Analiza teksta

Josip Begić
PMF-MO
Zagreb, Croatia

***Sažetak*—Rudarenje teksta (eng. text minining) i analiza stava (eng. sentiment analysis) važno su područje pretraživanja podataka (eng. data mining) koje je s porastom interneta i količine teksta koju računala mogu obraditi postalo izrazito važno u brojnim segmentima života. Želimo li preporučiti novi proizvod kupcu, odrediti kakvo mišljenje prevladava o nekoj temi, (smisleno) grupirati velike skupove podataka u manje cjeline ili odrediti glavne značajke nekog teksta vrlo vjerojatno ćemo morati koristiti neku od metoda analize stava, odnosno rudarenja teksta. U ovom radu dan nam je skup od 400000 recenzija piva sa internet stranice *www.ratebeer.com*, te pokušavamo dati odgovor na pitanje "Koje je najbolje pivo na svijetu?". U tu svrhu korištene su metode nadziranog i nenadziranog strojnog učenja. Poseban naglasak stavljen je na korištenje različitih metoda za selekciju obilježja.**

***Keywords*—strojno učenje, analiza stava, rudarenje teksta, KNIME**

I. Uvod

Za ovaj projekt korišten je skup od 379789 recenzija skupljenih sa internet stranice *www.ratebeer.com*. Podaci su dobiveni u sklopu data mining natjecanja Mozgalo. Prvi dio ovog rada bavi se njihovom analizom - na kojim su jezicima tekstovi napisani, u koje vrijeme, jesu li svi tekstovi jedinstveni ili se ponavljaju, filtriranjem neželjenih tekstova, proučavanjem kako su ocjene distribuirane, koliko različitih piva uopće gledamo, koliko ima korisnika... Ovaj dio projekta rađen je u SQL jeziku za upite (eng. SQL query language) Cloudera Impala, te u manjoj mjeri u programskom jeziku Excel i platformi Pentaho. Nakon što smo obavili potrebnu analizu baze podataka, koja će detaljnije biti objašnjena u idućoj sekciji, za ostatak projekta korištena je platforma za rudarenje podataka (eng. data mining platform) KNIME u kojoj su već implementirani brojni algoritmi strojnog učenja, te programi za analizu teksta.

II. Baza podataka

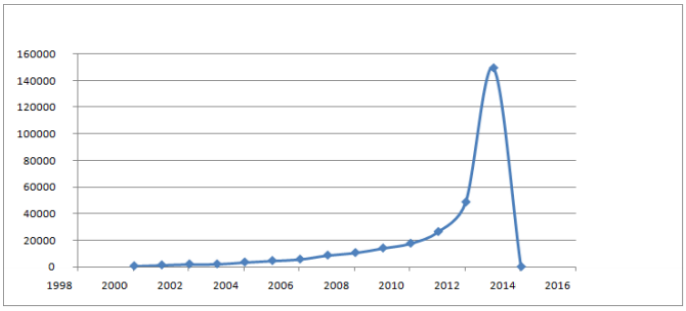
Dana baza sadržavala je, kao što je već spomenuto, 379789 recenzija o pivima, te je za svaku recenziju pisalo ime autora, datum objave teksta, ocjena koju je autor dao pivi, ključ koji je jedinstveno određivao svaku recenziju, jezik na kojem je recenzija napisana, ime recenziranog piva, te još neke manje bitne informacije - datum kada su podaci preuzeti (svi unutar 2, 3 dana), s koje su stranice skinuti (svi sa *www.ratebeer.com*) te url adresa danog teksta (obzirom da smo imali ključ koji je jedinstveno određivao svaku recenziju, ova informacija nam nije bila potrebna).

Odlučili smo, radi jednostavnosti, gledati samo tekstove na engleskom jeziku. Naime, većina programa napisanih za preprocesiranje teksta postoji samo za "veće" jezike kao što su engleski, njemački i francuski (pa je ostale bilo preteško analizirati). U danoj bazi podataka označena su 25 jezika na kojima su tekstovi napisani, ali ispostavilo se da je dobar dio tih podataka kriv. 232522 recenzije označene su kao engleske, 2900 njemačke, 2477 talijanske... Ove recenzije zaista su bile na navedenim jezicima, ali je zato bilo jasno da 59306 recenzija označenih kao estonskih (kratica et) vjerojatno nisu na estonskom. Razlog ovomu moglo bi biti da korisnici slučajno krivo označe estonski umjesto engleskog (kratica en). Detaljnom provjerom označenog jezika došli smo do zaključka da su tekstovi na estonskom, rumunjskom i mađarskom (svih je bilo barem oko 10000) zapravo napisani na engleskom, pa smo ih odlučili ostaviti u bazi, dok smo ostale obrisali. Time nam je ostalo 321676 tekstova na engleskom jeziku. Ovime je stupac koji je označavao jezik recenzije postao nebitan za daljnu analizu, pa ga više nismo koristili.

