Topic modeling pomoću LDA i primjene na klasifikaciju i klasteriranje

Marija Majda Perišić
Prirodoslovno-matematički fakultet
Matematički odsjek
Zagreb, Hrvatska

Tomislav Droždjek
Prirodoslovno-matematički fakultet
Matematički odsjek
Zagreb, Hrvatska

***Abstract*—U ovom radu, primjenom LDA metode, odredit ćemo izgled tema te distribuciju istih unutar danih dokumenata. Ukratko ćemo objasniti ideje na kojima se temelji LDA model. Nakon toga, provest ćemo eksperiment klasteriranja dokumenata, a potom, na problemu klasifikacije, usporediti kvalitetu separiranja grupa dokumenata na osnovi LDA značajki naspram vreće riječi.**

I. Uvod i opis problema

Topic modeling, odnosno modeliranje zastupljenosti tema u nekom dokumentu postalo je vrlo popularno područje u zadnjih nekoliko godina.[1] Očito je da se porastom broja dostupnih dokumenata pojavljuje potreba da im se, koristeći računalo, brzo i efikasno odredi tema. Također, izuzetno je korisno skup dokumenata klasificirati s obzirom na prevladavajuću temu. Upravo to su problemi kojima ćemo se baviti u ovom radu.

Cilj ovog rada je, koristeći metodu Latent Dirichlet Allocation (u daljnjem tekstu LDA), naći distribuciju tema nad pojedinim dokumentom (udio svake teme u tom dokumentu), gdje je tema distribucija nad fiksnim rječnikom. Nadalje, tako dobivene informacije pokušat ćemo iskoristiti kako bi dokumente klasificirali, te dobivene rezultate usporediti s rezultatima dobivenim koristeći "bag of words" model. Na podacima koji nisu unaprijed labelirani po temama provest ćemo klasteriranje kako bismo grupirali dokumente kojima pripadaju slični LDA vektori, te ih time na određeni način povezali po sličnosti tema u njima.

II. LDA model

LDA je generativni vjerojatnosni model kojim opisujemo proces generiranja skupa tekstualnih dokumenata. Model pretpostavlja da imamo K tema, pri čemu je K fiksan broj, koje opisuju cijeli skup dokumenata. Pod temom podrazumijevamo Dirichletovu distribuciju nad fiksnim rječnikom od V riječi. Sve teme označene su sa Φ, $\Phi \in M_{K\times V}$, a pojedine teme sa $\vec{\phi_k}$, $\vec{\phi_k} \sim Dir(\vec{\beta}\ )$. Ukupan broj dokumenata označavamo sa M, a broj riječi u dokumentu m sa N_m.

Naš model pretpostavlja da su dokumenti nastali sljedećim postupkom:

- Za svaki dokument odabiremo mješavinu tema zastupljenih u njemu. Preciznije, za m-ti dokument generiramo σ_m, $\sigma_m \sim Dir(\vec{\alpha})$.
- Iz odabrane mješavine tema za svaki dokument biramo temu $z_{m,n}$, $z_{m,n} \sim Multi(\vec{\sigma_m})$.
- Iz odabrane teme odaberemo riječ $w_{m,n}$, pri čemu $w_{m,n} \sim Multi(\vec{\phi}_{z_{m,n}})$.

pri čemu posljednja dva koraka ponavljamo dok ne dođemo do željenog broja riječi N_m, za dani dokument, a čitav postupak ponavljamo za svih M dokumenata.

Upravo opisan postupak prikazan je na slici 1. Krugovi na slici predstavljaju slučajne varijable, dvostruki krug je opažena slučajna varijabla, dok pravokutnici označavaju ponavljanje.

