspješnost izvršavanja zadatka zato što se može pretpostaviti da je distribucija određenih interpunkcija kroz dokumente uniformna, odnosno da nema nikakvu diskriminativnu vrijednost. Činjenica da interpunkcije kvare rezultat navode na zaključak da dokumenti imaju određenu količinu nečistoća koje se do određenog dijela izbacuju uključivanjem samo onih pojavnica koje počinju slovom. Nadalje, što se tiče uključivanja naslova u formalni prikaz dokumenta, zaključeno je kako je optimalno četiri puta ponoviti pojavnice iz naslova.

4.7 Odabir svojstava na razini pojavnica

U ovom se potpoglavlju eksperimentira odabirom svojstava među ponuđenima, odnosno vrijednostima sljedećih varijabli:

- isključivanje hapax legomena (IHL) - odabiru se sva svojstva osim onih koja se pojavljuju manje od dva, tri, odnosno četiri puta
- isključivanje funkcijskih riječi (IFR) - odabiru se sva svojstva osim onih koja su u određenom rangu popisa funkcijskih riječi računatom mjerom $\log IDF$

Tablica 4.16: Broj svojstava koja se pojavljuju manje od dva, odnosno tri puta u korpusu

broj pojava	broj svojstava	postotak
n	488,081	1.0
$n < 2$	200,954	0.412
$n < 3$	265,275	0.544

4.7.1 Isključivanje hapax legomena

Prvi eksperiment uključuje isključivanje hapax legomena, odnosno svojstava koja se u korpusu pojavljuju samo jednom. Od ove se intervencije u pravilu ne očekuje poboljšanje rezultata, no zbog Zipfove razdiobe svojstava u tekstu moguće je znatno pojednostavniti model prikaza dokumenta. Eksperimentira se s isključivanjem svojstava iz korpusa koja su se pojavila jednom, odnosno dvaput. Broj takvih svojstava prikazan je u tablici 4.16. Iz tablice je vidljivo da se na razini korpusa isključivši samo hapax legomena broj svojstava skoro prepolovio. Prema Zipfovoj razdiobi, uklanjanjem svojstava koja se pojavljuju određeni broj puta, broj svojstava logaritamski opada.

U prvom se eksperimentu izbacuje svojstva koja se u korpusu pojavljuju samo jednom ($n > 1$), dvaput ($n > 2$) ili triput ($n > 3$). Rezultati su prikazani u tablici 4.17.

Zaključak koji rezultati sugeriraju je u jednu ruku iznenađujući - u dva od tri uzorka izbacivanje najrjeđih svojstava, koja bi trebala biti i vrlo informativna, popravlja rezultate. Takav rezultat navodi na zaključak da se zapravo najrjeđa svojstva ne dijele među dokumentima koji opisuju isti događaj te time njihovo uključivanje u model dokumenta kvari rezultat. Jedini uzorak u kojemu izbacivanje hapax legomena kvari rezultat je "teški" uzorak 6-SVI. U slučaju tog uzorka očito vrlo rijetka svojstva ipak uspijevaju daljnje diskriminirati dokumente, te one koji ne opisuju isti događaj držati odvojenima. Generalno se može zaključiti kako je općenito korisno prihvatiti pravilo $n > 2$, odnosno da u popis ulaze samo svojstva koja su se u korpusu pojavila tri ili više puta s iznimkom težih zadataka gdje je poželjno ostaviti sva svojstva.

Činjenica da se teži zadaci redovito drugačije ponašaju od lakših navodi

Tablica 4.17: Utjecaj izbacivanja pojavnica koje se pojavljuju samo određeni broj puta na evaluacijske mjere F_1 i $F_{0.5}$ na uzorcima 4-SVI, 5-SVI i 6-SVI

	4-SVI		
svojstva	broj svojstava	F_1	$F_{0.5}$
$n > 0$	21,958	0.774	0.831
$n > 1$	20,954	0.775	0.832
$n > 2$	20,498	0.775	0.832
$n > 3$	20,168	0.775	0.832
	5-SVI		
	broj dimenzija	F_1	$F_{0.5}$
$n > 0$	34,226	0.723	0.798
$n > 1$	32,657	0.727	0.801
$n > 2$	31,864	0.73	0.802
$n > 3$	31,168	0.725	0.801
	6-SVI		
	broj dimenzija	F_1	$F_{0.5}$
$n > 0$	33,945	0.68	0.754
$n > 1$	32,498	0.677	0.748
$n > 2$	31,701	0.678	0.748
$n > 3$	31,102	0.675	0.743

Tablica 4.18: Popis svojstava s najmanjim logaritmom mjere IDF

u	0.045	s	0.498	ne	0.967	još	1.26	rekao	1.49
je	0.056	koji	0.567	do	0.967	sa	1.29	već	1.507
i	0.08	će	0.659	nije	0.984	po	1.308	no	1.518
na	0.133	kako	0.721	to	0.996	ali	1.309	bio	1.534
se	0.183	te	0.749	godine	1.118	biti	1.315	dok	1.537
za	0.269	iz	0.754	koje	1.124	prema	1.375	dana	1.544
su	0.299	o	0.77	koja	1.135	jer	1.415	prije	1.595
da	0.328	što	0.786	kao	1.139	sve	1.432	ili	1.624
od	0.394	bi	0.934	zbog	1.163	oko	1.464	uz	1.659
a	0.417	nakon	0.957	više	1.253	samo	1.467	danas	1.67

na zaključak kako bi vjerojatno bilo korisno moći razlikovati te slučajeve. Iz tog bi razloga bilo zanimljivo istražiti metode kojima se to razlikovanje može postići.

4.7.2 Isključivanje funkcijskih riječi

Nadalje se istražuje ima li izbacivanje najčešćih svojstava - funkcijskih riječi - pozitivan utjecaj na rješavanje zadatka. U pravilu bi trebalo očekivati da će ih TF-IDF mjera ovako i onako učiniti nebitnima te se izbacivanjem funkcijskih riječi ne bi trebao postići značajan napredak. Svojstva s najmanjim logaritmom IDF-a navedena su u tablici 4.18.

Rezultati izbacivanja različitog broja funkcijskih riječi prikazani su u tablici 4.19. Podaci navode na iznenađujući zaključak da izbacivanje ključnih riječi pomaže zadatku grožđenja, no ne nužno onih prvih, već najviše onih u rasponu od 100 i 150. Popis tih funkcijskih riječi je dan u tablici 4.7.2. Mogući razlog tome je neoptimalnost TF-IDF mjere u tom području, odnosno činjenica da TF-IDF mjera ne kažnjava dovoljno svojstva u tom rasponu. Zanimljivo je također za primijetiti da izbacivanje funkcijskih riječi iz tog raspona najviše pomaže u najtežim zadacima, u ovom slučaju u uzorku 6-SVI, dok u jednostavnijim zadacima kao što je 4-SVI ono ne čini razliku.

Smanjivanje broja svojstava prikazanih u prethodnim tablicama može iz-

Tablica 4.19: Utjecaj izbacivanja funkcijskih riječi prema mjeri logaritma IDF-a određenog raspona (IF) i smanjenog broja dimenzija ($-BD$) na evaluacijske mjere F_1 i $F_{0.5}$ na uzorcima 4-SVI, 5-SVI i 6-SVI

		4-SVI		5-SVI		6-SVI	
IF	$-BD$	F_1	$F_{0.5}$	F_1	$F_{0.5}$	F_1	$F_{0.5}$
0	0	0.774	0.832	0.723	0.798	0.68	0.754
0-50	50	0.774	0.832	0.721	0.797	0.678	0.754
0-100	100	0.774	0.832	0.721	0.797	0.681	0.758
0-150	150	0.774	0.832	0.724	0.799	0.683	0.759
0-200	200	0.774	0.832	0.724	0.799	0.681	0.758
0-250	250	0.774	0.832	0.72	0.796	0.683	0.759
50-100	50	0.774	0.832	0.727	**0.8**	0.68	0.754
100-150	50	0.774	0.832	**0.73**	**0.8**	**0.685**	**0.76**
150-200	50	0.774	0.832	0.723	0.798	0.681	0.757
50-150	100	0.774	0.832	0.727	0.8	0.685	0.759

Tablica 4.20: Popis funkcijskih riječi u rasponu od 100 do 150 čije izbacivanje pokazuje najbolje rezultate

kod	taj	pod	sati	neće
čak	ljudi	piše	kazao	kaže
tome	kojem	policija	kad	im
hina	mogu	izjavio	pet	bili
10	tom	ipak	hrvatska	četiri
treba	dio	radi	20	gotovo
15	odnosno	riječ	novi	svi
mi	eura	tek	vrlo	osim
nešto	dalje	naime	njih	cijena
pred	jučer	vlada	put	svoj

Tablica 4.21: Prosječna duljina stvarnog vektora u memoriji u slučaju neizbacivanja funkcijskih riječi (IF=0), odnosno izbacivanja onih u rasponu od 100 do 150 (IF=100-150) u sva tri uzorka

uzorak	IF=0	IF=100-150	postotak smanjenja
4-SVI	158.35	153.84	0.0285
5-SVI	153.06	148.54	0.0295
6-SVI	156.82	152.06	0.0304
prosjek	156.08	151.48	0.0295

gledati nebitno, no ono to ne mora biti iz razloga što se vektori u memoriji ne prikazuju kao stvarni vektori zato što je većina dimenzija jednaka nuli. Naime, u tim se vektorima bilježe samo one dimenzije za koje pojedina svojstva imaju vrijednost. Prosječna duljina stvarnog vektora u memoriji u sva tri uzorka prikazana je u tablici 4.21. Kako je prosječni broj dimenzija oko 155, izbacivanje 100 najnediskriminativnijih dimenzija značajno bi umanjilo veličinu prikaza pojedinog dokumenta.

Kako najbolje rezultate pokazuje izbacivanje funkcijskih riječi u rasponu od 100 do 150, izbacuje se zapravo samo 50 svojstava, i to ne onih najnediskriminativnijih, tj. najčešćih. Iz rezultata je vidljivo da je broj izbačenih svojstava redovito značajno manji te je za izbačenih 50 funkcijskih riječi broj svojstava po prikazu dokumenta umanjen za prosječno samo 4 svojstva, odnosno 3% veličine prikaza.

Iz ove je analize moguće zaključiti kako izbacivanje funkcijsih riječi nije isplativo raditi na cijelom popisu, već nešto dublje, u ovom slučaju u rasponu od 100 do 150. Isto tako se može zaključiti kako takav odabir svojstava ne smanjuje značajno kompleksnost prikaza dokumenta, već da samo pozitivno utječe na uspješnost rješavanja zadatka.

Općenito se o odabiru svojstava na razini pojavnica može zaključiti kako se izbacivanjem hapax legomena i funkcijskih riječi zadatak pronalaženja događaja riješava uspješnije. Izbacivanje hapax legomena prikaz dokumenta značajno pojednostavnjuje za razliku od izbacivanja funkcijskih riječi gdje se prikaz dokumenta zanemarivo pojednostavnjuje.

4.8 Morfološka normalizacija

U ovom su potpoglavlju prikazani rezultati eksperimentiranja morfološkom normalizacijom. Pod morfološkom se normalizacijom pretpostavlja svođenje svih oblika nekog leksema na jedan uniformni oblik. Prednost ovakvog prikaza jest očita - omogućuje se prepoznavanje koncepta bez obzira na paradigmatski oblik leksema u kojem se on pojavljuje.

Dva problema povezana s morfološkom normalizacijom su

- svaki takav postupak sa sobom donosi pogreške, odnosno uvođenje dodatnog šuma
- korištenjem morfološki normaliziranih oblika dolazi do gubitka informacije, naime, morfosintaktička kategorija određene pojavnice također ima informacijsku vrijednost

Kao mogućnost rješavanja drugog problema nameće se kombinacija morfološki normaliziranih i nenormaliziranih svojstava. Takav bi pristup za rezultat imao kompleksniji prikaz, naime broj svojstava bi bio veći, no moguće je da bi sa sobom nosio i određeni napredak. Navedeni pristup prelazi granice ovog rada te se njime neće eksperimentirati.

4.8.1 Korjenovanje

Korjenovanje je najjednostavniji oblik morfološke normalizacije koji se sastoji od svođenja svake pojavnice na njen korijen. U pravilu se razlikuju dvije razine korjenovanja:

- flektivno korjenovanje u kojemu se uklanjaju samo nastavci paradigme (razum—om, razum—a)
- tvorbeno korjenovanje u kojemu se pojavnica svodi na leksički morfem čime se često probija granica paradigme jednog leksema (rad—no, rad—imo)

Korištenje tvorbenog korjenovanja ima svoje zagovornike i protivnike, no u većini istraživanja nije pokazalo ohrabrujuće rezultate. Iz tog će se razloga u ovom doktorskom radu koristiti samo algoritmi za flektivno korjenovanje.

Jedna dodatna važna značajka korjenovanja je ta što je ono u većini slučajeva kontestualno nezavisno. Naime, u svođenju pojavnice na neki osnovni oblik ne uzima se u obzir okolina u kojoj se pojavnica nalazi. Nasuprot tomu, morfosintaktičko je označavanje izrazito kontekstualno čime je bitno kompleksnije, no i jezično točnije.