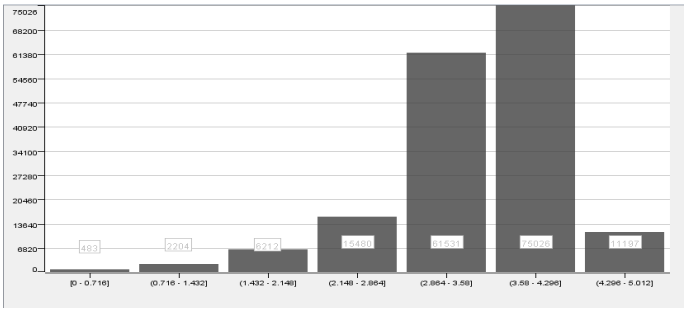
U bazi su se nalazili tekstovi o 30135 piva, koje je napisalo 15752 različitih korisnika. Također, primjećeno je da se 79 piva više ne proizvodi (označene kao "umirovljene", eng. retired), pa smo pripadne recenzije, njih 19909 odlučili izbrisati - očekujemo da će najbolje pivo biti neka koja još uvijek postoji.

Idući korak bio je odrediti koliko stare recenzije želimo gledati. Stranica *www.ratebeer.com* osnovana je prije 14 godina, i pitanje je koliko su recenzije iz prošlog desetljeća relevantne za isto pivo danas. Kvaliteta piva mogla je pasti, moglo se prestati proizvoditi, mogao se promijeniti cijeli stil piva, način pripreme... Također, prema [1], primijećeno je da pojedine riječi u različitim vremenskim razdobljima postaju važnije, ili da im se mijenja značenje, ovisno o aktualnom trendu. Npr., danas riječi poput "YOLO!", "OMG" imaju neko značenje, dok su prije desetak godina izgledale kao nasumični niz znakova. Slično je i na različitim zajednicama na internetu, konkretno na *www.ratebeer.com* riječ "aroma" oko 2003. godine bila je vrlo korištena (10% korisnika je koristi) a riječ "smell" (koja se u kontekstu piva smatra sinonimom, obje riječi označavaju miris) je rijetko korištena (manje od 2% korisnika je upotrebljava). 10 godina poslije trend se promijenio, i sada riječ "aroma" ne koristi više nitko, a riječ "smell" se koristi kako bi se opisao miris piva. Na slici 1 možemo vidjeti distribuciju broja recenzija po godinama. Vidimo da se broj recenzija u zadnjih nekoliko godina eksponencijalno povećao, te motivirani prethodnim razmišljanjem zaključili smo da nećemo izgubiti previše važnih informacija ako izbrišemo objave starije od 2010. godine.

Nadalje, postavlja se pitanje imaju li nam duljina same recenzije nekakvu važnost? Jesu li sve recenzije smislene, neovisno o tome koliko riječi sadrže? Naravno, ispostavilo se da nisu sve recenzije jednako "dobre" - većina tekstova sa manje od 40 znakova bila je besmislena (npr. napisana na čudnim slovima koja su podsjećala na kineska ili arapska, ponavljala se jedno slovo 40 puta, ponavljala se jedna riječ više puta isl.). Slično, tekstovi sa preko 6000 znakova bili su previše

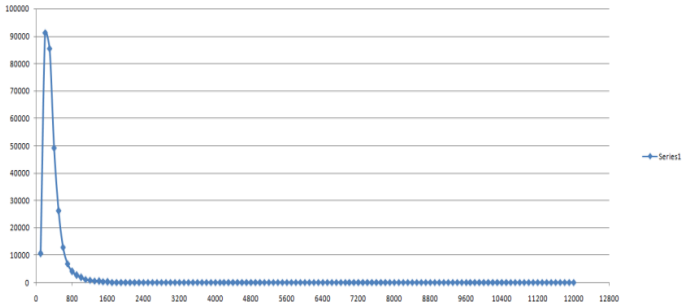


Slika 1. Distribucija broja recenzija po godinama



Slika 3. Histogram ocjena

dugački (nerijetko isti tekst se samo kopirao iznova i iznova) i iz njih se nije mogla dobiti korisna informacija o opisanoj pivi - čak ni nama nije bilo jasno što je autor htio reći, pa nije bilo za očekivati da će računalu biti jasnije. Smatrali smo da će ovakvi tekstovi predstavljati šum na podacima, i da ih se bolje riješiti. Pogledamo li sliku 2, na kojoj je prikazana distribucija duljine teksta, vidimo da se ipak najveći broj tekstova grupira oko 150 riječi, dok je ovih sa manje od 40 i više od 1000 vrlo malo. Zato nam se činilo smislenim izbrisati sve recenzije osim onih koje imaju između 50 i 800 znakova.