Iz grafičkog prikaza evidentno vrijedi:

$$p(\Phi, \Theta, \vec{z}, \vec{w}|\alpha, \beta) =$$

$$p(\Phi|\beta) \prod_{m=1}^{M} p(\vec{\theta_m}|\alpha) \prod_{n=1}^{N_m} p(z_{m,n}|\vec{\theta_m}) p(w_{m,n}|z_{m,n}, \Phi) =$$

$$\prod_{k=1}^{K} p(\vec{\phi_k}|\beta) \prod_{m=1}^{M} p(\vec{\theta_m}|\alpha) \prod_{n=1}^{N_m} p(z_{m,n}|\vec{\theta_m}) p(w_{m,n}|\vec{\phi}_{z_{m,n}})$$

pri čemu koristimo svojstvo Bayesove mreže (pretpostavka da su čvorovi uvjetno nezavisni uz danog roditelja).

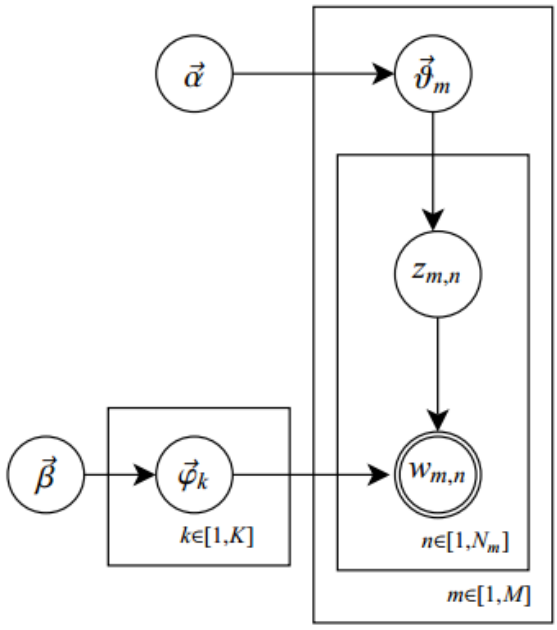

Fig. 1. Generiranje dokumenata za LDA

Cilj nam je, na osnovi opaženih varijabli modela (riječi), doći do podataka o skrivenim varijablama, tj. naći teme te mješavinu tema za svaki dokument, što možemo napraviti tako da prvo izračunamo distribuciju $p(\vec{z}|\vec{w})$, odnosno vjerojatnost tema pojedine riječi.

Preciznije, kako bismo procijenili distribuciju $p(\vec{z}|\vec{w})$, koristimo Gibbsovo uzorkovanje. Više o Gibbsovom uzorkovanju u [2]. Kako je, zbog svojstva Dirichletove distribucije da svaka tema može generirati svaku riječ, Markovljev lanac za $p(\vec{z}|\vec{w})$ nastao Gibbsovim uzorkovanjem regularan, postoji jedinstvena distribucija u koju on konvergira za svako početno stanje, a to je upravo $p(\vec{z}|\vec{w})$. Budući da je Dirichletova distribucija konjugatna apriorna distribucija za multinomnu distribuciju, nakon što procijenimo vrijednosti $\vec{w}$ i $\vec{z}$ možemo izračunati mješavine tema $\vec{\sigma_m}$ i teme $\vec{\phi_k}$ na sljedeći način

$$p(\vec{\theta}_m|\vec{z}_m,\vec{\alpha}) = Dir(\vec{\theta}_m|\alpha+\vec{c}_m)$$

$$p(\vec{\phi}_k|\vec{w}_m,\vec{z}_m,\vec{\beta}) = Dir(\vec{\phi}_k|\vec{\beta}+\vec{n}_k)$$

Napomena 1. *Za raspodjelu Φ nad familijom raspodjela θ kažemo da je konjugatna apriorna distribucija za θ (eng. conjugate prior) ako za opservacije X distribucija $p(\theta|X)$ ima isti funkcijski oblik kao raspodjela $p(\theta|\phi)$ te parametre koji uključuju X. X je skup opservacija dobiven na temelju raspodjele θ. U slučaju Dirichetove i multinomne distribucije, ako je $\vec{\theta} \sim Dir(\vec{\alpha})$ i $X \sim Multi(\vec{\theta})$, vrijedi:*

$$p(\vec{\theta}|X) \sim Dir(\vec{\alpha}+\vec{n})$$

gdje je $\vec{n} = (n_1, n_2, \ldots, n_K)$, n_k je broj pojavljivanja broja k u uzorku.