U sklopu ove grupe eksperimenata vrše se dvije razine korjenovanja:

- ona jednostavna, jezično nezavisna koja se vodi pravilom da uklanja kraj pojavnice do posljednjeg samoglasnika uz uvjet da nije više uklonjeno nego što je preostalo
- složenija koja sadrži pravila za imenice i pridjeve hrvatskog jezika

Složeniji je algoritam za korjenovanje razvijen na Odsjeku za informacijske znanosti te se trenutno sastoji od 178 pravila zamjene završetka riječi.

Nominalna varijabla korjenovanja K, dakle, ima tri moguće vrijednosti

- N - ne vrši se korjenovanje
- J - vrši se jednostavno korjenovanje
- S - vrši se složeno korjenovanje

Rezultati usporedbe različitih razina korjenovanjem prikazani su u tablici 4.22.

Rezultati ukazuju na činjenicu da korjenovanje znatno pojednostavnjuje model prikaza dokumenta prepolavljajući apsolutni broj svojstava. Jednostavno, jezično nezavisno korjenovanje značajnije umanjuje broj svojstava.

Što se evaluacije na zadatku tiče, i jednostavno i složeno korjenovanje nešto kvare rezultat. Jednostavno korjenovanje pritom redovito postiže lošije rezultate.

Ponovno je moguće prijmijetiti trend da dodatne metode kod jednostavnijih problema kvare rezultat, dok kod težih one nemaju negativan utjecaj ili

Tablica 4.22: Utjecaj različitih razina korjenovanja (K) na broj dimenzija (BD) te evaluacijske mjere F_1 i $F_{0.5}$ na uzorcima 4-SVI, 5-SVI i 6-SVI

	4-SVI			5-SVI			6-SVI		
K	BD	F_1	$F_{0.5}$	BD	F_1	$F_{0.5}$	BD	F_1	$F_{0.5}$
B	21,958	0.774	0.832	34,226	0.723	0.798	33,945	0.68	0.754
J	14,608	0.777	0.81	21,707	0.738	0.778	21,324	0.697	0.724
S	17,454	0.776	0.818	26,667	0.743	0.796	26,376	0.707	0.745

je on čak lagano pozitivan. Tako se može primijetiti da složeno korjenovanje na prva dva uzorka kvari rezultat, dok ga u slučaju trećeg, teškog uzorka, nešto poboljšava. Ista se tendencija može primijetiti i kod jednostavnog korjenovanja.

Što se tiče razlike u rezultatima evaluacijskih mjera F_1 i $F_{0.5}$, mjera F_1 se korjenovanjem redovito popravlja, dok $F_{0.5}$ od te metode pati. To navodi na zaključak kako zapravo korjenovanje pozitivno utječe na potpunost smanjujući preciznost. Takav se rezultat može i pretpostaviti, naime, svakom vrstom unifikacije elementi koji se uspoređuju postaju sličniji te se time može očekivati i povećanje potpunosti rješenja.

Metode korjenovanja općenito u sklopu ovih eksperimenata ne donose značajno poboljšanje rezultata te je time odlučeno kako daljnje korjenovanje neće biti provođeno.

4.8.2 Morfosintaktičko označavanje

Drugi oblik morfološke normalizacije - morfosintaktičko označavanje - statistička je nadzirana metoda koja na temelju označenog uzorka posjeduje izračunate vjerojatnosti niza određenih morfosintaktičkih kategorija. Na temelju morfosintaktičkog leksikona i statističkog modela za predviđanje ona vrši označavanje svake pojavnice pripadajućom morfosintaktičkom kategorijom te lemom. Označivač korišten u ovom doktorskom radu razvijen je u Zavodu za lingvistiku Filozofskog fakulteta u Zagrebu.

Rezultat usporedbe triju prikaza dokumenata - onog pojavnicama, lemama, odnosno i pojavnicama i lemama - prikazan je u tablici 4.23.

Tablica 4.23: Utjecaj uključivanja pojavnica (P), lema (L), odnosno pojavnica i lema (P+L) u prikaz dokumenata na broj dimenzija (BD) te evaluacijske mjere F_1 i $F_{0.5}$ na uzorcima 4-SVI, 5-SVI i 6-SVI

	4-SVI			5-SVI			6-SVI		
	BD	F_1	$F_{0.5}$	BD	F_1	$F_{0.5}$	BD	F_1	$F_{0.5}$
P	21,958	0.774	0.832	34,226	0.723	0.798	33,945	0.68	0.754
L	14,849	0.785	0.824	23,236	0.741	0.788	22,500	0.704	0.737
P+L	36,807	0.775	0.832	57,462	0.737	0.805	56,445	0.682	0.73

Rezultati ukazuju da se broj dimenzija u slučaju prikaza dokumenta lemama smanjuje u prosjeku za 30 posto. U slučaju uključivanja i pojavnica i lema u prikaz dokumenta broj dimenzija, naravno, raste, i to za očekivanih 40 posto.

Evaluacijske mjere u slučaju korištenja samo lema pokazuju mali napredak, dok se taj napredak u slučaju uključivanja i pojavnica i lema zapravo ne primjećuje. Jedan od mogućih razloga za tu razliku je uvećanje broja dimenzija u slučaju drugog prikaza što u pravilu negativno utječe na rezultat.

Općenito je moguće zaključiti kako morfosintaktičko označavanje ne pridonosi redovito rješavanju problema pronalaženja događaja. Uzevši u obzir kompleksnost te metode, sa sigurnošću se ova metoda može odbaciti.

Jedan od mogućih razloga neuspjehu obje metode morfološke normalizacije leži vjerojatno i u prirodi problema - pokušava se otkriti sve dokumente koji izvještavaju o određenom događaju. Moguće je pretpostaviti da će, govoreći o nekom događaju, autori redovito uz iste lekseme koristiti te lekseme i u istim morfosintaktičkim kategorijama. Time se morfosintaktička kategorija može pretpostaviti kao vrijedan podatak. Korištenje obaju podataka također ne utječe pozitivno na rješenje. U tom je slučaju vjerojatni krivac znatan rast broja dimenzija prikaza dokumenta.

Tablica 4.24: Čestotna razdioba oznaka osobnih imena u uzorcima 4-SVI, 5-SVI i 6-SVI

oznaka	čestota
<Prezime>	4615
<Ime>	4574
<Naselje>	3097
<osoba>	2420
<Osoba>	2195
<Zanimanjefun>	279
<Subjekt>	279
<Zanimanjefun2>	161
<Titulaprefiks>	95
<subjekt>	84
<Funkcija>	61
<Titulasufiks>	7

4.9 Prepoznavanje osobnih imena

Prepoznavanje osobnih imena (engl. *named entity recognition*) često je korištena metoda u mnogim područjima obrade prirodnog jezika, poput pretraživanja informacija, klasifikacije dokumenata, ekstrakcije informacija i sl. Ovdje se istražuje mogući utjecaj te metode na zadatak pronalaženja događaja. Prepoznavanje osobnih imena izvršeno je od strane Zavoda za poslovna istraživanja od kojega je preuzet i uzorak za istraživanje. U tri korištena uzorka oznake kojima su označena osobna imena imaju radzdiobu prikazanu u tablici 4.24.

Prepoznavanja osobnih imena vršena su pomoću regularnih gramatika. Kao pomoć pri oblikovanju regularnih gramatika korišteni su morfološki leksikoni imena, prezimena, naseljenih mjesta, poslovnih subjekata i drugih.

Oznakom <Prezime> označena su sva prezimena u tekstu, primjerice <Prezime>Maček</Prezime>.

Oznakom <Ime> označena su sva imena u tekstu, primjerice <Ime>Jasminku</Ime>.

Oznakom <Naselje> označena su sva naselja u tekstu, primjerice

<Naselje>Slatini</Naselje>.

Oznakom <osoba> označene su sve poznate osobe (osobe u bazi podatata), te je toj oznaci redovito pridružen atribut oso_id, primjerice <osoba oso_id="103427751"><Ime>Mira</Ime> <Prezime>Biljan Komorčec</Prezime></Osoba>. Oznakom <Osoba> označene su nepoznate osobe, odnosno niz imena i prezimena koji ne postoji u bazi podataka, primjerice <Osoba><Ime>Rudolf</Ime> <Prezime>Mayer</Prezime></Osoba>.

Oznakama <Zanimanjefun> i <Zanimanjefun2> označeni su navodi zanimanja, odnosno funkcija označenih osoba, i to prije (oznaka <Zanimanjefun>), odnosno poslije imena (oznaka <Zanimanjefun2>). Primjeri oznaka jesu <Zanimanjefun>gradonačelnik</Zanimanjefun> i <Zanimanjefun2>sudac</Zanimanjefun2>.

Oznakom <Subjekt> označeni su nepoznati poslovni subjekti, primjerice turističke agencije <Subjekt>Collegium</Subjekt>, dok su oznakom <subjekt> s atributom sub_id označeni poznati poslovni subjekti, primjerice <subjekt sub_id="68520">KPMG Croatia d.o.o.</Subjekt>

Oznakama <Titulaprefiks> i <Titulasufiks> označene su pripadajuće titule prepoznatim imenima ispred, odnosno iza imena. Primjeri oznaka su <Titulaprefiks>dr.</Titulaprefiks> i <Titulasufiks>producent</Titulasufiks>.

Oznakama <Funkcija> označene su funkcije pripadajućih imena, primjerice <Funkcija>ministar</Funkcija>.

Od svih oznaka najviše se može očekivati od oznaka <osoba> i <subjekt> jer jedine vrše ne samo prepoznavanje osobnog imena, već i njegovu identifikaciju, odnosno jednoznačno određuju entitet. Iz tog su razloga za početak uključene samo te oznake. U slučaju da te oznake ne pokažu napredak, može se pretpostaviti da ni ostale oznake neće polučiti isti. U eksperiment su, dakle, uključena svojstva poput <osoba oso_id="101451608">, odnosno <subjekt sub_id="68520">. Takvih svojstva u sva tri uzorka pronađeno je 2504. Rezultati tog eksperimenta prikazani su u tablici 4.25.

Iz rezultata je vidljiva već prije prepoznata činjenica - na lakše zadatke dodatne metode nemaju nikakav ili negativan utjecaj, dok je na teže zadatke

Tablica 4.25: Utjecaj uključivanja prepoznatih osobnih imena osoba i poslovnih subjekata (POI) u prikaz dokumenata na evaluacijske mjere F_1 i $F_{0.5}$ na uzorcima 4-SVI, 5-SVI i 6-SVI

	4-SVI		5-SVI		6-SVI	
POI	F_1	$F_{0.5}$	F_1	$F_{0.5}$	F_1	$F_{0.5}$
NE	0.763	0.825	0.747	0.817	0.685	0.759
DA	0.763	0.825	0.739	0.812	0.686	0.759

(uzorak 6-SVI) taj utjecaj pozitivniji. Općenito je moguće primijetiti da utjecaj prepoznavanja osobnih imena nije značajan ni kad se uključe najkorisnija svojstva. Iz tog se razloga odustaje od daljnjih eksperimenata.

Mogući razlog za ovakve rezultate je onaj sličan neuspjehu morfološke normalizacije. Naime, samo pojavljivanje imena već je očito dovoljno informativno te daljnje označavanje tog imena jednoznačnim identifikatorom ne daje dovoljno nove informacije. Iz prošlih eksperimenata s morfološkom normalizacijom može se pretpostaviti da će se osobna imena, opisujući određeni događaj, pojavljivati u identičnim morfosintaktičkim kategorijama što prepoznavanje tih osobnih imena čini dodatno nepotrebnim.

4.10 Sintaktička obrada

U ovom potpoglavlju eksperimentira se s elementima teksta većima od pojedine pojavnice, odnosno s višečlanim izrazima. Popis svojstava se postupno proširuje digramima koji su hi-kvadrat testom pokazali najveću mjeru neslučajnog supojavljivanja.

Rezultati zadatka pronalaženja događaja uz uključivanje N digrama s najvišim hi-kvadrat rezultatom su dani u tablici 4.26

Rezultati pokazuju da uključivanje višečlanih svojstava dobivenih statističkom metodom ne pospješuju rezultate.

Općenito je moguće zaključiti kako niti jedna naprednija tehnika u zadatku pronalaženja događaja nije polučila pozitivne rezultate. Razlog tome je moguće tražiti u činjenici da u ovakvoj paradigmi prikaza dokumenta,

Tablica 4.26: Utjecaj uključivanja najjačih N dvočlanih kolokacija prema hi-kvadrat testu u prikaz dokumenata na evaluacijske mjere F_1 i $F_{0.5}$ na uzorcima 4-SVI, 5-SVI i 6-SVI

	4-SVI		5-SVI		6-SVI	
N	F_1	$F_{0.5}$	F_1	$F_{0.5}$	F_1	$F_{0.5}$
0	0.763	0.825	0.747	0.817	0.685	0.759
500	0.763	0.825	0.747	0.817	0.685	0.759
700	0.763	0.825	0.747	0.817	0.685	0.759
900	0.763	0.825	0.748	0.817	0.681	0.757
1,100	0.772	0.826	0.738	0.81	0.681	0.755
1,300	0.767	0.822	0.735	0.807	0.683	0.758
1,500	0.769	0.823	0.733	0.78	0.677	0.753

dakle dokumenta kao jednostavnog vektora svojstava koje se potom grupira grožđenjem, dodatne jezične metode ne predstavljaju razliku. Rezultati navode na zaključak kako je potrebno raditi na novoj paradigmi prikaza značenja dokumenta i njegove obrade kako bi ovakve metode mogle ostvariti uspjeh.