Slika 2. Distribucija duljine recenzija

Za navedene radnje koristili smo platformu Pentaho, te u manjoj mjeri programski jezik Excel. Također se ispostavilo da se neki tekstovi ponavljaju više puta, odnosno da su *spam*, pa smo i njih izbrisali jer nisu sadržavali nikakvu korisnu informaciju. Odlučili smo iz baze izbrisati i sva piva koje su recenzirane manje od 6 puta, jer smo smatrali da skup od svega 6 osoba ne može dovoljno objektivno ocjeniti neku pivu. Na kraju, ne možemo nikome tvrditi da je pivo koje je probalo svega nekoliko osoba najbolje na svijetu.

U ovom trenutku, kada smo izbrisali sve po nama nebitne, baza je sadržavala 172 133 članaka koji su opisivali 20 485 različitih piva i napisalo ih je 6 286 različitih korisnika. Dobivenu tablicu u Pentahu prebacili smo u csv fromat, te smo s njom nastavili raditi u platformi za rudarenje podataka, KNIME.

III. Model

Na dobivenom skupu podataka razvijen je BoW model (eng. bag of words, vreća riječi), koji će detaljnije biti objašnjen u idućim sekcijama. BoW modelom od danog skupa tekstova dobili smo vektorski prostor, na kojemu smo mogli koristiti algoritme nenadziranog i nadziranog strojnog učenja. Obzirom na pitanje koje je postavljeno na početku projekta ("Koje je najbolje pivo na svijetu?"), trebalo je definirati što znači najbolje pivo i kako isto dobiti iz ovako velikog skupa podataka. Obzirom da je uz recenziju svakog piva bila dana i njena ocjena koju je autor teksta smatrao prikladnom, najjednostavniji pristup bio je izračunati srednju ocjenu za svako pivo, sortirati ih silazno po ocjenama i prvu u tablici proglasiti najboljom. Ipak, u tablici 1 jasno se vidi problem ovakvog pristupa - najbolja piva bila su one koja su imale svega jednu ili dvije (subjektivne) recenzije. Reći da je najbolje pivo neko koje je probala nekolicina ljudi bilo je besmisleno i valjalo je napraviti nešto drugo. U tablici 4 može se vidjeti najbolje rangirana piva koje je ocjenilo više od 10 osoba - ova lista puno je smislenija i na njoj se zaista nalaze kvalitetna, općeprihvaćena piva.

Ime	Srednja ocjena	Broj recenzija
Founders Nitro	5	1
Orladno Brewing	5	1
Westbrook Rummy	5	1
Cask Larder Tomok	4.8	1
Lost Coast Barley	4.8	1
New Albanian Citra Ass clown	4.8	1

Tablica 1. Najbolje ocjene i broj recenzija

Ime	Srednja ocjena	Broj recenzija
Founders KBS Kent	4.44	60
Stone Enjoy By Ipa	4.41	11
Russian River Pliny The Elder	4.4	85
Cantillon Saint Lamvinus	4.39	28
New Glarus R& D Sour Ale	4.39	42
Westvleteren 12X	4.39	23

Tablica 2. Najbolje ocjene za piva s više od 10 recenzija

Ipak, odmah uočavamo da se na toj listi nalaze vrlo elitna piva, uglavnom sve jake, tamna piva čija je osnovna karakteristika veća gorčina i intenzivniji okus. Na slikama 4 i 5 možemo vidjeti najčešće atribute koji opisuju piva sa ocjenama iznad 4.5. Postavlja se pitanje koliko je točno reći da je najbolja piva neka koju preferira manji skup iskusnih ljudi kao što su oni sa *www.ratebeer.com*. Preciznije, ima li smisla preporučiti nekom neiskusnijem, novom kušaču gorku tamnu pivu, ako njegovo nepce nije naviknuto na takav okus i pritom očekivati da će i njemu navedena piva biti najbolja? Prema [3], u internet zajednicama poput *www.ratebeer.com* svaki korisnik prolazi nekoliko stadija iskustva, tijekom kojih se njegov ukus razvija od amatera do eksperta. Naravno, u različitim razdobljima korisnicima se sviđaju različita piva, i zato bi dobar model trebao razlikovati iskustvo pojedinog

člana i na temelju toga davati odgovor koje je najbolje pivo na svijetu. Nažalost, zbog ograničenosti danog skupa podataka, nedovoljnog znanja i nedostatka potrebnog alata, nismo bili u mogućnosti provesti sličan model.

U spomenutom članku [3] primjećene su neke zanimljive činjenice vezane uz internet zajednice, od koje je za našu temu važna bila ona da iskusniji korisnici daju ekstremnije ocjene proizvodima (ili ocjene bliže 5 ako je pivo dobro, ili bliže 1 ako je loše), dok početnici ocjenjuju sve proizvode umjerenije. Još je zanimljivije da će početnik i pivo koje mu se ne svidi ocjeniti visokom ocjenom, ako je ono tako ocijenjeno od strane iskusnih korisnika. Stoga smo zaključili, što su i potvrdili naši rezultati, da će najbolje pivo vjerojatno biti ono koje preferiraju iskusniji članovi zajednice.