Napomena 2. *Ako je $\vec{x} \sim Dir(\vec{\alpha})$ onda očekivanje dimenzije x_i jest:*

$$E[x_i|\vec{\alpha}] = \frac{\alpha_i}{\sum_{k=1}^{K}\alpha_k}$$

Iz formule za očekivanje Dirichletove distribucije slijedi

$$\theta_{m,k} = \frac{\alpha_k + c_m^k}{\sum_{l=1}^{K}\alpha_l + c_m^l}$$

$$\phi_{k,t} = \frac{\beta_t + n_k^t}{\sum_{v=1}^{V}\beta_v + n_k^v}$$

III. Primjene

A. Klasifikacija

1) Opis problema: U ovom poglavlju pokušat ćemo eksperimentom dati odgovor na pitanje koliko dobro možemo klasificirati tekstualne dokumente na osnovi značajki iz vektora mješavina tema $\vec{\theta}_m$.

2) Skup podataka i pretprocesiranje: Eksperiment je proveden na bazi od 20000 dokumenata koji su podijeljeni u 20 jednakih grupa te labelirani. Točnije, radi se o standardnoj bazi 20-newsgroups[1] koja sadrži poruke sa 20 Usenet grupa. Tablica I prikazuje imena i teme tih 20 grupa.

Baza je pretprocesirana na sljedeći način:

1) Brisanje metapodataka.
2) Micanje potpisa na kraju poruke.
3) Micanje URL-ova.
4) Brisanje nealfabetskih znakova.
5) Pretvaranje svih slova u mala slova.
6) Micanje stop-riječi.

Za prva 4 koraka koristili smo PowerGrep [2], alfabetske znakove smo u mala slova pretvorili korištenjem Malleta [3], kao i micanje stop-riječi.

[1] http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html
[2] http://www.powergrep.com
[3] http://mallet.cs.umass.edu/

grupa	tema
alt.atheism	ateizam
comp.graphics	grafika
comp.os.ms-windows.misc	Microsoft Windows
comp.sys.ibm.pc.hardware	PC hardware
comp.sys.mac.hardware	Macintosh hardware
comp.windows.x	X Window sustav
sci.crypt	kriptografija
misc.forsale	prodaja
rec.autos	automobili
rec.motorcycles	motori
rec.sport.baseball	bejzbol
rec.sport.hockey	hokej
sci.electronics	elektronika
sci.med	medicina
sci.space	svemir
soc.religion.christian	kršćanstvo
talk.politics.guns	politika o vatrenom oružju
talk.politics.mideast	Bliski istok
talk.politics.misc	politika
talk.religion.misc	religija

TABLE I. BAZA 20-NEWSGROUPS I TEME ZA SVAKU GRUPU

Baseball	year hit run won baseball player good ball average league run lost braves morris team
Homoseksualnost	men sex homosexual women homosexuality people sexual love gay male show hate marriage behavior wife
Oružje	law state rights gun goverment police laws weapons court guns legal bill control amendment constitution
Svemir	space earth nasa launch orbit moon shuttle mission satellite solar planet spacecraft system sky flight
Konverzacija	don doesn give didn isn ve wouldn big thing remember back hell guy guess wrong
Kršćanstvo	god jesus church christ bible lord sin man faith christian love father heaven spirit hell

TABLE II. NEKE TEME DOBIVENE LDA MODELOM I 15 VODEĆIH RIJEČI ZA SVAKU TEMU

3) Konstrukcija LDA modela: LDA model konstruirat ćemo pomoću softverskog paketa Mallet, koristeći Gibbsovo uzorkovanje opisano u poglavlju 2. Hiperparametri algoritma postavljeni su na $\alpha = 50/K$ i $\beta = 0.01$. Broj iteracija postavljen je na 10000. Konstruiran je model sa $K = 50$ tema. Vodeće riječi za neke izabrane teme nalaze se u tablici II.