4.11 Odnos veličine referentnog korpusa na TF-IDF mjeru

Posljednja grupa pokusa odnosi se na odnos veličine referentnog korpusa i uspješnosti mjere težine svojstava koja zahtijeva referentni korpus. U poglavlju 4.4 od svih predloženih mjera težine svojstava odabrana je mjera TF-IDF tako da će ovdje biti eksperimentirano s utjecajem veličine referentnog korpusa na tu mjeru, odnosno uspješnost te mjere da prikaže težinu svojstva u zadatku pronalaženja događaja.

Eksperimentira se s dva različita algoritma - jednim koji očekuje da će u referentnom korpusu biti sva svojstva koja se pojavljuju u dokumentima koje treba analizirati, te drugim gdje ta pretpostavka ne mora biti zadovoljena. U slučaju da za neko svojstvo ne postoji mjera iz referentnog korpusa, vrši se najjednostavniji oblik izravnjavanja (engl. *smoothing*) - pretpostavlja se da

Tablica 4.27: Utjecaj veličine referentnog korpusa (VRK) te algoritma za računanje mjera na referentnom korpusu na evaluacijske mjere F_1 i $F_{0.5}$ na uzorku 5-SVI

	ALG1		ALG2	
VRK	F_1	$F_{0.5}$	F_1	$F_{0.5}$
1,000	0.71	0.797	0.745	0.811
2,000	0.715	0.797	0.747	0.817
3,000	0.725	0.803	0.729	0.805
4,000	0.73	0.8076	0.726	0.805
5,000	0.717	0.798	0.726	0.805
6,000	0.726	0.805	0.721	0.801
7,000	0.721	0.801	0.721	0.801
8,000	0.721	0.801	0.72	0.8
9,000	0.721	0.801	0.719	0.798
10,000	0.721	0.801	0.722	0.8
20,000	0.718	0.798	0.72	0.801
cijeli	0.723	0.798	0.723	0.798

se to svojstvo pojavilo jedanput, odnosno pridružuje mu se minimalna mjera iz popisa mjera referentnog korpusa.

Drugi je algoritam jednostavniji iz razloga što nije potrebno paziti da su dokumenti koje je potrebno analizirati dio referentnog korpusa. To sa sobom nosi i činjenicu da je, u slučaju da se pokaže da nije potrebno imati što veći mogući referentni korpus, moguće jednokratno izračunati mjere na referentnom korpusu te se njih ne mora nanovo računati.

Rezultati eksperimenata na uzorku 5-SVI prikazani su u tablici 4.27.

Iz podataka je vidljivo da uz upotrebu prvog algoritma već korpus od 1,000 dokumenata daje rezultate vrlo bliske višestruko većem korpusu. U slučaju korištenja drugog algoritma pri manjim korpusima, odnosno kada je češći slučaj da ne postoji mjera referentnog korpusa za neko svojstvo te se koristi maksimalna vrijednost, moguće je primijetiti određeno poboljšanje rezultata. Moguća argumentacija tih rezultata jest da se time sve pojavnice koje nisu prečeste jednako nadograđuje, odnosno ne čini posebna razlika između pojavnica koje su vrlo rijetke, odnosno razmjerno rijetke. Kako se u ovim eksperimentima koristi mjera referentnog korpusa IDF, zaključak ovih

Tablica 4.28: Utjecaj veličine referentnog korpusa (VRK) na evaluacijske mjere F_1 i $F_{0.5}$ na uzorku 5-SVI

VRK	F_1	$F_{0.5}$
10	0.717	0.769
20	0.722	0.786
30	0.719	0.787
40	0.719	0.787
50	0.722	0.789
100	0.719	0.791
200	0.714	0.791
300	0.718	0.792
500	0.723	0.8
700	0.736	0.805
900	0.741	0.81
1,000	0.745	0.811

rezultata pokazuje iz još jednog kuta manjkavosti te vrlo popularne i moćne mjere.

Između ova dva algoritma očiti je pobjednik drugi, i to iz dva razloga:

- drugi algoritam postiže bolje rezultate i od prvog algoritma i od pristupa gdje se cijeli korpus koristi za računanje mjera referentnog korpusa
- ovim se algoritmom uz značajno smanjenje količine podataka potrebnih za računanje mjera referentnog korpusa ostvaruje dodatna velika prednost - izbor podataka te računanje mjera nad odabranim podacima moguće je učiniti jednokratno te do daljnjega koristiti izračunate mjere.

Kako je već u prvom eksperimentu vrlo mala količina podataka pokazala vrlo dobre rezultate, sljedeći će eksperiment istražiti uključivanje od samo 10 do 1000 dokumenata u referentni korpus. Rezultati ovog eksperimenta prikazani su u tablici 4.28.

Na rezultatima je moguće primijetiti da je već 30 dokumenata dovoljna količina podataka za dobru procjenu mjere referentnog korpusa. S 500 do-

Tablica 4.29: Utjecaj veličine referentnog korpusa (VRK) na evaluacijske mjere F_1 i $F_{0.5}$ na uzorcima 4-SVI, 5-SVI i 6-SVI

	4-SVI		5-SVI		6-SVI	
VRK	F_1	$F_{0.5}$	F_1	$F_{0.5}$	F_1	$F_{0.5}$
1,000	0.763	0.825	0.745	0.81	0.68	0.746
2,000	0.763	0.825	0.747	0.817	0.685	0.759
3,000	0.755	0.82	0.729	0.805	0.675	0.754
cijeli	0.774	0.832	0.723	0.798	0.68	0.754

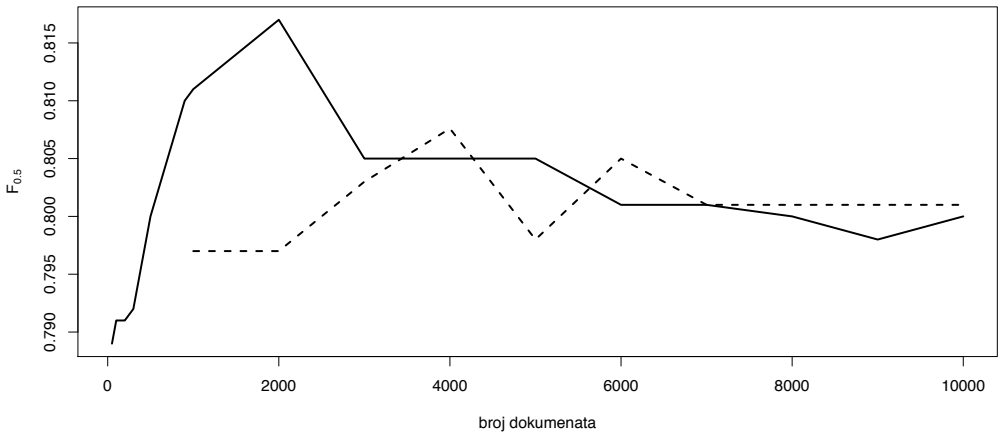
kumenata procjena postaje zaista dobra. Unatoč tome, 1,000 dokumenata ipak daje bolje rezultate.

Kako bi se pronašla dobra procjena optimalne veličine referentnog korpusa izvršen je i treći eksperiment, i to nad sva tri uzorka. Rezultati tog eksperimenta prikazani su u tablici 4.29.

Zaključno je na slici 4.1 prikazan odnos $F_{0.5}$ mjere i veličine referentnog korpusa te načina oblikovanja referentnog korpusa. Iscrtkana krivulja predstavlja rezultat prvog algoritma koji zahtijeva prisutnost dokumenata koji se analiziraju dok puna krivulja predstavlja rezultat drugog gdje analizirani dokumenti nisu dio referentnog korpusa. I iz ovog je prikaza vidljivo da je optimalna veličina referentnog korpusa oko 2,000 dokumenata. Isto je tako vidljivo da drugi, jednostavniji algoritam redovito pobjeđuje kompleksniji. Usto je moguće primijetiti trbuh na početku pune krivulje gdje ta krivulja značajno odskače od iscrtkane. Naknadno se krivulje izjednačuju. Argumentacija tih rezultata ide u smjeru nepokrivanja velikog broja specifičnijih svojstava od strane referentnog korpusa što za posljedicu ima da se tim svojstvima dodjeljuje maksimalna vrijednost iz referentnog korpusa. Ovaj zaključak predstavlja dodatnu kritiku u ovom slučaju korištene TF-IDF mjere koja s većim i informativnijim referentnim korpusom specifičnija svojstva više ne nagrađuje dovoljno.

Zaključak koji polučuju prikazani rezultati jest da je optimalna veličina referentnog korpusa otprilike 2,000 dokumenata što odgovara veličini od cca. pola milijuna pojavnica, i to ne nužno nad dokumentima koje se trenut-

Slika 4.1: Odnos $F_{0.5}$ mjere i veličine referentnog korpusa s obzirom na način oblikovanja referentnog korpusa



no analizira. To omogućuje jednokratnu analizu manjeg korpusa te daljnje korištenje jednokratno izračunatih mjera. Činjenica da za TF-iDF mjeru u ovom zadatku vrijedi "manje je bolje" ukazuje na dodatnu nesavršenost popularne TF-IDF mjere.

Ovim je skupom eksperimenata odgovoreno na pitanje koliko referentni korpus treba sadržavati dokumenata no i ne koliko ga često treba osvježavati. Na to bi pitanje odgovore trebali dati dodatni eksperimenti.

Poglavlje 5

Zaključak

U ovom je doktorskom radu empirijski provjeren, odnosno optimiziran niz metoda za koje je moguće pretpostaviti da mogu uspješno riješiti zadatak pronalaženja događaja u višestrukim izvorima informacija. Zadatak je, dakle pri više izvora informacija koji višestruko izvještavaju o događajima identificirati dokumente u kojima se izvještava o istom događaju. Oblikovanjem samih grupa zvanih grozdovi vrši se i pronalaženje određenih događaja. Jedna od pretpostavki koja pojednostavnjuje zadatak je ta da se u jednom dokumentu izvještava o jednom događaju, odnosno da se dokument može smatrati prikazom jediničnog događaja.

Osnovne metode koje se primjenjuju u ovom radu su formalni prikaz dokumenta vektorom svojstava, korištenje mjera udaljenosti između vektora kako bi se izračunale udaljenosti u sadržaju dokumenta, odnosno sličnost dokumenata. Na temelju tih sličnosti algoritmom grožđenja se dokumenti grupiraju u grozdove te se pokušava postići takav poredak koji bi odgovarao onome oblikovanom od strane ljudkog označitelja - stručnjaka koji je unaprijed dokumente grupirao prema događaju o kojem izvještavaju.

Kako bi se istražila kompleksnost zadatka koji obavlja ljudski označitelj, nezavisno su označena dva uzorka te je izračunata mjera dogovora između označitelja, i to na dva načina. Mjera koja je osjetljiva na veliku razliku u broju elemenata u skupovima, odnosno različiti broj grozdova koji se uspoređuju daje vrijednost od 0.648, dok vrijednost koja razliku u broju elemenata za-

nemaruje, odnosno nije osjetljiva na različit broj oblikovanih grozdova daje strop od 0.91. Obje su vrijednosti relevantne - druga vrijednost je vjerojatno bliže stvarnom stropu zadatka, odnosno ne može se očekivati od algoritma da evaluacijskom mjerom prijeđe ovu vrijednost. Prva pak vrijednost ukazuje na kompleksnost zadatka, odnosno činjenicu da se rezultat dvaju ljudskih označitelja koji su dobili identične upute slaže samo u 65% slučajeva.

Kako bi se različite metode mogle usporediti potrebna je mjera sličnosti rezultata stvorenog od strane čovjeka, odnosno algoritma. Tome služe evaluacijske metode. U ovom je doktorskom radu korišteno sedam različitih evaluacijskih metoda.

Teorijski je dokazano kako čistoća kao evaluacijska metoda nije poželjna mjera za ovakav zadatak. Ona, naime, ne kažnjava pretjerano dijeljenje grozdova, odnosno smatra optimalnim rezultatom i onaj gdje je svaki dokument u svom grozdu.

Normalizirana međusobna informacija i rand indeks nemaju navedeni problem, oni, naime, preferiraju rezultate s manjim brojem grozdova. Njima se može zamjeriti činjenica da redovito daju vrlo visoke vrijednosti te da time ne diskriminiraju dovoljno dobro među rezultatima (standardna devijacija rezultata je vrlo mala).

Evaluacijske mjere koje pokazuju najbolje rezultate su klasične mjere preciznosti, potpunosti i F mjere. Preciznost i potpunost pokazuju standardnu manjkavost - svoju suprostavljenost. Upravo taj problem rješava se F mjerama. Parametri β F mjera koji se istražuju su 1 i 0.5. Općenito obje mjere daju uvid u uspješnost neke metode rješavanja danog zadatka, no prednost se mora dati $F_{0.5}$ mjeri iz razloga što ona daje prednost preciznosti. Naime, priroda zadatka nalaže da se rezultat što je točniji no manje potpun smatra boljim od onoga koji je potpuniji, no manje točan.

Zaključno se o evaluacijskim mjerama može reći da sve osim čistoće daju upotrebljive rezultate te da su najinformativnije preciznost i potpunost, odnosno iz njih izračunate F mjere. Mjera $F_{0.5}$ se smatra najkorisnijom zbog prirode zadatka čija se rješenja evaluiraju.