Na kraju, postavilo se pitanje je li moguće nekako grupirati piva obzirom na članke? Možemo li, na temelju opisa pojedinog piva, svrstati je u kategoriju, i ako da kakvu (npr. piva koju preferiraju iskusni članovi, ili piva obzirom na okus ili boju)? Obzirom da nismo imali nikakve unaprijed određene grupe, jedini način bio je koristiti neku metodu nenadziranog strojnog učenja, poput k-means. O opisu algoritma, implementaciji i rezultatima više se govori u idućim sekcijama.

A. Predprocesiranje i BoW model

U ovom odjeljku opisan je postupak predprocesiranja teksta i BoW model. Predprocesiranje je postupak kojim pripremamo postojeći tekst za računalnu analizu - uklanjamo nebitne dijelove teksta poput veznika i interpunkcijskih znakova. BoW modelom dobivamo vektorski prostor, a u ovom odjeljku pokazujemo i važnost odabira riječi (odnosno značajki, eng. features selection) koje ćemo koristiti kao vektore. Za sve potrebne operacije korištena je ekstenzija za predprocesiranje u platformi KNIME.

1) Predprocesiranje: Problem koji se javlja pri analizi stava je da nerijetko dobivamo ogromne količine teksta koje nije jednostavno, ili je čak nemoguće, pohraniti u računalu i dalje analizirati. Zato je potrebno ukloniti sve dijelove teksta koji u sebi ne sadrže nikakvu informaciju: veznike, čestice, interpunkcijske znakove, stop - riječi, priloge... Također, riječi poput "dark" i "darky" želimo gledati kao iste - značenje im se bitno ne razlikuje, pa u takvom slučaju nema potrebe povećati dimenziju prostora i time otežati rješavanje problema. Za to se brinu stemmeri, algortmi koji sve riječi stavljaju u njihov osnovni oblik (npr. nominativ jednine i infinitiv za imenice i glagole). Ukratko opisujemo korištene čvorove

- **POS tagger, POS filter** - POS (eng. part of speech) tagger je čvor koji svakoj riječi pridodaje njenu vrstu (imenica, glagol, veznik, prilog, pridjev...). POS filter služi kako bi od svih riječi ostavili samo one koje smatramo korisnima u tekstu. U našem slučaju to su bile imenice i pridjevi, smatrali smo (slično je korišteno i u drugim radovima poput [2]) da oni sadrže najvažniju informaciju o tekstu.
- **Stop word filter** - ovaj čvor traži u tekstu tzv. stop riječi, one koje nemaju nikakvo značenje
- **Case converter** - ovaj čvor sva slova pretvara u mala, riječi koje se razlikuju samo u tome jesu li napisane velikim ili malim slovima trebaju biti predstavljene istim vektorom (npr. "Black" i "black")
- **Porter stemmer** - stemmeri služe kako bi svaku riječ prebacili u njen osnovni oblik, tj. u korijen riječi. Time se značajno smanjuje dimenzija prostora
- **Punctuation erasure** - koristimo kako bi obrisali interpunkcijske znakove

2) BoW model: Nakon učinjenih koraka, dimenzija prostora je i dalje prevelika (preko 500000), pa da izbjegnemo problem prokletstva dimenzije (eng. curse of dimensionality) želimo se riješiti riječi koje ne opisuju piva, niti nam daju ikakvu informaciju o istima. Više je načina na koji se ovo može napraviti, a u ovom radu korištena su 3 vrlo česta: računanje frekvencije pojmova i inverzne frekvencije dokumenata (tzv. TF*IDF), χ^2 izvlačenje riječi (eng. χ^2 Keyword Extractor) i izvlačenje riječi metodom Keygrapha (eng. Keygraph keyword extractor). Ukratko objašnjavamo svaku metodu.

TF∗IDF - Kratica TF označava frekvenciju pojma (eng. term frequency), njome mjerimo koliko se pojedini pojam pojavio puta u u dokumentu, dok IDF (eng. inverse document frequency) predstavlja težinu koja smanjuje važnost riječima koje se pojavljuju prečesto, a računa se po formuli

$$IDF(t, D) = log\frac{N}{1 + |\{d \in D : t \in d\}|}$$

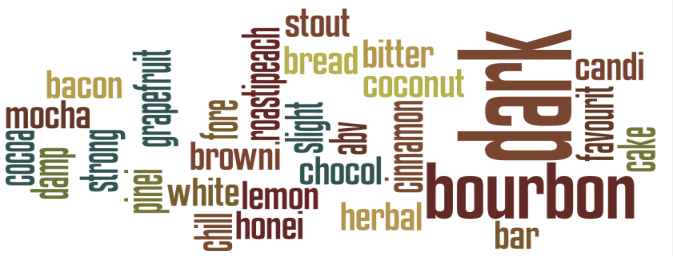
pri čemu N označava broj dokumenata, dok izraz u nazivniku označava broj dokumenata u kojima se pojavljuje pojam t. Pomnožimo li $tf * idf$, dobivamo mjeru koliko je pojedina riječ važna za dokument.