Primijetimo da neke teme već poprilično dobro aproksimiraju početne teme (npr. svemir, oružje, baseball), neke teme se mogu pridružiti više kategorija, dok se neke sastoje od općenitih konverzacijskih pojmova.

4) Klasifikacija: Za svaku grupu postavili smo problem binarne klasifikacije: u pozitivnoj klasi nalaze se svi dokumenti iz dane grupe, a u negativnoj dokumenti iz ostalih grupa. Tako, za svaku grupu, dobivamo približno 19000 podataka u negativnoj i 1000 podataka u pozitivnoj klasi. Kako su klase neuravnotežene, F1 predstavlja bolju mjeru uspješnosti od točnosti te stoga želimo naučiti model koji će dati što bolju F1 mjeru na temelju LDA značajki. Uspoređujemo dobivene rezultate s rezultatima dobivenim za model čije značajke su vreća riječi, dakle vektor kojem koordinate odgovaraju broju pojavljivanja odgovarajuće riječi u dokumentu. Kao algoritam za klasifikaciju odabrali smo Naive Bayes, točnije implementaciju Naive Bayesa u KNIME-u[4].

grupa	LDA-F1	vreća riječi - F1
alt.atheism	0.421	0.658
comp.graphics	0.525	0.649
comp.os.ms-windows.misc	0.511	0.793
comp.sys.ibm.pc.hardware	0.398	0.788
comp.sys.mac.hardware	0.365	0.798
comp.windows.x	0.64	0.747
misc.forsale	0.493	0.878
rec.autos	0.723	0.916
rec.motorcycles	0.843	0.910
rec.sport.baseball	0.879	0.935
rec.sport.hockey	0.892	0.910
sci.crypt	0.792	0.947
sci.electronics	0.481	0.84
sci.med	0.699	0.934
sci.space	0.621	0.928
soc.religion.christian	0.553	0.93
talk.politics.guns	0.460	0.82
talk.politics.mideast	0.742	0.89
talk.politics.misc	0.447	0.673
talk.religion.misc	0.335	0.61

TABLE III. BAZA 20-NEWSGROUPS I REZULTATI

Tablica III prikazuje rezultate klasifikacije. Preciznije, F1 mjeru za LDA model i vreću riječi po svim grupama. Vidljivo je da je bag of words model mnogo bolje separirao teme, što je i bilo očekivano budući da sadrži mnogo veći broj značajki od LDA modela. Uočimo da su najniže F1 mjere dobivene za teme ateizam i religija, odnosno PC hardware i Macintosh hardware. Naime, zbog bliskosti tematike te sličnih riječi korištenih u dokumentima iz navedenih grupa, teško ih je pravilno odijeliti. No, za klase koje su dobro separirane od ostalih, i LDA vektori omogućuju dobro razdvajanje klasa.

Sljedeći eksperiment proveli smo da bi provjerili koliko dobro LDA vektori separiraju klase kad su one dobro odvojene jedna od druge. Tablicu I preradili smo na način da smo spojili grupe sa sličnom tematikom. Konkretno, spojili smo sve grupe comp.* u grupu Računala, talk.politics.* u grupu Politika, talk.religion.christian, talk.religion.misc i alt.atheism u grupu Religija, rec.autos i rec.motorcycles postali su grupa Automoto, a rec.sport.baseball i rec.sport.hockey Sport. U Tablici IV nalaze se dobiveni rezultati.