Nadalje je eksperimentirano s algoritmima grožđenja. Dokazano je kako je za zadatak pronalaženja događaja algoritam grožđenja jednim prolaskom

jednako uspješan kao i složeniji hijerarhijski algoritmi za grožđenje. Razlog tomu leži u prirodi rješenja koje dokument prikazuje kao vektor svojstava - pojavnica izvan njihovog konteksta te kompleksnost hijerarhijskih algoritama ovdje ne dolazi do izražaja.

Kod mjera udaljenosti najbolje je rezultate pokazala kosinusna mjera te odmah u nastavku i *Jaccard* i *Dice* koeficijenti te *Jensen-Shannon*. Potonji dosljedno pokazuju nešto slabije rezultate. *Manhattan* udaljenost pokazuje bitno slabije rezultate, dok je najlošija mjera udaljenosti zasigurno Euklidova udaljenost koja se još jednom potvrdila kao loša mjera u slučaju postojanja stršećih vrijednosti, odnosno velikih razlika u vrijednostima pojedinih dimenzija.

Dvije pretpostavljene heuristike - dokumenti koji opisuju isti događaj objavljuju se u istom danu i ne postoje dva dokumenta iz istog izvora koji opisuju isti događaj - su dokazane i *in vitro* - na označenim podacima - i *in vivo* - na zadatku pronalaženja događaja. Obje su heuristike, naime, uspješno dokazane analizom podataka, a u praksi, pri rješavanju problema pronalaženja događaja, kao što je i očekivano, pospješuju potpunost, a smanjuju preciznost, no i dalje pozitivno utječu na evaluacijsku mjeru $F_{0.5}$. Nadalje, prva heuristika čini cijeli problem bitno jednostavnije izračunljivim. Za pretpostaviti bi bilo da bi se prva heuristika primjenjivala i da kvari rezultat zato što njeno neprimjenjivanje znatno otežava izračun zbog kombinatorne eksplozije broja parova dokumenata koje treba uspoređivati. Primjena druge heuristike nešto otežava izračun no bitno manje nego što pospješuje rezultat.

Među mjerama težine svojstva kao najbolja se mjera dokazala i najpopularnija TF-IDF mjera postižući redovito bolje rezultate od t-testa, druge po redu optimalne vrijednosti varijable MTS. Uvjetna vjerojatnost - nelogaritmirani t-test - nešto zaostaje za t-testom dok je očiti gubitnik ovog eksperimenta vjerojatnost - jedina mjera koja ne uzima u obzir razdiobu svojstava u cijelom korpusu s obzirom na razdiobu svojstava u pojedinom dokumentu.

Što se tiče određivanja svojstava na razini pojavnica, interpunkcije su se pokazale kao štetne. Za pretpostaviti bi bilo da one nemaju distinktivnu razliku te da ne bi smjele utjecati na rezultat, no očito je uključivanjem samo pojavnica koje odgovaraju pravilu pojavnice dodatno očišćen uzorak.

Veličina slova se ispostavila kao štetan podatak. Korištenjem jednostavne statističke metode određivanja veličine slova rezultat je bolji no kad se pojavnice ostavljaju onakve kakve jesu, no i dalje najbolje rezultate postižu pojavnice pretvorene u mala slova.

Jednostavna metoda određivanja važnosti naslova za dokument ponavljanjem pojavnica iz naslova određeni broj puta pokazala se najuspješnijom kada se pojavnice ponove četiri puta.

Što se tiče odabira svojstava na razini pojavnica, isključivanje hapax legomena, dakle svojstava koja se pojavljuju samo jednom, dva, odnosno tri puta, pomaže zadatku te usto znatno smanjuje broj svojstava. Zaključeno je kako je optimalno ne uključiti svojstva koja se pojavljuju jednom, dakle prave hapax legomena.

Isključivanje funkcijskih riječi, odnosno onih svojstava koja imaju najmanju vrijednost logaritma mjere IDF, općenito popravlja rezultat. Najveći pomak se može primijetiti na više uzoraka kad se isključe svojstva na popisu u rasponu od 100 do 150 što navodi na zaključak kako IDF mjera nije optimalna za računanje "nevažnosti" svojstva. Što se tiče pojednostavnjenja prikaza dokumenta, odnosno smanjenja broja svojstava, ova metoda ostvaruje zanemarivo pojednostavnjenje.

Jednostavniji oblik morfološke normalizacije - korjenovanje - se pokazalo u slučaju jednostavnog korjenovanja štetnim, dok se u slučaju kompleksnijeg, jezično ovisnog korjenovanja pokazalo niti korisnim niti štetnim.

Primijećuje se, kao i do sada, trend da se dodatne, kompleksnije metode pozitivno odražavaju na teže probleme, dok kod lakših problema rijeđe pokazuju napredak ili pak prouzročuju nazadak.

Kompleksnija i jezično bitno točnija metoda morfosintaktičkog označavanja također generalno ne pokazuje značajan napredak u rješavanju zadatka. Uzevši u obzir njenu kompleksnost i cijenu razvoja, za ovu razinu prikaza dokumenta i ovaj zadatak ona nikako nije za preporučiti.

Najveći razlog neuspjehu morfološke normalizacije je priroda zadatka - pronalaženje dokumenata koji izvještavaju o istom događaju. Naime, za pretpostaviti je da će autori opisujući neki događaj uz iste lekseme te lekseme koristiti i u istoj morfosintaktičkoj kategoriji.

Uključivanje višečlanih izraza u popis svojstava doživljava neuspjeh sličan onome kao kod morfološke normalizacije. U popis svojstava je postepeno uvođeno sve više i više najjačih kolokacija izračunatih hi-kvadrat testom uz prag čestote svakog elementa kolokacije od tri pojavljivanja te nikakav napredak nije primijećen. Uzevši u obzir neuspješnost prethodnih statističkih i jezičnih metoda kao i ove statističke, moguće je zaključiti kako ovakav pristup analizi, odnosno modeliranju sadržaja, tj. značenja dokumenta, posebno za zadatak pronalaženja dokumenata koji izvještavaju o istom događaju, ne profitira od dubljih jezičnih analiza te najbolje rezultate postiže koristeći što jednostavnije statističke mjere. Za očekivati je kako bi ovakve metode kao i neke kompleksnije poput izgradnje sintaktičkih stabala ili analize semantike ili pragratike mogle imati pozitivan utjecaj, no samo uz promjenu, odnosno nadogradnju osnovne paradigme vektorskog prostora i algoritama za grožđenje.

Što se tiče utjecaja veličine referentnog korpusa na uspješnost rješavanja zadataka, zaključeno je kako "više je bolje" u ovom slučaju ne vrijedi te da je optimalna veličina referentnog korpusa 2,000 dokumenata. Istražujući je li bolje posjedovati mjeru iz referentnog korpusa za svako svojstvo u dokumentu koji se analizira, zaključeno je kako to nije potrebno. Time je postignuto znatno pojednostavljenje - moguće je jednokratko oblikovati referentni korpus te izvršiti na njemu potrebna mjerenja. Te je mjere nadalje moguće uvijek direktno pozivati, a sve nepostojeće vrijednosti se, dokazano je, mogu poistovjetiti s onima hapax legomena u referentnom korpusu.

Vrijeme trajanja takvog izmjerenog referentnog korpusa dalje je nepoznato. Za njegovo određivanje trebalo bi provesti vremenska istraživanja koja zasigurno nadilaze spektar interesa ovog doktorskog rada.

Općenito je moguće zaključiti kako je vektorski model prikaza dokumenata te grupiranje podatkovnih točaka jednostavnim algoritmom grožđenja vrlo uspješna metoda pronalaženja pojedinih događaja u skupu dokumenata. Statističke metode koje uzimaju u obzir čestotnu razdiobu svojstava u referentnom korpusu značajno pospješuju rezultat. Referentni korpus ne treba biti izrazito velik, odnosno ne vrijedi "što više, to bolje". Naprednije statističke i lingvističke jezične metode u pravilu ne pomažu zadatku pronalaženja

događaja. Daljnje istraživanje ovog područja mora težiti promjeni paradigme prikaza dokumenta. Ovakav je prikaz, naime, očito ostvario svoj vrhunac te ne profitira uvođenjem svojstava određenih i odabranih nekim složenijim metodama.

Poglavlje 6

Dodaci

6.1 Englesko-hrvatski glosar manje poznatih stručnih termina

accuracy - točnost
anaphora resolution - razrješavanje anafore
average-link clustering - grožđenje prosječnom vezom
bag of lemmata - vreća lema
bag of words - vreća riječi
baseline - osnovni pristup
bottom-up - od dna prema gore
ceiling - strop
chunking - razdjeljivanje
cluster - grozd
clustering - grožđenje
complete-link clustering - grožđenje potpunom vezom
computational linguistics - računalna lingvistika
conditional probability - uvjetna vjerojatnost
contingency table - tablica slučajeva
cosine measure - kosinusna mjera
cost - trošak
cost function - funkcija troška

cut-off point - točka rezanja
data stream - podatkovni tok
divergence - odstupanje
external quality criterion - vanjski kriterij kvalitete
false negative - lažno negativno rješenje
false positive - lažno pozitivno rješenje
feature - svojstvo
feature extraction - određivanje svojstava
feature selection - odabir svojstava
feature space - prostor svojstava
feature weight - težina svojstva
gold standard - zlatni standard
hierarchical clustering - hijerarhijsko grožđenje
human annotator - ljudski označitelj
inter-annotator agreement - dogovor između označitelja
internal quality criterion - unutarnji kriterij kvalitete
joint probability - zajednička vjerojatnost
knowledge management - upravljanje znanjem
lemmatization - lematizacija
lower bound - donja granica
machine learning - strojno učenje
maximum likelihood estimate - procjena najveće vjerojatnosti
named entity recognition - prepoznavanje imena entiteta
natural language processing (NLP) - obrada prirodnog jezika (OPJ)
normalized mutual information - normalizirana međusobna informacija
objective function - objektivna funkcija
outlier - stršeći podatak
parsing - raščlamba
penalizing - kažnjavanje
phrase - sveza
POS tagging - POS označavanje, označavanje vrste riječi, morfosintaktičko označavanje
precision - preciznost

pruning - podrezivanje
purity - čistoća
rand index - rand indeks, točnost
recall - potpunost
single-link clustering - grožđenje pojedinačnom vezom
single-pass clustering - grožđenje jednim prolaskom
smoothing - izravnjavanje
speech-to-text - govor u tekst
stopping, stop-word removal - uklanjanje stop riječi
supervised machine learning - nadzirano strojno učenje
text-to-speech - tekst u govor
threshold - prag
top-down - od vrha prema dolje
trade-off - kompromis
true negative - istinito negativno rješenje
true positive - istinito pozitivno rješenje
upper bound - gornja granica
unsupervised machine learning - nenadzirano strojno učenje
weight - težinski faktor
weighting - težinsko faktoriranje
word sense disambiguation - razrješavanje leksičke višeznačnosti

6.2 Primjer rezultata krajnjeg algoritma nad uzorkom 5-SVI

Prikazani su grozdovi koji sadrže više od jednog dokumenta. Navedeni su naslov i izvor dokumenta.

—

Bolivija: prošao referendum za veću autonomiju Santa Cruza / totalportal.hr
Bogati napuštaju Moralesa / glas-slavonije.hr
Najbogatija bolivijska pokrajina izglasala veću autonomiju / business.hr

Prošao referendum za veću autonomiju Santa Cruza / tportal.hr
Bolivija: Prošao referendum za veću autonomiju Santa Cruza / vecernji.hr
Bolivijska najbogatija pokrajina želi autonomiju / javno.com
Hrvat vodi "pobunu" protivv Moralesa / jutarnji.hr
Najbogatija bolivijska provincija proglasila autonomiju / rtl.hr
Morales: referendum je promašena separatistička mjera / seebiz.eu
Bolivija: Prošao referendum za veću autonomiju Santa Cruza / dnevnik.hr
Najveća pokrajina Bolivije na referendumu izabrala autonomiju / index.hr
—
Obama vratio veću potporu među demokratskim biračima / jutarnji.hr
Ankete: Obama vratio veću potporu među demokratskih biračima / vecernji.hr
Obama hita prema sve većoj potpori demokrata / javno.com
Obama vratio veću potporu među demokratskih biračima / tportal.hr
Obama povećao prednost pred Hillary Clinton / totalportal.hr
Obama vratio veću potporu među demokratskih biračima / seebiz.eu
Obama ponovo u anketama pobjeđuje Clinton / business.hr
Obama vratio veću potporu među demokratskih biračima - ankete / dnevnik.hr
Obama vratio veću potporu među demokratskim biračima / nacional.hr
Obama vratio potporu glasača i povećao prednost pred Clinton / index.hr
—
Borisu Tadiću prijete metkom u čelo zbog izdaje Srbije / index.hr
Tadiću prijete ubojstvom zbog izdaje / javno.com
Prijetnje smrću srbijanskom predsjedniku Borisu Tadiću / totalportal.hr
Borisu Tadiću prijete 'metkom u čelo" / dnevnik.hr
Borisu Tadiću prijete smrću / nacional.hr
Predsjedniku Srbije Tadiću prijete "metkom u čelo" / business.hr
Borisu Tadiću prijete metkom u čelo / tportal.hr
Borisu Tadiću prijete smrću / rtl.hr
Borisu Tadiću prijete smrću / mojportal.hr
Borisu Tadiću prijete metkom u čelo / jutarnji.hr
—