χ^2- *izvlačenje riječi* - Ova metoda opisana je u [6], a radi na način da najfrekventnije pojmove grupira u klastere s obzirom na L_1 normu koja predstavlja udaljenost između njihove vjerojatnosne distribucije. Jednom grupirani pojmovi zatim se rangiraju u padajućem poretku ovisno o tome koliko odudara očekivano pojavljivanje u klasteru od stvarnog pojavljivanja. Ključne riječi su one sa najvećom razlikom.

Slika 4. Riječi koje opisuju najpozitivnije tekstove dobivene χ^2 metodom

Keygraph metoda - Najfrekventnije riječi predstavljaju početne čvorove u grafu, te se zatim računa jačina veza između čvorova. Povezani podgrafovi danog grafa predstavljaju klastere. Novi pojmovi zatim se dodaju grafu ovisno o tome koliko su blizu kojem čvoru. Ova metoda detaljnije je opisana u [5].

Na idućoj tablici možemo vidjeti najčešće riječi dobivene pojedinom metodom, te njihovu težinu. Zadnji korak bio je iskoristiti čvor *Document vector* koji je svakome dokumentu



Slika 5. Riječi koje opisuju najpozitivnije tekstove dobivene Keygraph metodom

pridružio vektor sa riječima koje se u njemu pojavljuju. Time smo dobili matricu čiji su retci predstavljali dokument, a stupci vektore koji se u njemu pojavljuju. Ovim metodama dobiven je vektorski prostor prihvatljive dimenzije te smo sada mogli iskoristi poznate algoritme strojnog učenja.

TF*IDF	χ^2(χ vrijednost)	keygraph (score)
coffer(3.11)	dark(16.65)	cocoa (49)
worst(3.01)	bottl(15.8)	browni(45)
lightcreami(2.51)	huge(15.8)	cake(45)
favorit (2.31)	smooth(15.75)	bourbon(37)
crap(2.29)	black(14.08)	coconut(37)
bottl(2.11)	regular(13.65)	bitter(36)
alchocholicsh(2.1)	mouthfeel(13.65)	chocol(36)
cinamon(2.04)	nice(12.6)	dark(33)
lemonad (2.01)	batch(12.18)	white(32)

Tablica 3. Najčešće riječi i pripadna težina

B. *Strojno učenje*

U ovom odjelju ukratko su opisane metode strojnog učenja korištene u projektu. Metoda potpornih vektora (SVM) i Naivni Bayes (NB) korišteni su kako bi klasificirali tekstove u pozitivnu ili negativnu klasu. Naučeni modeli zatim se mogu pozvati na još neocjenjenom skupu podataka dobivenog sa navedene ili neke druge stranice. K-means klasteriranje korišteno je kako bi vidjeli možemo li grupirati piva u manje grupe ovisno o tekstu koji ih opisuje. Sve algoritme pokrenuli smo na skupu riječi dobivenih metodama $TF * IDF$, χ^2 i keygraph, te na kraju usporedili dobivene rezultate.

Naivni Bayes (eng. Naive Bayes) jedna je od osnovnih metoda strojnog učenja, koja unatoč jakim pretpostavkama često daje vrlo dobre rezultate. Ideja metode je odrediti vjerojatnost da određeni tekst pripada klasi (dobar, loš, neutralan) na temelju pojmova koji se u tekstu pojavljuju. Preciznije, ako imamo $c_1, \ldots, c_k$ mogućih klasa, vjerojatnost da dokument x pripada klasi $c_j, j = 1, \ldots, k$ dana je Bayesovim pravilom:

$$\mathbb{P}(c_j|x) = \frac{\mathbb{P}(x|c_j)\mathbb{P}(c_j)}{\mathbb{P}(x)}$$

Pretpostavka metode je uvjetna nezavisnost pojmova x uz danu klasu c_j, koja ne vrijedi uvijek, ali metoda daje vrlo dobre rezultate u praksi.

Metoda potpornih vektora (eng. Support vector machine) metoda bazira se na traženju $n-1$ dimenzionalne hipperravnine kojom želimo razdvojiti tekstove na pozitivne i negativne, ovisno o danoj ocjeni. Prednost ove metode je činjenica da ne tražimo bilo koju razdvajajuću hiperravninu, nego onu koja ima najveću marginu razdvajanja (rubne rečenice biti će najdalje moguće). Naravno, dimenzija prostora je broj pojmova koje smo dobili metodama za izvlačenje ključnih riječi, a svaka riječ predstavlja jedan potprostor u tom prostoru. U platformi KNIME moguće je uzeti nekoliko različitih jezgri, a mi smo koristili RBF, odnosno Gaussovu jezgru, varirajući parametar σ.