Dobiveni rezultati potvrđuju našu slutnju, F1 mjera za LDA metodu kod dobro separiranih klasa je usporediva s F1 mjerom dobivenom za vreću riječi. Također valja napomenuti da je velika prednost LDA

[4]http://www.knime.org

grupa	LDA-F1	vreća riječi - F1
Računala	0.83	0..886
sci.crypt	0.769	0.843
misc.forsale	0.47	0.567
Automoto	0.845	0.904
Sport	0.94	0.967
sci.electronics	0.477	0.608
sci.med	0.743	0.875
sci.space	0.616	0.864
Politika	0.802	0.842
Religija	0.81	0.884

TABLE IV. PRERAĐENA BAZA 20-NEWSGROUPS S DOBIVENIM REZULTATIMA

metode veća brzina izvršavanja klasifikacije zbog puno manje dimenzionalnosti, pa je ponekad, kod dobro odvojenih tema, LDA čak i bolji izbor od obične vreće riječi, unatoč malo manjoj točnosti.

B. Klasteriranje

1) Opis problema: Ideja klasteriranja je podijeliti dokumente na klastere na osnovi slične tematike. Taj postupak može se koristiti kad nam prioritet nije naći dominantnu temu u dokumentu, nego grupirati one dokumente koji imaju sličnu raspodjelu po svim temama.

2) Skup podataka i pretprocesiranje: Eksperiment je proveden na bazi od 2247 članaka Associated Pressa, preuzetu sa [5]. Pretprocesiranje podataka sastojalo se od sljedećih koraka:

1) Maknuli smo HTML tagove koji su se pojavljivali u dokumentu koristeći PowerGrep.
2) Obrisali smo prazne redove (kako ih ne bi očitali kao dokumente kod LDA). Također smo koristili PowerGrep.
3) Pretvaranje svih slova u mala slova. To radimo u Malletu.
4) Obrisali smo standardne stop-riječi iz engleskog jezika. Taj postupak napravili smo također napravili u Malletu prije samog LDA.

3) Konstrukcija LDA modela: LDA model konstruirat ćemo pomoću softverskog paketa Mallet, koristeći Gibbsovo uzorkovanje opisano u poglavlju 2. Hiperparametri algoritma postavljeni su na $\alpha = 50/K$ i $\beta = 0.01$. Broj iteracija postavljen je na 10000. Konstruiran je model sa $K = 50$ tema. Vodećih 15 riječi iz odabranih tema prikazane su u tablici V. Vidljivo je da su riječi smisleno grupirane u teme.

[5]http://www.cs.princeton.edu/ blei/lda-c/ap.tgz

Rat u Iraku	iraq military united kuwait iran war troop bush president force today invasion oil invasion crisis
Medicina	health medicine aids drug hospital care research disease dr treatment blood doctor patient heart cancer
Pravo	court law judge decision case federal legal state order supreme ruling right case issue civil
Politika	house committee bill senate congress reagan rep sen chairman white legislation member president defense administration
Televizija	television show tv film network abc nbc cbs movie series time news broadcast story week
Svemir	space project shuttle mission launch earth nasa test complete site scientist find facility million ground

TABLE V. PRIMJERI NEKIH TEMA I NAJVJEROJATNIJIH RIJEČI DOBIVENIH IZ MODELA SA 50 TEMA.

4) Klasteriranje: Klasteriranje smo proveli pomoću k-means algoritma kojeg smo proveli na LDA vektorima dobivenim gore opisanim postupkom. Detalji o k-means algoritmu mogu se naći na [3]. Klasterirali smo podatke na 4 klastera.

5) Rezultati: U ovom poglavlju prikazat ćemo neke dokumente koji su najbliži centroidama svakog od dobivenih klastera. Ti bi dokumenti trebali biti bliski po tematici. Zbog sažetosti, za svaki pojedini klaster ćemo prikazati samo prvih nekoliko rečenica odgovarajućih dokumenta.

Vidljivo je da su članci poprilično slični, te da se klastering u načelu može koristiti za grupiranje sličnih dokumenata, pa čak i kao primitivni recommender sustav za slične članke. Međutim, treba imati na umu ograničenja ove metode. Eksperimenti su pokazali da je k-means LDA vektora dosta osjetljiv na različite inicijalizacije. Također, ako se dosta odmaknemo od centroide dokumenti čiji se LDA vektori nalaze blizu ruba pripadnog klastera bit će sličniji dokumentima iz drugog, njemu susjednog, klastera.