Prosvjed 150 pirotehničara na Trgu bana Jelačića / business.hr
Pirotehničari prosvjeduju zbog radne snage iz BiH / totalportal.hr
Prosvjed pirotehničara na Trgu bana Jelačića / dnevnik.hr
Prosvjed pirotehničara na Trgu bana Jelačića / vecernji.hr
VIDEO: Prosvjed pirotehničara na Jelačićevom trgu / javno.com
Pirotehničari na rubu egzistencije / tportal.hr
Pirotehničari prosvjedovali na Trgu bana Jelačića protivv uvoza stranih radnika / index.hr
Prosvjed pirotehničara na Trgu bana Jelačića / mojportal.hr
Prosvjed pirotehničara na Trgu bana Jelačića / jutarnji.hr
Prosvjed piroteničara / vijesti.hrt.hr

—

Dr. Ognjen Šimić pred sutkinjom Ikom Šarić / glas-slavonije.hr
Dr. Ognjen Šimić ne osjeća se krivim za primanje mita i pranje novca / vecernji.hr
Šimić pred riječkim sudom: Nisam kriv! / totalportal.hr
Šimić: Ne smatram se krivim / nacional.hr
Počelo suđenje kirurgu Šimiću za primanje mita / business.hr
Šimićev kolega: Nudio mi je da dijelimo novac / jutarnji.hr
Rijeka: Počelo suđenje kirurgu Šimiću za primanje mita / dnevnik.hr
Riječki kirurg Ognjen Šimić: Nisam kriv / index.hr
Kirurg Šimić ne osjeća se krivim za primanje mita / javno.com

—

Brazil: potonuo brod s 80-tak putnika; najmanje 15 mrtvih / totalportal.hr
Brazil: Potonuo brod s 80-tak putnika; najmanje 15 mrtvih / mojportal.hr
U Amazoniji potonuo brod koji je prevozio putnike na zabavu, 15 poginulih / index.hr
Potonuo brod s 80 putnika; najmanje 15 mrtvih / dnevnik.hr
U brodolomu u Amazoniji najmanje 15 mrtvih / tportal.hr
U brodolomu na Amazoni najmanje 15 poginulih / business.hr
Brazil: potonuo brod s 80-ak putnika; najmanje 15 mrtvih / jutarnji.hr
Najmanje 15 mrtvih prilikom potonuća broda / nacional.hr
U Brazilu potonuo brod sa 80 putnika / javno.com

—

Prema novim podacima, u Mianmaru oko 4.000 poginulih / jutarnji.hr
Mianmar: Ciklon ubio 10.000 ljudi / rtl.hr
Ciklon odnio 3969 žrtava u Mianmaru / business.hr
Mianmar: ciklon Nargis odnio 3.969 života / totalportal.hr
U Mianmaru 4,000 mrtvih, 3,000 nestalih / dnevnik.hr
Snažan ciklon do sada usmrtio deset tisuća ljudi u Mianmaru / index.hr
Ciklon u Mianmaru odnio 3.969 života / javno.com
U ciklonu Naris poginulo 3.969 osoba / nacional.hr
Ciklon usmrtio četiri tisuće ljudi / tportal.hr

—

Tijela triju beba pronađena u frižideru / rtl.hr
Tijela troje djece pronađena u zamrzivaču / javno.com
Još jedan horor podrum: Njemica ubila troje vlastite novorođenčadi / index.hr
Tijela 3 bebe otkrivena u zamrzivaču u Njemačkoj / tportal.hr
Išao jesti pa u zamrzivaču pronašao tijela troje novorođenčadi / dnevnik.hr
U Njemačkoj u zamrzivaču pronađena tijela triju beba / jutarnji.hr
Njemačka: Tri dječja tijela nađena u zamrzivaču obiteljske kuće / mojportal.hr
U zamrzivaču pronađena tijela triju beba / nacional.hr
U podrumu kuće u Njemačkoj policija je pronašla tri mrtve bebe, za koje se ispostavilo da ih je ubila vlastita majka, a pronašao ih je njezin sin kada je išao po nešto za jelo. / net.hr

—

Danas počeo "bijeli štrajk" u T-HT-u / liderpress.hr
Bijeli štrajk u T-HT-u / vijesti.hrt.hr
Radnici T-HT-a započeli bijeli štrajk / nacional.hr
Bijeli štrajk prvih 2000 zaposlenika T-HT-a / totalportal.hr
"Bijeli štrajk" u T-HT-u / mojportal.hr
Počeo "bijeli štrajk" u T-HT-u / vecernji.hr
Danas počeo "bijeli štrajk" u T-HT-u / poslovni.hr
Počeo "bijeli štrajk" u HT-u koji će usporiti pružanje usluga korisnicima /

business.hr

—

SDP u saborsku proceduru uputio interpelaciju vezanu uz slučaj Pukanić / vecernji.hr

SDP će u Sabor uputiti interpelaciju u slučaju Mirjane Pukanić / totalportal.hr

Policija je zataškala prijetnje Mirjani Pukanić / jutarnji.hr

SDP: Policija je ponižavajućim scenama privođenja pred novinarima narušila prava Mirjane Pukanić / index.hr

Slučaj Mirjane Pukanić u saborskim klupama / tportal.hr

SDP je u saborsku proceduru uputio interpelaciju kojom se od Vlade traži kažnjavanje odgovornih u slučaju prisilne hospitalizacije Mirjane Pukanić, zbog kršenja njenih prava. / net.hr

SDP u saborsku proceduru uputio interpelaciju vezanu uz slučaj Pukanić / dnevnik.hr

"Ministar zdravstva postupanjem u slučaju Pukanić narušio povjerenje u vladavinu prava" / business.hr

—

Kovačević razočaran / vijesti.hrt.hr

Kovačević razočaran jer nema sučeljavanja pred SDP-ovim članstvom / mojportal.hr

Dragan Kovačević protivv sučeljavanja u Otvorenom / index.hr

Kovačević razočaran jer je Milanović izbjegao sučeljavanje pred članstvom SDP-a / business.hr

Kovačević: Gdje je nestao čovjek? / javno.com

Kovačević: Gdje je nestao čovjek? / seebiz.eu

Kovačević: "U SDP-u nije zaživjela demokratska praksa javnog sučeljavanja" / vecernji.hr

Večeras u Otvorenom "javno sučeljavanje" kandidata SDP-a / nacional.hr

—

Pucano na kćerku i unuka vlasnika Međimurjepleta, čakovečku HDZ-ovku / business.hr

Pucano na Nikolinu Babić. potpredsjednicu čakovečkog HDZ-a / nacional.hr

Metak je okrznuo rukav čakovečke HDZ-ovke / javno.com
Pucano na Nikolinu Babić, direktoricu Pane i potpredsjednicu čakovečkog HDZ-a / vecernji.hr
Pucali na potpredsjednicu čakovečkog HDZ-a / jutarnji.hr
Policija traga za osobama koje su pucale na Nikolinu Babić / mojportal.hr
U napadu na potpredsjednicu čakovečkog HDZ-a sudjelovale četiri osobe / dnevnik.hr
Potpresjednica čakovečkog HDZ -a dva puta izbjegla smrt / index.hr

—

Arsenal nudi 12 milijuna funti za Niku Kranjčara? / rtl.hr
Arsenal nudi 12 milijuna funti za Niku Kranjčara / javno.com
Arsenal Portsmouthu za Kranjčara nudi 12 milijuna funti / business.hr
Arsenal nudi 12 milijuna funti za Niku Kranjčara / liderpress.hr
Arsenal za Kranjčara nudi 12 milijuna funti / nacional.hr
Arsenal nudi 12 milijuna funti za Niku Kranjčara? / dnevnik.hr
Daily Mirror: Arsenal nudi 12 milijuna funti za Niku Kranjčara / mojportal.hr

—

Protiv bivših dužnosnika Osječke pivovare podignuta optužnica / index.hr
Podignuta optužnica protivv odgovornih dužnosnika osječke Pivovare / dnevnik.hr
Osijek: podignuta optužnica protivv dužnosnika osječke pivovare / totalportal.hr
Vrh osječke Pivovare optužen za krivotvorenje te zlouporabu položaja / business.hr
Podignuta optužnica protivv odgovornih dužnosnika osječke Pivovare / mojportal.hr
Podignuta optužnica protivv odgovornih dužnosnika osječke Pivovare / vecernji.hr
Podignuta optužnica protivv odgovornih dužnosnika osječke Pivovare / liderpress.hr

—

Dugoselski poduzetnik od Eduarda traži devet milijuna kuna / business.hr

Eduardu tužba od 8,7 milijuna kuna / mojportal.hr
Da Silva prijavio bivšeg savjetnika zbog prijevare / javno.com
Radivojević tuži Eduarda Da Silvu za 8,7 milijuna / nacional.hr
Tužba protivv Dudua: "Savjetnik" traži 8,7 milijuna kuna / totalportal.hr
"Jedino što su oni napravili jest da su Eduardu jednom kupili tenisice" / dnevnik.hr
Radivojević tužbu temelji na ugovoru o poslovnoj suradnji u kojem stoji da mu je, kao osobnom savjetniku, Dudu dužan isplatiti 10 posto provizije na bilo koji iznos uplaćen na njegovo ime do 2010. / net.hr

—

Primorac saslušao studente / vijesti.hrt.hr
Primorac pokušava izbjeći studentske prosvjede razgovorom s Rektorskim zborom / business.hr
Studenti ne odustaju od prosvjeda zbog Bolonje / jutarnji.hr
Studenti ne odustaju od najavljenog prosvjeda / dnevnik.hr
Studenti nakon sastanka s Primorcem ipak ne odustaju od prosvjeda / index.hr
Nema odustajanja od prosvjeda / nacional.hr
Primorac primio organizatore studentskog štrajka / javno.com

—

Slaven Bilić potpisao novi ugovor s HNS-om / seebiz.eu
Bilić produžio ugovor i objavio putnike na Euro / nacional.hr
Bilić na Euro vodi Sharbinija i Pamića / dnevnik.hr
Bilić potpisao novi dvogodišnji ugovor / business.hr
Bilić pozvao Pokrivača i Sharbinija, Mandžukić "izvisio" / totalportal.hr
Bilićev popis za Euro: Ušli Sharbini i Pamić / rtl.hr
Bilić: Sharbini i Pamić idu na priprema za Euro 2008. / mojportal.hr

—

Nisam znao da su transvestiti / javno.com
Ronaldo: Učinio sam nešto jako glupo / mojportal.hr
Ronaldo: Učinio sam nešto jako glupo / nacional.hr
Ronaldo: "Učinio sam nešto jako glupo" / rtl.hr
Ronaldo: Učinio sam nešto jako glupo / vecernji.hr

Ronaldo: Učinio sam nešto jako glupo / dnevnik.hr

—

Britanski umjetnik provest će dva dana zakopan u pijesku / totalportal.hr
Umjetnik dva dana zakopan u pijesku / tportal.hr
Britanski umjetnik provest će dva dana zakopan u pijesku / business.hr
Zakopao se u pijesak kako bi spasio plažu / javno.com
Britanski umjetnik zatrpan u pijesku / nacional.hr
Britanski umjetnik provest će dva dana zakopan u pijesku / mojportal.hr

—

Jedna osoba poginula u eksploziji plina u Sarajevu / totalportal.hr
Eksplozija plina u centru Sarajeva; jedna osoba poginula / mojportal.hr
Poginula žena u eksploziji plina u Sarajevu / tportal.hr
Jedna osoba poginula, a troje ozlijeđeno u eksploziji plina u centru Sarajeva / index.hr
Sarajevo: Jedna osoba poginula, troje ozlijeđeno u eksploziji plina / dnevnik.hr
Jedna osoba poginula, troje ozlijeđeno u eksploziji plina u Sarajevu / vecernji.hr

—

Dalj: Peticijom protivv ulice srpskog znanstvenika / javno.com
Dalj: Mještani protivv preimenovanja ulice poginulog branitelja Kemenjija / dnevnik.hr
Daljani ne žele ulicu Milutina Milankovića / tportal.hr
Srbi u Dalju protivv ulice s imenom poginulog branitelja / jutarnji.hr
Dalj: Mještani protivv preimenovanja ulice / mojportal.hr
Peticijom protivv SDSS-a koji želi mijenjati ime ulice poginulog hrvatskog policajca / index.hr

—

U Rijeci počinje suđenje kirurgu zbog primanja mita / mojportal.hr
Liječniku Šimiću počinje suđenje za primanje mita / totalportal.hr
Osumnjičen za mito i pranje 2,5 milijuna kuna / jutarnji.hr
Počinje suđenje kirurgu / vijesti.hrt.hr
Počinje suđenje riječkom kirurgu Ognjenu Šimiću / index.hr

Rijeka: Suđenje zbog mita / rtl.hr

—

Sanader: Poslodavci i sindikati za dogovor o minimalnoj plaći imaju samo tjedan dana / business.hr
Sanader: Vlada ima rješenje za minimalac / totalportal.hr
Sanader: Ako ne bude dogovora o minimalnoj plaći, odlučit će Vlada / jutarnji.hr
Vlada će odrediti minimalac? / nacional.hr
Sanader o socijalnom partnerstvu i bipartizmu / poslovni.hr
Sanader: Jačati socijalno partnerstvo i bipartizam / liderpress.hr