K-means K-means klasteriranje metoda je nenadziranog strojnog učenja koja se bazira na tome da odredimo broj klastera k, te algoritam grupira postojeće podatke u tih k grupa tako da udaljenost između svakog elementa i centra klastera bude minimalna moguća. U našem slučaju, sve pojmove koji su se pojavili u svim tekstovima o pojedinom pivu stavili smo u jedan vektor, gledajući srednju vrijednost pojavljivanja svakog pojmova. Time je svaka piva bila reprezentirana jednim vektorom, te je algoritam pokušavao pronaći podskupove u prostoru za koje će udaljenost njihovih elemenata biti minimalna moguća. Algoritam je radio do 99 iteracija, a broj k smo varirali.

IV. Eksperimenti i rezultati

Ideja je bila naučiti klasifikator da razlikuje samo dvije klase, kako bi dobili ekstremnije ocjene za svaki komentar - ili je dobar, ili je grozan. Prvo smo podjelili tekstove na one izrazito pozitivne (ocjena iznad 4.4) i izrazito negativne (ocjena ispod 2.1), te smo na njima odlučili trenirati i testirati algoritam. Dani skup podataka sadržavao je 13721 tekstova, od toga 6122 negativna i 7599 pozitivna. Iz ovog broja tekstova dobili smo BoW model dimenzije 211 117, koji smo zatim opisanim metodama u sekciji III, A reducirali na neki prihvatljive veličine. U tablici 4 možemo vidjeti dimenzije prostora dobivene iz ovog skupa tekstova.

metoda:	BoW	Keygraph ekstraktor	χ^2 ekstraktor	TF*IDF
dimenzija:	211 117	123 611	123 571	84 341

Tablica 4. Dimenzija prostora dobivena pripadnom metodom izvlačenja ključnih riječi

Performanse računala omogućile su nam treniranje na samo 10% skupa, odnosno 1372 slučajno odabrana članka. U tablici 5 možemo vidjeti uspješnost pojedinog klasifikatora ovisno o tome koju metodu ekstracije ključnih riječi koristi. Iz dobivenih rezultata zaključili smo da SVM uz parametar $\sigma = 0.5$ i Keyword Keygraph ekstrakciju ključnih riječi daje najbolje rezultate. Zanimljivo je primjetiti da u prostoru dobivenom χ^2 metodom nismo dobivali bolje rezultate od onog koji smo dobili pomoću TF*IDF metode, unatoč tome da je dimenzija prostora bila gotovo duplo veća. Napominjemo da smo pokretali SVM i za druge parametre σ, ali rezultati su bili jednaki ili lošiji od navedenih.

	Linearna regresija χ^2 ekstraktor	*SVM uz $\sigma = 0.5$ χ^2 ekstraktor*	*Naive Bayes χ^2 ekstraktor*	*Naive Bayes keygraph ekstraktor*	*SVM uz $\sigma = 0.5$ Keygraph ekstraktor*	*SVM uz $\sigma = 0.5$ TF*IDF*
recall:	0.71	0.52	0.57	0.6	0.84	0.75
precision:	0.75	0.56	0.58	0.76	0.84	0.83
sensitivity:	0.71	0.52	0.57	0.6	0.84	0.75
specificy:	0.71	0.52	0.57	0.6	0.84	0.75
accuracy:	69%	57%	59%	64%	84%	78%
F1 measure:	0.68	0.56	0.57	0.55	0.85	0.76

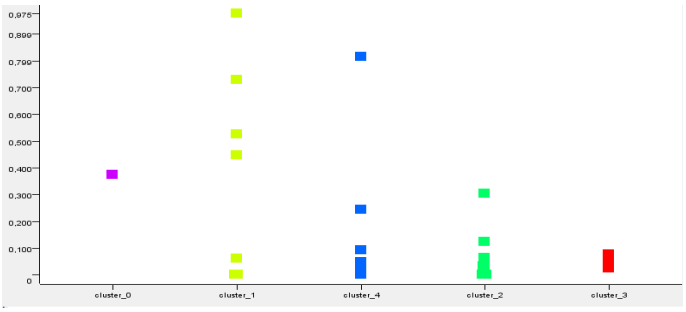
Tablica 5. Učinak pojedinog klasifikatora

Ostalo je napraviti još klasteriranje skupa podataka. Obzirom da u platformi KNIME nema implementirana niti jedna metoda za računanje unutarnje kvalitete klasteriranja (poput DB indexa ili C-indexa, *internal cluster validation index*), te da dani skup podataka nije bio ni na koji način određen (npr. po vrsti piva, cijeni piva, zemlji iz koje dolazi...) odlučili smo pokrenuti program na manjem skupu podataka koji ćemo moći "na ruke" provjeriti koliko je smislen (nekoliko pokušaja klasteriranja većeg skupa prošlo je vrlo loše, pogotovo metodom TF*IDF, gdje je 99% pivi smješteno u jedan klaster).