IV. ZAKLJUČAK

Proučavajući problem klasifikacije LDA modelom dobili smo rezultate koji, iako lošiji od modela vreće riječi, solidno separiraju članke u teme. Naravno, kako vreća riječi koristi mnogo više značajki od LDA, takav rezultat je bio i očekivan. No, kod skupa podataka gdje su teme dobro odijeljene (drugi dio eksperimenta) F1 mjera za LDA model se približava F1 mjeri za vreću riječi. Dakle, kod problema odvajanja dokumenata čije teme su jasno odijeljene svakako valja razmotriti korištenje LDA, budući da koristi mnogo manje računalnih resursa poput memorije te je zamjetno brži od modela s vrećom riječi. Još jedan eksperiment s usporedbom LDA i vreće riječi proveden je u [4], s razlikom što je za klasifikaciju korišten model s potpornim vektorima, te su u konačnici dobiveni rezultati slični našima.

Klaster 1
Share prices took the largest drop of the year today on the Tokyo Stock Exchange, and the dollar fell sharply against the Japanese yen. The Nikkei Stock Average of 225 selected issues, a 251.67-point loser... The dollar surged against the Japanese yen Monday, and stock prices were sharply lower due to concern over higher oil prices and the possibility of a U.S. interest rate increase. At the end of trading, the dollar was quoted at 122.00 yen... Share prices on the London Stock Exchange were little changed Thursday amid some profit-taking and uncertainty over the market's direction...
Klaster 2
Czechoslovakia began talks today with the Soviet Union on the withdrawal of about 75,000 Soviet troops, and one source said Czechoslovak officials want at least half of the force out by May. On the eve of the talks, about 6,000 people protested... South Korea launched regular ferry service with China today for the first time since the Korean peninsula was divided in 1945. A 4,300-ton... Thousands of Czechs joined U.S. veterans Saturday for a sun-kissed streetfest in the first celebration ever of the liberation of this beer-brewing city by American troops in World War II...
Klaster 3
Here are highlights of Thursday's session of the 28th Communist Party congress, its 10th day. The congress is scheduled to continue Friday. Here is a look at the accomplishments of the 28th Communist Party congress after eight days, as well as what remains to be done: ACCOMPLISHMENTS: Mikhail S. Gorbachev was re-elected as party leader Tuesday in a one-sided race... Highlights of Wednesday's session of the 28th Communist Party congress, in its third day:
Klaster 4
Here's a sampling of what newspapers abroad are saying about the Democratic National Convention: Here are the latest, unofficial results in the Republican races for the U.S Senate.

TABLE VI. PRIMJERI DOKUMENATA IZ SVAKOG KLASTERA

Na nelabeliranom skupu dokumenata proveli smo klasteriranje s ciljem grupiranja onih koji su tematski bliski. Dobiveni su smisleni rezultati, no valja naglasiti nedostatke navedene metode. Naime, teško je ocijeniti koji bi bio odgovarajući broj tema te koliko klastera uzeti. Također, dolazi do situacije da su dva dokumenta međusobno bliska, no razvrstani su u različite klastere (nalaze se na granicama istih). O korisnikovim preferencijama i željenim rezultatima ovisi kako odabrati prikladan broj klastera te kako mjeriti uspješnost algoritma.

REFERENCES

[1] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, January 2003.

[2] G.Cassela, E. I. George. Explaining the Gibbs Sampler. The American Statistician, 46:167-174, December 1990.

[3] http://web.math.pmf.unizg.hr/nastava/su/materijali/

[4] D. Korenčić. Statističke metode za dubinsku obradu podataka: Latentna Dirichletova alokacija.

[5] . Steyvers, T. Griffiths. Probabilistic Topic Models. Handbook of latent semantic analysis, 427(7):424-440,2007.