—

Preminuo bivši potpredsjednik zagrebačke Gradske skupštine Dragutin Štiglić / index.hr
Preminuo potpredsjednik zagrebačke Gradske skupštine Dragutin Štiglić / mojportal.hr
Preminuo potpredsjednik zagrebačke Gradske skupštine Dragutin Štiglić / business.hr
Preminuo gradski zastupnik Dragutin Štiglić / javno.com
Preminuo potpredsjednik zagrebačke Gradske skupštine / tportal.hr

—

Srbija ne odustaje od potrage za ratnim zločincima / javno.com
"Srbija ne odustaje od potrage za ratnim zločincima" / totalportal.hr
Srbija vjeruje da će potraga za Karadžićem i Mladićem uroditi plodom / business.hr
Vukčević: Srbija ne odustaje od potrage za ratnim zločincima / dnevnik.hr
Vukčević: Srbija ustrajna u potrazi za ratnim zločincima / nacional.hr

—

Audi želi proizvoditi električne automobile / jutarnji.hr
Audi će ponuditi električni automobil za 5 do 10 godina / business.hr
Audi namjerava proizvoditi električne automobile / poslovni.hr
Audi izbacuje električni automobil za 5 do 10 godina / totalportal.hr
Audi će ponuditi električni automobil za 5 do 10 godina / liderpress.hr

—

Ernest Rađen izjavio da se ne osjeća krivim / dnevnik.hr
Optuženi za zločine u Škabrnji ne osjeća se krivim / javno.com
Ernest Rađen rekao da se ne osjeća krivim / vecernji.hr
Rađen rekao da se ne osjeća krivim / tportal.hr
Optuženik za ratne zločine u Škabrnji: Nisam kriv / index.hr

—

Njemački političar možda će prodati bradu / totalportal.hr
Milijun eura za bradu njemačkog političara / tportal.hr
Njemački političar daje bradu u dobrotvorne svrhe / business.hr
Političar razmišlja o prodaji brade / vecernji.hr
Njemački političar možda će prodati bradu / dnevnik.hr

—

Hrvatsku zastavu zapalila dva slovenska vojnika? / index.hr
Slovenci ne vjeruju da je zastavu zapalila vojska / javno.com
Hrvatsku zastavu zapalili slovenski vojnici? / mojportal.hr
Umag: Slovenski vojnici zapalili hrvatsku zastavu? / vecernji.hr
Slovenski vojnici zapalili hrvatsku zastavu / seebiz.eu

—

Štićenici koji budu kršili zabranu pušenja mogu biti izbačeni iz staračkih domova / index.hr
Zagreb: Najavljen rigorozni sigurnosni sustav u umirovljeničkim domovima / dnevnik.hr
Oštrija sankcioniranja u domovima umirovljenika / javno.com
Zabrana pušenja u domovima umirovljenika / mojportal.hr
Umirovljenike će zbog pušenja izbacivati iz domova / nacional.hr

—

Mesić prozvao Vladu / vijesti.hrt.hr
Ministarstvo vanjskih poslova zapošljava podobne / javno.com
"Oni se ne zaustavljaju u VII. upravi, već im je ona ulaz, a onda se odmah premještaju", rekao je predsjednik Mesić i pozvao medije da istraže tko su ti koji ulaze tako u diplomaciju. / net.hr
Mesić: Sedma uprava zapošljava podobne bez natječaja / nacional.hr
Mesić optužio Ministarstvo vanjskih poslova za varanje države / index.hr

—

WHO: smrtonosni virus u Kini nije prijetnja OI / totalportal.hr
Smrtonosni virus neće ugroziti Olimpijske igre / tportal.hr
Smrtonosni virus u Kini nije prijetnja OI / javno.com
WHO: Smrtonosni virus u Kini nije prijetnja OI / dnevnik.hr

—

Alitalia odlazi u ruke Lufthanse? / totalportal.hr
Lufthansa i UniCredit razgovaraju o suradnji na preuzimanju Alitalije / poslovni.hr
Lufthansa i UniCredit preuzimaju Alitaliju? / business.hr
Lufthansa i UniCredit razgovaraju o preuzimanju Alitalije / liderpress.hr

—

Dvojac osumnjičen za atentat na Karzaija / javno.com
Uhićena dvojica zbog umiješanosti u atentat na Karzaija / tportal.hr
Kabul: Uhićena dvojica pod sumnjom za umiješanost u atentat na Karzaija / vecernji.hr
Uhićena još dvojica osumnjičenih za umiješanost u atentat na Karzaija / dnevnik.hr

—

Bačene bombe na OŠ Jurja Dalmatinca u Šibeniku / vecernji.hr
Bačene eksplozivne naprave na školu u Šibeniku / javno.com
Prošle noći bačene eksplozivne naprave na školu u Šibeniku / totalportal.hr
Bačene bombe na školu u Šibeniku / tportal.hr

—

Deutsche Telekom želi preuzeti Sprint? / seebiz.eu
Deutsche Telekom "bacio oko" na američki Sprint Nextel / index.hr
Deutsche Telekom razmišlja o preuzimanju američkog Sprinta / liderpress.hr
Deutsche Telekom preuzima američki Sprint? / business.hr

—

Sindikalna košarica u travnju 0,17 posto skuplja / tportal.hr
Sindikalna košarica u travnju 0,17 posto skuplja nego u ožujku / poslovni.hr
Sindikalna košarica u travnju 0,17 posto skuplja nego u ožujku / vecernji.hr
Životni troškovi četveročlane obitelji 6206 kuna / business.hr

—

”Postavljanje američkog proturaketnog štita u Češkoj velik je korak za europski obrambeni sustav” / index.hr
De Hoop Scheffer: Proturaketni štit je korak prema europskom obrambenom sustavu / vecernji.hr
Proturaketni štit važan za europski obrambeni sustav / nacional.hr
Scheffer: Štitom do europskog obrambenog sustava / javno.com

—

Dino Dvornik se oprašta od ‘Pape‘ / javno.com
Dino Dvornik snima pjesmu za oca / mojportal.hr
Dino Dvornik uglazbit će stihove Jakše Fiamenga i snimiti pjesmu u spomen na oca Borisa / vecernji.hr
I Dino Dvornik se odlučio od svog oca oprostiti stihovima. Ovih će dana krenuti u studio, gdje bi trebao snimiti pjesmu pod nazivom Pape. Tekst za pjesmu piše Jakša Fiamengo. / net.hr

—

ArcelorMittal u pregovorima s Angang Steelom / poslovni.hr
ArcelorMittal i Angang Steel pregovaraju o spajanju / seebiz.eu
Indija i Kina ujedinjene u čeliku / jutarnji.hr
Mittal kupnjom 25 posto Angang Steela ulazi na kinesko tržište / business.hr

—

Razbojnik opljačkao konduktera HŽ-a / rtl.hr
Opljačkao konduktera i odnio torbu s 200 tisuća kuna / dnevnik.hr
Opljačkan kondukter HŽ-a / nacional.hr
Kondukteru HŽ-a ukrao torbu s 200.000 kuna / tportal.hr

—

ZSE: Rast Crobexa, HT najniži ikad / javno.com
Crobex iznad 3.800 bodova / totalportal.hr
ZSE: Crobex ponovno iznad 3800 bodova / business.hr
Crobex porastao 1,14 posto / liderpress.hr

—

Pankretić traži strožu kontrolu vlasnika pasa / business.hr
Ministar Pankretić ponovno razmatra Pravilnik o opasnim psima / index.hr

Pankretić tražio dodatnu provjeru Upisnika pasa / nacional.hr
Ministar Pankretić zatražio je od veterinarske inspekcije provjeru imaju li vlasnici sve potrebne dokumente, te da se napravi evidencija slučajeva ugriza i napada pasa. / net.hr

—

Virovitica: U prometnoj nesreći poginuo mopedist / vecernji.hr
U nesreći kod Virovitice poginuo mopedist / javno.com
U nesreći kod Virovitice poginuo motorist / index.hr
Virovitica: poginuo mopedist / totalportal.hr

—

Nižu se nesreće na snimanju novog nastavka o Jamesu Bondu / mojportal.hr
Ukleti James Bond: Na snimanju nožem izboden tehničar! / dnevnik.hr
Ukleto snimanje: Izboden filmski tehničar koji radi na snimanju novog nastavka James Bonda / index.hr
Prokletstvo Jamesa Bonda: Filmski tehničar izboden kuhinjskim nožem / vecernji.hr

—

PPS: Navijači su napadnuti samo zato što su Srbi / jutarnji.hr
Napad na Delije izvela Kohorta? / totalportal.hr
Leskovar osudio napad na "Delije" u Boboti / javno.com
Kohorta napala Delije u Boboti / glas-slavonije.hr

—

Počela evakuacija više od 700 putnika s nasukanog cruisera u latvijskom akvatoriju / vecernji.hr
Brod s tisuću putnika nasukao se pred latvijskom obalom / mojportal.hr
Putnički brod nasukao se pred latvijskom obalom / javno.com
Brod s tisuću putnika nasukan na latvijsku obalu / business.hr

—

BiH neće izručiti Ćurkovića i Vrbata Hrvatskoj / dnevnik.hr
BiH neće izručiti Ćurkovića i Vrbata Hrvatskoj / tportal.hr
BiH neće izručiti Ćurkovića i Vrbata Hrvatskoj / jutarnji.hr
Stigle odbijenice: BiH neće isporučiti Ćurkovića i Vrbata radi dvojnog državljanstva / index.hr

—

Suhopolje: U prometnoj nesreći smrtno stradala suvozačica / vecernji.hr
Vozač teško ozlijeđen, a njegova supruga mrtva / tportal.hr
Suhopolje: u prometnoj nesreći smrtno stradala suvozačica / totalportal.hr
Suhopolje: U prometnoj nesreći smrtno stradala suvozačica / dnevnik.hr

—

Ciklon u Mianmaru odnio najmanje 351 život / business.hr
Najmanje 351 poginuli u ciklonu u Mianmaru / mojportal.hr
Razorni ciklon odnio više od 350 života / rtl.hr
Razorni ciklon ne smeta održavanju referenduma / javno.com

—

Prag: Druga izvedba "izgubljene" Vivaldijeve opere nakon 278 godina / dnevnik.hr
Prag: Druga izvedba "izgubljene" Vivaldijeve opere nakon 278 godina / totalportal.hr
Prag: Druga izvedba "izgubljene" Vivaldijeve opere nakon 278 godina / vecernji.hr

—

Magma: za isplatu dividende 9,59 milijuna kuna / totalportal.hr
Magma za isplatu dividende izdvaja 9,59 milijuna kuna / business.hr
Magma za isplatu dividende odredila 9,59 milijuna kuna / liderpress.hr

—

"Ivo Tijardović" postuhumno dodijeljen Dvorniku / javno.com
Nagrade HNK Split - 'Ivo Tijardović" posmrtno Borisu Dvorniku / totalportal.hr
Dvorniku posmrtno dodijeljen "Ivo Tijardović" / tportal.hr

—

Građani Hrvatske kupili čak 32 posto više automobila / totalportal.hr
Najveća mjesečna prodaja automobila do sada / javno.com
Hrvati i dalje najviše kupuju Opele / business.hr

—

Turska: Erdogan će osnovati novu stranku ako se ukine AKP-a / totalportal.hr

Turska: Erdogan će osnovati novu stranku u slučaju ukidanja AKP-a / jutarnji.hr
Turski premijer namjerava osnovati novu stranku / javno.com
—
Značajni turistički porasti u prva tri mjeseca i ožujku / poslovni.hr
Broj noćenja u prva tri mjeseca porastao čak 21,2Hrvatska bilježi dvoznamekasti rast turističkog prometa / business.hr
—
ATP: mali napredak Ljubičića i Čilića / totalportal.hr
Mali napredak Ljubičića i Čilića / javno.com
ATP ulazna lista: Mali napredak Ljubičića i Čilića / dnevnik.hr
—
Ronaldo plakao nakon afere s transvestitima / business.hr
Ronaldo strahuje da je njegov ugled nepovratno narušen / mojportal.hr
Ronaldo strahuje da je njegov ugled narušen / javno.com
—
U prometnoj nesreći teško ozlijeđene četiri osobe / dnevnik.hr
U prometnoj nesreći četvero teško ozlijeđenih / javno.com
Četiri osobe ozlijeđene teško u prometnoj nesreći u Gornjem Bazju / vecernji.hr
—
Pronađene strvine šest zaštićenih tuljana / tportal.hr
Oregon: Pronađene strvine šest zaštićenih morskih lavova / dnevnik.hr
U Oregonu pronađeno šest mrtvih tuljana / javno.com
—
Toyota diže cijene u Sjevernoj Americi / totalportal.hr
Amerikancima će Toyote biti skuplje / business.hr
Toyota diže cijene u Sjevernoj Americi / liderpress.hr
—
Zbog problema u opskrbi, porasle cijene nafte / javno.com
Strahovanja od poremećaja u opskrbi poskupjela barel / seebiz.eu
Strah od nestašice podigao cijene nafte / business.hr
—

Mianmar: Vlast želi referendum unatoč razornom ciklonu / dnevnik.hr
Vojna hunta želi referendum unatoč ciklonu / javno.com
Mianmar: vlasti žele referendum unatoč razornom ciklonu / totalportal.hr

—

Moguć povratak srbijanskih veleposlanika natrag na Kosovo / nacional.hr
Jeremić: Na jesen moguć povratak veleposlanika u zemlje koje su priznale Kosovo / totalportal.hr
Srbija bi na jesen mogla vratiti zbog Kosova povučene veleposlanike / business.hr