Klasteriranje smo radili na nekoliko skupova podataka (npr. samo najbolji i najgori tekstovi, samo najbolji tekstovi, tekstovi koji opisuju uglavnom tamnije pive), ali niti na jednom skupu nismo dobili nikakve smislene podjele. Ovdje navodimo samo jedan od pokušaja, na skupu najboljih tekstova ocijenjenih ocjenom većom od 4.0. Koristili smo skup od 15 078 članaka koji su opisivali 133 različita piva, te smo već opisanim postupkom od danog skupa napravili vektorski prostor (pomoću Keygraph ekstraktora, TF*IDF ekstraktora i χ^2 ekstraktora). Keygraph ekstraktor opet se pokazao najboljim, ostale dvije metode gotovo sve primjerke grupirale su u jedan veći klaster. Ideja je sada bila sve tekstove grupirati ovisno o imenu pive, te za svaki vektor gledati srednju vrijednost njegovih bodova (*eng. score*) u tekstu. Nadali smo se da se kod tamnijih piva češće pojavljuju riječi poput "dark", "black", "strong" i sl., ali dani rezultati nisu to potvrdili. Varirajući broj klastera k nismo dobili nikakve značajnije rezultate. Na idućoj slici stavljamo jedan primjer klasteriranja, koji nije dao nikakvu informaciju: na osi x nalaze se oznake klastera, a na osi y bodovi riječi "white" za svaki od klastera. Iako nekakva podjela postoji, provjerivši dobivene rezultate dobili smo da se u svakom klasteru nalaze i tamnija i svijetlija, i bolja i lošija, i skuplja i jeftinija, i američka i neamerička piva. Provjerivši rezultate u literaturi poput [2], vidjeli smo da je većina rezultata dobivena na temelju usporedbe sa već unaprijed označenim (labeliranim) skupom, što mi nismo bili u mogućnosti napraviti.

V. Zaključak

U ovom radu naglasak je stavljen na metode ekstrakcije ključnih riječi iz teksta, te njihovu efikasnost na algoritmima za klasifikaciju i klasteriranje većeg skupa podataka. TF*IDF metoda koja se najčešće spominje u kontekstu analize teksta (barem prema autorovom dosadašnjem iskustvu) daje slabije



Slika 6. Pojavljivanje riječi *white* po klasterima

rezultate od preostale dvije metode, ali također i drugačije rezultate - pogledamo li tablicu 3, vidimo da χ^2 i keygraf ekstraktori daju riječi koje opisuju piva, dok TF*IDF kao riječi s najvećom težinom daju one koje najviše odudaraju od očekivanih - u tekstovima s dobrim ocjenama to su bili worst, crap, fucking... Obzirom na konstrukciju TF*IDF mjere to je i očekivano, ali zato ova metoda u ovom kontekstu možda i nije najbolja. Naime, na stranicama poput *www.ratebeer.com* pridjevi koji opisuju dobre proizvode vrlo su slični, pa traženje riječi koja odudaraju od ostalih često može biti pogrešno. SVM se pokazao kao najuspješnija metoda za klasifikaciju teksta, pogotovo kombiniran s keygraph metodom ekstrakcije riječi. Zanimljiva je i činjenica da χ^2 metoda, iako daje dvostruko veću dimenziju prostora od TF*IDF, ne daje znatno bolje rezultate. Uzrok ovom ponašanju autor nije otkrio. Nažalost, zbog ograničenih mogućnosti platforme KNIME i danih podataka koje smo imali, *K-means* algoritam nije davao nikakve interpretabilne odgovore na pitanja vezana uz klasifikaciju teksta.

KNIME (Konstanz Information Miner) je vrlo dobar program za obradu i analizu manje količine teksta. Nažalost, već za skup od 20000 recenzija postaje gotovo nemoguće koristiti program, čak i na jačim računalima (8 GBram, 64-bit procesor). Stoga mislimo da je za analizu zaista velikih skupova podataka potrebno koristiti programe koji imaju direktniji pristup memoriji računala (Python, Matlab). Nadalje, za davanje preciznijeg, potpunijeg i općenitijeg odgovora na pitanje "Koje je najbolje pivo na svijetu?" bilo bi od velike pomoći raditi na informativnijem skupu podataka na kojemu bi mogli učiti. Neke od korisnih informacija kojima autor nije raspolagao bile bi iskustvo korisnika pri ocjenjivanju piva, kao u [3], cijeli "životni ciklus" korisnika na stranici kao u [1], informacija o tome što korisnik opisuje (miris, okus, izgled, konačni dojam) u pojedinim rečenicama, kao u [4] (ova ideju koristili su pobjednički timovi na natjecanju) ili označen skup u kojemu su određene različite vrste piva (tamno, svijetlo, pšenično, IPA, stout, porter...), kako bi mogli točnije ocjenjivati atribute za pojedine klase piva.