—

Afganistan: U tri slučajne eksplozije 6 mrtvih, troje djece / vecernji.hr
Afganistan: u tri slučajne eksplozije 6 mrtvih / totalportal.hr
Kabul: Slučajna eksplozija bombi ubila šest ljudi / javno.com

—

EK odredila uvjete za prihvaćanje "trećeg puta" u energetici / seebiz.eu
EK odredila uvjete za prihvaćanje "trećeg puta" u energetici / poslovni.hr
EU će dozvoliti E.ON-u i EDF-u kontrolu nad distribucijom energije / business.hr

—

Istraga u slučaju incesta u Austriji pred zaključenjem / totalportal.hr
"Josef Fritzl je osoba s nevjerojatnom kriminalnom energijom" / dnevnik.hr
Istraga o austrijskom monstrumu je pri kraju / javno.com

—

Putin potpisao zakon o ograničavanju stranih ulaganja / dnevnik.hr
Putin ograničava strana ulaganja u Rusiji / javno.com
Putin ograničava strani kapital / liderpress.hr

—

Bosanski Srbi obilježili 63. obljetnicu pokušaj proboja iz logora Jasenovac / dnevnik.hr
Obilježena obljetnica neuspjelog proboja / javno.com
Bosanski Srbi obilježili 63. obljetnicu pokušaj proboja iz logora Jasenovac / vecernji.hr

—

”Rusija će u petak pokazati svoj vojni potencijal” / business.hr
Putin najavio veliku vojnu paradu u Moskvi / tportal.hr
Putin: Rusija će u petak pokazati svoj vojni potencijal / totalportal.hr

—

Ingra skupila 79 posto dionica Mavrova / seebiz.eu
Ingra stekla 79 posto dionica Mavrova / liderpress.hr
Ingra drži 79 posto GP Mavrova / poslovni.hr

—

Kvartalna neto dobit Končara 19,1 milijun kuna / poslovni.hr
Grupa Končar ostvarila neto dobit 19,1 milijun kuna / liderpress.hr
Dobit grupe Končar 19,1 milijun kuna / business.hr

—

Za pregledavanje 8,5 milijuna udžbenika - 30 milijuna kn / tportal.hr
Preporod traži 30 milijuna kuna za pregledavanje udžbenika / business.hr
Pregled udžbenika stoji 60 milijuna kn / vecernji.hr

—

Australski liječnik za legalnu prodaju organa / dnevnik.hr
U Australiji predložena legalna prodaja organa / tportal.hr
Australski liječnik za legalizaciju prodaje organa / nacional.hr

—

Predsjednik Mesić posjetio Inu / poslovni.hr
Dragičević o Ininim cijenama: Protiv korupcije se može boriti tržišnim cijenama / business.hr
Oskrba sigurna - a cijene? / vijesti.hrt.hr

—

Zbog ubojstva svećenika Rafaja saslušano pet svjedoka / tportal.hr
U istražnom postupku za ubojstvo svećenika Rafaja saslušano pet svjedoka / dnevnik.hr
Još nema optužnice protivv osumnjičenog za ubojstvo svećenika Josipa Rafaja / index.hr

—

Manje proizvoda u zalihama / jutarnji.hr
Smanjene zalihe gotovih industrijskih proizvoda / poslovni.hr

Zalihe industrijskih proizvoda porasle 0,2 posto / business.hr

—

Nakon što je prije 22 godine otišao iz Zagreba, Mirko Ilić se privremeno vraća velikom retrospektivnom izložbom Mirko Ilić / strip / ilustracija / dizajn / multimedija 1975-2008. / net.hr
Mirko Ilić: Kome to treba?!? / seebiz.eu
Mirko Ilić otkantao hrvatske izdavače / tportal.hr

—

CSKA po šesti put pobjednik Eurolige / totalportal.hr
Po šesti put CSKA - europski prvak / javno.com
Moskovski CSKA šesti put prvaci Europe / dnevnik.hr

—

Imena bucmaste djece izvjesili na oglasnoj ploči / jutarnji.hr
U vrtiću izvješen popis bucmaste djece / nacional.hr
Velika Gorica: u vrtiću izvjesili popis s težinom djece / totalportal.hr

—

Viro odlučuje o isplati 20 kuna dividende / business.hr
Viro: za dividendu 27,63 milijuna kuna / poslovni.hr
Viro predlaže isplatu 20 kuna dividende po dionici / liderpress.hr

—

Ubio ženu nakon što je na Facebooku objavila da ga ostavlja / mojportal.hr
Obiteljska tragedija zbog poruke na Facebooku / tportal.hr
Ubio suprugu nakon što je na Facebooku objavila da ga ostavlja / dnevnik.hr

—

Dvojica dječaka brutalno ubijena pa zapaljena, policija sumnjiči oca / dnevnik.hr
Glasgow: dvojica dječaka spaljena u automobilu / totalportal.hr
Dvojica dječaka u Škotskoj brutalno ubijena pa zapaljena u automobilu / index.hr

—

Prlić: Herceg Bosna poštivala suverenitet BiH / jutarnji.hr
Prlić: Herceg Bosna je poštivala suverenitet BiH / javno.com

—

Milito šest mjeseci izvan terena / totalportal.hr
Milito šest mjeseci izvan terena / javno.com

—

Subversive Film Festival od 18. do 24. svibnja u Zagrebu / vecernji.hr
Subversive Film Festival u Zagrebu / tportal.hr

—

Manic Street Preachers na Radar Festivalu / vecernji.hr
Manic Street Preachers dolaze na Radar festival / javno.com

—

Real prvak Španjolske / totalportal.hr
Španjolska: Real 31. put prvak / mojportal.hr

—

Nakon Ljubičića, ispao i Čilić / javno.com
ATP Rim: Nakon Ljubičića ispao i Čilić / dnevnik.hr

—

Palestinski aktivisti ispalili rakete prema izraelskom teritoriju / totalportal.hr
Pojas Gaze: Palestinski aktivisti ispalili rakete prema izraelskom teritoriju / dnevnik.hr

—

Wall Street na valu pozitivnih vijesti / business.hr
Na Wall Streetu se ovoga tjedna očekuje rast / totalportal.hr

—

Radnici T-HT-a nisu se odazvali pozivu na "bijeli štrajk" / totalportal.hr
HT: Radnici se nisu odazvali štrajku koji smatramo nezakonitim / business.hr

—

Kina i tibetska vlada u egzilu dogovorili nove razgovore / dnevnik.hr
Tibetanska vlada u egzilu dogovorila pregovore / javno.com

—

Njemačka uvodi strože kazne: Za vožnju pod utjecajem alkohola 1000 eura / mojportal.hr
Njemačka drastično povećava kazne u prometu / business.hr

—

Erste uvodi MultiCash uslugu / seebiz.eu
Erste banka uvodi MultiCash uslugu za korporativne klijente / business.hr

—

Najstariji čovjek na svijetu: 118-godišnjak ulazi u Guinessovu knjigu rekorda / mojportal.hr
Najstariji čovjek na svijetu je 118-godišnji Rus / javno.com

—

U urušavanju krijumčarskog tunela poginuo muškarac / totalportal.hr
U rušenju krijumčarskog tunela poginuo Palestinac / javno.com

—

Engleska: Arsenal - Everton 1:0 / mojportal.hr
Chelsea ostao u utrci za prvaka Engleske / javno.com

—

Dolar oslabio prema euru u oskudnoj trgovini na tržištima / business.hr
Dolar slabi, cijene hrane najozbiljniji problem / javno.com

—

EU brine o sigurnosti dječjih igračaka / business.hr
EK ovoga mjeseca potpisuje sporazum s industrijom o sigurnosti igračaka / totalportal.hr

—

Tuerk: Joras je na slovenskom teritoriju / javno.com
Tuerk drži da će izbor arbitra za hrvatsko-slovensku granicu biti težak / dnevnik.hr

—

Novčani fondovi u plusu, PBZ Dollar fond jedini gubitnik / liderpress.hr
Većina investicijskih fondova bilježila dobitke / totalportal.hr

—

ZSE: Financijski rezultati dižu cijene dionica? / business.hr
U fokusu ulagača tromjesečna financijska izvješća / liderpress.hr

—

U brodolomu u Amazoniji najmanje 12 poginulih / javno.com
U Amazoniji najmanje 15 poginulih / vijesti.hrt.hr

—

Incident u autokampu: Slovenska vojska ne vjeruje medijima / mojportal.hr
Slovenija: Ispituje se jesu li slovenski vojnici zapalili hrvatsku zastavu / dnevnik.hr

—

Njemačka mladež sve se više opija / business.hr
Mladi Nijemci sve više piju, a manje puše i konzumiraju lake droge / mojportal.hr

—

Hrvatska će u 2008. zaposliti više od 8.000 stranaca / totalportal.hr
Strancima više od osam tisuća radnih dozvola / nacional.hr

—

Većina dionica u plusu / liderpress.hr
ZSE: Još je prerano za ocjenu da se trend preokrenuo / business.hr

—

Primirje je gotovo / jutarnji.hr
Gotovo primirje Mesić-Sanader? / vijesti.hrt.hr

—

Guverneri središnjih banaka nemoćni pred poskupljenjima hrane / business.hr
Monetarni čelnici oprali ruke od poskupljenja hrane / jutarnji.hr

—

Chromos boje i lakovi preuzeli mostarsku Astru Dubravku / business.hr
Chromos boje i lakovi: ugovor o dokapitalizaciji Astra Dubravke / poslovni.hr

—

Iran unatoč molbama nastavlja obogaćivanje urana / javno.com
Iran se ne odriče prava na nuklearnu tehnologiju / totalportal.hr

—

Boston u drugom krugu / javno.com
NBA: Boston u sedmoj utakmici deklasirao Atlantu / mojportal.hr

—

Srbin nakon svađe polio oca benzinom i zapalio / index.hr

Mladić zapalio svog oca živog nakon svađe / javno.com

—

Tisuće građana na Titovom grobu u Beogradu / javno.com
Obilježena godišnjica smrti Josipa Broza / glas-slavonije.hr

—

Nadan Vidošević negira optužbe za zloporabu / jutarnji.hr
Nadan Vidošević poriče optužbe za malverzaciju prilikom privatizacije Ikom Kovnice / index.hr

—

Sudionici incidenta u Boboti pušteni na slobodu / tportal.hr
Dvojica napadača u Boboti puštena na slobodu / dnevnik.hr

—

U tučnjavi u Boboti lakše ozlijeđeno pet osoba i dijete / mojportal.hr
Bobota: u tučnjavi ozlijeđeno osmero ljudi, od kojih dva djeteta / totalportal.hr

—

Gascoigne ponovno u bolnici / dnevnik.hr
Gascoigne se pokušao ubiti / javno.com

—

Hanžeković: Mislim ozbiljno s Veterinom / seebiz.eu
Hanžeković objavio ponudu za preuzimanje Veterine / poslovni.hr

—

Njemica priznala da je "možda ostavila djecu u zamrzivaču" / index.hr
Majka jedina osumnjičena za ubojstva troje djece / javno.com

—

Prijateljicu nevjenčanog supruga pretukla palicom i opljačkala / vecernji.hr
U stanu ju udarila palicom po glavi i opljačkala / index.hr

—

RS optužuje Silajdžića da zastupa samo Bošnjake / totalportal.hr
"Silajdžić u Ohridu nije predstavljao BiH nego samo Bošnjake" / index.hr

—

Lindsey Lohan gostuje u seriji "Ružna Betty" / totalportal.hr
Lindsay Lohan gostuje u "Ružnoj Betty" / vecernji.hr

—

Galliani: Flamini seli u Milan / javno.com
Flamini definitivno u Milanu / totalportal.hr

—

Bolivijski konzervativci slave pobjedu / javno.com
Morales: Referendum je promašena separatistička mjera / tportal.hr

—

Mesić s vatrogascima / vijesti.hrt.hr
Sanader kod Mesića / javno.com

—

Slovenska paraglajderica poginula na Svilaji / mojportal.hr
Sinj: Poginula paraglajderica / glas-slavonije.hr

—

U Staroj Drenčini novo sisačko groblje sa šest tisuća mjesta / jutarnji.hr
Novo sisačko groblje ne smije više čekati / vecernji.hr

—

Vezao se za Božicu pravde jer liječnike smatra odgovornim za smrt kćerke / dnevnik.hr
Zbog smrti kćeri vezao se za Božicu pravde / tportal.hr

—

Kljaković Gašpić izvrsno startao / tportal.hr
EP Finn: Kljaković-Gašpić četvrti u prvom plovu / totalportal.hr

—

Ribica igra nogomet, košarku i pleše limbo / jutarnji.hr
Video: Riba koja pleše, igra košarku, nogomet...ali stvarno! / dnevnik.hr

—

Kesovija: Neću na HRF s curama koje mašu sisicama / javno.com
Tereza Kesovija više nije na 12. HRF-u / totalportal.hr

—

Nova gazdarica Farme, Jelena, je dokazala da i žena i te kako zna udariti šakom po stolu. S velikom lakoćom kao iz topa je "odvalila" da će ovoga tjedna sluškinje biti Hana i Dvina. Video! / net.hr
Osveta: Hana i Dvina nove sluškinje / dnevnik.hr

—

Kohorta: Ne treba nam slika Arkana, sam skup Delija je provokacija / totalportal.hr

Policija privela dva napadača na Delije / jutarnji.hr

—

Slučaj Močvara bez sretnog završetka: Vrata zauvijek zatvara koncert Leta 3 / index.hr

Let 3 zatvara Močvaru / javno.com

Bibliografija

[Agić and Tadić, 2006] Agić, Ž. and Tadić, M. (2006). Evaluating morphosyntactic tagging of croatian texts. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*.