Literatura

[1] J. Leskovec, C. Potts, C. D. N. Mizil, R. West, D. Jurafsky, *No Counrtry for Old Members: User Lifecycle and Linguistic Change in ONline Communities*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

[2] S. Ezzat, M. N. El Gayar, M. M. Ghanem, *Sentiment Analysis of Call Centre Audio Conversations using Text Classification*, 2012.

[3] J. McAuley, J. Leskovec, *From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews*, Standford University

[4] J. McAuley, J. Leskovec, D. Jurafsky, *Learning Attitudes and Attributes from Multi-Aspect Reviews*, Standford University

[5] Masao, Nakada, Yuko, *Document clustering based of similarity of subjects using integrated subject graph*, 2006.

[6] Yutaka Matsuo i Mitisuru Ishizuka, *Keyword extraction from a Single Document using Word Cooccurrence Statistical Information*, 2003.

VI. DODATAK

Kako je cijeli projekt započo pitanjem "Koje je najbolje pivo na svijetu?" u sklopu natjecanja Mozgalo, u ovom dodatku kratko spominjemo naš rezultat u vezi tog projeka. Nismo htjeli detaljnije ulaziti u to obzirom da nema prevelike veze sa strojnim učenjem, pa smo ga smatrali irevelantnim za ovaj rad. Zbog ograničenosti skupa podataka, našeg nedovoljnog programerskog znanja i ograničenih mogućnosti korištenih računala, uspjeli smo napraviti samo jednostavni dictionary model, odnosno model koji se bazira na riječniku u kojem smo označili koje su nam pozitivne, a koje negativne riječi u kontekstu opisa piva.

Kako bi uopće mogli pokrenuti program, trebali smo reducirati skup od 172 000 tekstova na skup od 20 000 tekstova, što je bila gornja ograda količine podataka koju smo mogli učitati. Stoga je bilo potrebno uvesti neke radikalnije pretpostavke kako bi mogli doći do nekog smislenog zaključka o tome koje je najbolje pivo na svijetu. Prvo smo odlučili promatrati samo tekstove koji su ocijenjeni ocjenom većom od 3.9. Ovo smo odlučili uz pretpostavku da je najbolja piva ona koja će rijetko (gotovo nikada) biti ocijenjena prosječnom ili ispodprosječnom ocjenom. Dodatno opravdanje za ovaj potez bila je i činjenica da tražimo jedno najbolje pivo na svijetu, a ne skup od više njih, pa vjerujemo da su samo izuzetno dobra piva mogla biti u ovoj klasi. Ohrabrio nas je i već spomenuti članak [3], u kojemu je primijećeno da najboljim pivima iskusniji korisnici uvijek daju iznimno dobre ocjene, dok im i noviji daju iznadprosječne ocjene ili prosječne ocjene, iako možda nisu potpuno naviknuti na jače okuse. Zatim smo izračunali srednju vrijednost ocjena preostalih piva i primjetili da ih je velik broj ocijenjen ocjenom blizu 5.0, iako je njihova stvarna ocjena (težinska ocjena sa stranice *www.ratebeer.com*) puno niža. Naime, znalo se dogoditi da neke pive imaju svega nekoliko recenzija u vrijeme skupljanja podataka (ili da smo dio obirsali u procesu čišćenja baze) i ostalo je samo par previše subjektivnih članaka. Iz tog razloga odlučili smo izbrisati sve članke koji su imali manje od 10 recenzija i dobiveni rezultati bili su u skladu sa očekivanima.

U ovom trenutku imali smo bazu od 40716 recenzija 629 različitih piva. Također, koristili smo rječnik od 20 riječi koje smo odabrali prema uputama sa stranice www.bjcp.com. Ovu stranicu napravili su profesionalni suci za ocjenjivanje piva (bjcp znači Beer Judges Certificate Program, odnosno program kojim se stječe titula sudca za pivu). Sve riječi označili smo kao pozitivne, a koje su se točno u njemu nalazile može se vidjeti na Slici 7. Preptostavili smo (opet malo drastičnije, ali iz informacija koje smo dobili iz brojnih članaka poput [3],[4], [1]) da će se među najboljim pivama naći one jake i intenzivne (pa smo uzimali i takve riječi), te da se u tekstovima neće nalaziti previše negativnih riječi nego da će se autori bazirati na opis same pive. Na ovom skupu sada smo napravili standardni proces za analizu teksta, označili smo ključne riječi, napravili BoW model, izračunali frekvencije te potom gledali uz koje se pive vežu najpozitivniji dojmovi. Za konačni rezultat dobili smo da je "najbolja" piva Hoppin' Frog DORIS The Destroyer, tamno pivo (Russian Imperial Stout) koje potječe iz SAD-a. To je, obzirom na sve pretpostavke, očekivano rješenje - kao što smo već spomenuli, iskusniji korisnici preferiraju tamnije i jače pive, a vjerojatno je i činjenica da pivo potječe iz SAD-a (od kud je većina korisnika stranice RateBeer) pomoglo u rastu njegove popularnosti.