[Agirre and Edmonds, 2007] Agirre, E. and Edmonds, P., editors (2007). *Word Sense Disambiguation: Algorithms and Applications*. Springer.

[Allan, 2002] Allan, J., editor (2002). *Topic Detection and Tracking: Event-based Information Organization*. Springer.

[Allan et al., 1998a] Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. (1998a). Topic detection and tracking pilot study: Final report. In *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218.

[Allan et al., 1998b] Allan, J., Papka, R., and Lavrenko, V. (1998b). On-line new event detection and tracking. In *In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 37–45.

[Alpaydin, 2004] Alpaydin, E. (2004). *Introduction to Machine Learning*. MIT Press.

[Bakarić, 2009] Bakarić, N. (2009). Aplikacija za ručno stvaranje klastera dokumenata prirodnog jezika. Master's thesis, Odsjek za informacijske znanosti, Filozofski fakultet, Sveučilište u Zagrebu.

[Bar-Hillel, 1960] Bar-Hillel, Y. (1960). The present status of automatic translation of languages. *Advances in Computers*, 1:91–163.

[Carbonell et al., 1999] Carbonell, J., Yang, Y., Lafferty, J., Brown, R., Pierce, T., and Liu, X. (1999). Cmu report on tdt2: Segmentation, detection and tracking. In *Proceedings of DARPA Broadcast News Workshop*.

[Cover and Thomas, 1991] Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley and Sons, Inc.

[Curran, 2004] Curran, J. R. (2004). *From Distributional to Semantic Similarity*. PhD thesis, University of Edinburgh.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., , and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

[Dharanipragada et al., 1999] Dharanipragada, S., Franz, M., McCarley, J., Roukos, S., and Ward, T. (1999). Ibm slide presentation at the the darpa broadcast news workshop. Presentation.

[Dice, 1945] Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Journal of Ecology*, 26(3):297–302.

[Doddington, 1999] Doddington, G. (1999). Presentation slides at the darpa broadcast news workshop. Presentation.

[Fellbaum, 1998] Fellbaum, C., editor (1998). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

[Forster, 2006] Forster, R. (2006). *Document Clustering in Large German Corpora Using Natural Language Processing*. PhD thesis, University of Zurich.

[Gowda and Diday, 1991] Gowda, K. C. and Diday, E. (1991). Symbolic clustering using a new dissimilarity measure. *Pattern Recognition*, 24(6):567–578.

[Grefenstette, 1994] Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA.

[Jain and Dubes, 1988] Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice-Hall advanced reference series. Prentice-Hall Inc., Upper Saddle River, NJ.

[Jain et al., 1999] Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323.

[Jurafsky and Martin, 2008] Jurafsky, D. and Martin, J. H. (2008). *Speech and Language Processing (2nd Edition)*. Prentice Hall Series in Artificial Intelligence. Prentice Hall, 2 edition.

[Lee, 1999] Lee, L. (1999). Measures of distributional similarity. In *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics*, pages 25–32, Vancouver, B.C. Association for Computational Linguistics.

[Li et al., 2005] Li, Z., Wang, B., Li, M., and Ma, W.-Y. (2005). A probabilistic model for retrospective news event detection. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 106–113, New York, NY, USA. ACM.

[Lombard, 1986] Lombard, L. B. (1986). *Events: A Metaphysical Study*. Routledge & Kegan Paul, Boston, MA.

[Lowe, 1999] Lowe, S. A. (1999). The beta-binomial mixture model and its applications to tdt tracking and detection. In *Proceedings of DARPA Broadcast News Workshop*, pages 127–132.

[Lyman and Varian, 2003] Lyman, P. and Varian, H. R. (2003). How much information? Technical report, School of Information Management and Systems.

[Macqueen, 1967] Macqueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA.

[Manning et al., 2008a] Manning, C. D., Raghavan, P., and Schütze, H. (2008a). *Information retrieval*. Cambridge University Press.

[Manning et al., 2008b] Manning, C. D., Raghavan, P., and Schütze, H. (2008b). *Information retrieval*, chapter Evaluation of Clustering. Cambridge University Press.

[Manning and Schütze, 1999a] Manning, C. D. and Schütze, H. (1999a). *Foundations of Statistical Natural Language Processing*, chapter Topics in Information Retrieval. The MIT Press.

[Manning and Schütze, 1999b] Manning, C. D. and Schütze, H. (1999b). *Foundations of Statistical Natural Language Processing*, chapter Part-of-Speech Tagging. The MIT Press.

[Manning and Schütze, 1999c] Manning, C. D. and Schütze, H. (1999c). *Foundations of Statistical Natural Language Processing*, chapter Clustering. The MIT Press.

[Manning and Schütze, 1999d] Manning, C. D. and Schütze, H. (1999d). *Foundations of Statistical Natural Language Processing*. The MIT Press.

[Manning and Schütze, 1999e] Manning, C. D. and Schütze, H. (1999e). *Foundations of Statistical Natural Language Processing*, chapter Collocations. The MIT Press.

[Mayeux, 1996] Mayeux, P. E. (1996). *Broadcast News: Writing & Reporting*, page 79. Brown & Benchmark Publishers, Guilford, CT.

[Papka, 1999] Papka, R. (1999). *On-line New Event Detection, Clustering and Tracking*. PhD thesis, University of Massachusetts.

[Papka and Allan, 1998] Papka, R. and Allan, J. (1998). On-line new event detection using single pass clustering. Technical report, University of Massachusetts, Amherst,MA.

[Pereira et al., 1993] Pereira, F., TIshby, N., and Lee, L. (1993). Distributional clustering of english words. In *Proceedings of the 31st annual meeting*

of the Association for Computational Linguistics, pages 183–190, Columbus, OH1.

[Popper, 1968] Popper, K. R. (1968). *The Logic of Scientific Discovery*. Harper & Row Publishers, New York, NY, USA.

[Salton, 1988] Salton, G., editor (1988). *Automatic text processing*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA.

[Schultz and Liberman, 1999] Schultz, J. M. and Liberman, M. (1999). Topic detection and tracking using idf-weighted cosine coefficient. In *Proceedings of DARPA Broadcast News Workshop*.

[Systran, 2009] Systran (2009). Systran - online translation, translation software and tools.

[Tadić, 2003] Tadić, M. (2003). *Jezične tehnologije i hrvatski jezik*. Ex Libris.

[Tanimoto, 1958] Tanimoto, T. T. (1958). An element mathematical theory of classification. Technical report, IBM Research, New York, NY.

[TDT, 2004] TDT (2004). *TDT 2004 - Annotation Manual*. Linguistic Data Consortium.

[van Rijsbergen, 1979] van Rijsbergen, C. J. (1979). *Information retrieval*. Butterworths, London, 2 edition.

[Voorhees, 1986] Voorhees, E. M. (1986). *The effectiveness and efficiency of agglomerative hierarchic clustering in document retrieval*. PhD thesis, Cornell University, Ithaca, NY.

[Walls et al., 1999] Walls, F., Jin, H., SIsta, S., and Schwartz, R. (1999). Topic detection in broadcast news. In *Proceedings of DARPA Broadcast News Workshop*, pages 193–198.

[Weaver, 1955] Weaver, W. (1955). Machine translation of languages. In Locke, W. N. and Boothe, A. D., editors, *Translation*, pages 15–23. MIT Press, Cambridge, MA.

[Wei and Lee, 2004] Wei, C.-P. and Lee, Y.-H. (2004). Event detection from online news documents for supporting environmental scanning. *Decis. Support Syst.*, 36(4):385–401.

[Wikipedia, 2009a] Wikipedia (2009a). Cluster analysis — wikipedia, the free encyclopedia. pristupljeno 18. siječnja 2009.

[Wikipedia, 2009b] Wikipedia (2009b). Taxicab geometry — wikipedia, the free encyclopedia. pristupljeno 20. veljače 2009.

[Willett, 1988] Willett, P. (1988). Recent trends in hierarchic document clustering: a critical review. *Information Processing and Management*, 24(5):577–597.

[Yang et al., 1998] Yang, Y., Pierce, T., and Carbonell, J. (1998). A study on retrospective and on-line event detection. In *In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 28–36.

[Zahn, 1971] Zahn, C. T. (1971). Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers*, 20(1):68–86.

[ZAPI, 2009] ZAPI (2009). Zavod za poslovna istraživanja.

Sažetak

Osnovni problem koji se u ovoj doktorskoj disertaciji obrađuje je problem pronalaženja događaja u višestrukim izvorima informacija. Uzorak na kojemu se istraživanje provodi sadrži 2,486 dokumenata objavljenih na 17 hrvatskih internetskih portala u vremenskom rasponu od tri dana. Uzorak je označen upotrebom posebno razvijene aplikacije. Ljudskim su označiteljima nuđeni unaprijed izračunati dokumenti kandidati. Uzorak je analiziran i izračunata su dva κ koeficijenta. Za evaluaciju postupaka korišteno je šest evaluacijskih mjera redovito korištenih za evaluaciju rezultata grožđenja. Optimalnom se mjerom pokazala $F_{0.5}$ mjera zbog veće važnosti preciznosti s obzirom na dani zadatak. Čistoća se pokazala neprimjerena mjera za neparticijske algoritme, a NMI i RI kao neprimjerene mjere za evaluaciju ovog zadatka zbog velikog broja istinito negativnih parova dokumenata. Empirijski je ispitan cijeli niz varijabli. Uspoređena su tri hijerarhijska algoritma grožđenja i algoritam jednim prolaskom. Posljednji se pokazao jednako uspješnim kao i hijerarhijski te je odabran kao optimalan iz razloga što je vremenski manje kompleksan od hijerarhijskih. Uspoređeno je šest mjera udaljenosti te je odabrana kosinusna mjera s redovito boljim rezultatima i manjom vremenskom kompleksnosti. Dvije postavljene heuristike vezane uz vrijeme i mjesto objave dokumenata su ispitane *in vitro* i *in vivo* te su se u oba slučaja pokazale korisnima. Između pet mjera težine svojstava odabran je TF-IDF. Istraženo je pet metoda odabira i određivanja svojstava na razini pojavnica te četiri metode na višim jezičnim razinama. Općenito su se metode na razini pojavnica pokazale korisnima za razliku od metoda na višim jezičnim razinama. Referentni korpus od pola milijuna pojavnica se pokazao optimalnim. Optimizacijom cijelog postupka pronalaženja događaja postignuta je $F_{0.5}$ mjera od ~ 0.82.

Ključne riječi: pronalaženje događaja, grožđenje, mjere udaljenosti, mjere težine svojstava, formalizacija dokumenta

Summary

The research in this dissertation is focused on the problem of event detection in parallel information sources. The data sample used in the research contains 2,486 documents collected from 17 Croatian news portals published in a time span of three days. The sample is tagged by human annotators using an application developed for this purpose. Human annotators are given document candidates calculated in advance. The tagged sample is analyzed and two κ coefficients are calculated. Six typical clustering evaluation measures are used in the research. The $F_{0.5}$ measure has proved itself as optimal for this task since it favors precision over recall. Purity is not applicable for non-partitional clustering algorithms, while NMI and RI are not suitable for this task because of the high number of true negatives. A list of variables is empirically tested. Three hierarchical clustering algorithms and one single-pass algorithm are compared. The latter is proven to be as efficient as the hierarchical ones that are more complex. Six distance measures are compared and the cosine measure is chosen as the optimal one with better results and lesser time complexity. Two heuristics concerning the time and place the documents were published are proven useful both *in vitro* and *in vivo*. From five feature weight measures the classical TF-IDF is chosen. Five methods of feature selection and extraction on the token level and four methods on higher language levels are also evaluated. In general the simpler methods on token level are more efficient for the given task than the more complex ones. A reference corpus of half a million of tokens is proven to be most efficient. By optimizing the whole procedure of event detection, an $F_{0.5}$ score of ~ 0.82 is achieved.

Keywords: event detection, clustering, distance measures, feature weight measures, document formalization

Životopis

Rođen sam 26. ožujka 1979. godine u Zagrebu. Nižu osnovnu školu pohađam u Zagrebu te Heidelbergu (Njemačka). Višu osnovnu školu u klasičnom razredu pohađam u Zagrebu nakon čega upisujem XV. matematičku gimnaziju u Zagrebu. Paralelno s osnovnom i srednjom školom završavam i osnovnu te srednju glazbenu školu, instrument klarinet.

Po završenom srednjoškolskom obrazovanju upisujem studij njemačkog jezika i književnosti te informacijskih znanosti na Filozofskom fakultetu u Zagrebu.

Po diplomi upisujem poslijediplomski studij informacijskih znanosti na Odsjeku za informacijske znanosti Filozofskog fakulteta gdje pola godine kasnije postajem i znanstveni novak.

Sudjelujem na više znanstvenih projekata i objavljujem desetke znanstvenih radova. Osnovno područje interesa mi je obrada prirodnog jezika, odnosno računalno jezikoslovlje.

Sadržaj

Popis tablica

Popis slika