Metode modeliranja tema — latentna Dirichletova alokacija

Andrej Ciganj
Matematički odsjek
Prirodoslovno-matematički fakultet
Sveučilište u Zagrebu
e-mail: aciganj@student.math.hr

Sanjin Ružić
Matematički odsjek
Prirodoslovno-matematički fakultet
Sveučilište u Zagrebu
e-mail: sruzic@student.math.hr

***Sažetak*—U radu opisujemo jedan model modeliranja tema — latentnu dirichletovu alokaciju (LDA): generativni probabilistički model za kolekcije diskretnih podataka kao što su korpusi tekstualnih dokumenata. Više pažnje je posvećeno matematičkoj pozadini LDA i inačicama modela kao što su nadzirani LDA i dinamički LDA. Postojeće implementacije tih modela u C-u i C++-u smo isprobali na unaprijed pripremljenim podacima. Što se tiče dobivenih rezultata, osnovni LDA kao nenadzirani model bilo je teže evaluirati tako da je ovdje fokus bio na rezultatima nadziranog LDA.**

I. Uvod

A. Motivacija

Živimo u svijetu koji se neprestano mijenja i sve brže razvija. To se posebice odnosi na elektroničku tehnologiju za koju bismo s pravom mogli reći da predstavlja kotač zamašnjak razvoja našeg društva. S pojavom Interneta, a posebice nakon njegove sveopće rasprostranjenosti, količina informacija koju proizvodimo je nezamisliva. Novosti, časopisi, knjige, znanstveni članci, video i audio zapisi, ... Problem više nije prenijeti informaciju, nego doći baš do one koju tražimo. Stoga je postalo od iznimne važnosti moći se snalaziti u toj ogromnoj količini podataka. Potrebno je osigurati način kojim bismo mogli organizirati, pretraživati i razumjeti spremljene podatke. Jasno je da zbog samog obima podataka to ne može čovjek napraviti "na ruke". Zato je potrebno razviti modele i algoritme koji će mu pomoći u ostvarenju tog cilja.

Neki od njih su pokazali veliku popularnost. To su prije svega PageRank kao temelj Googleove tražilice. On se temelji na poveznicama među dokumentima koje možemo najjednostavnije shvatiti kao glasovanje (ako stranica p ima link koji vodi na stranicu r, tada stranica p daje svoj glas za stranicu r). Pomoću toga možemo odrediti koja je stranica "najvažnija" za dani upit. Međutim, usprkos velikom uspjehu kojeg je postigao taj pristup, mogli bismo mu tražiti manu u činjenici da ne pokušava pronaći sakrivenu tematsku podlogu pojedinih dokumenata. U tom smjeru se razvijala teorija vjerojatnosnog modeliranja tema (engl. *probabilistic topic modelinga*), odnosno algoritama koji pokušavaju otkriti sakrivenu tematiku velike kolekcije dokumenata. Algoritmi modeliranja tema su statističke metode koje analiziraju riječi pojedinih dokumenata kako bi otkrile koje teme prožimaju dokument, te njihovu međusobnu povezanost i razvoj kroz vrijeme. Oni ne zahtijevaju nikakvo prethodno označavanje podataka, već samo iz originalnih dokumenata otkrivaju tematske pozadine istih.

B. Ciljevi

Ovim projektom htjeli smo se prije svega upoznati sa još jednim oblikom strojnog učenja, a to je, naravno, problem modeliranja tema. On nam se činio zanimljivim iz više razloga.

Prije svega radi se o jednom relativno novijem modelu koji ima niz otvorenih pitanja i ima sve veću primjenu na razna područja. Osim toga, kako smo se na predavanjima uglavnom bavili algoritmima nadziranog učenja, htjeli smo naučiti nešto i o nenadziranom učenju - problemima koji se kod njega javljaju te načinima kako ih se može uspješno savladati.

Shodno našim ciljevima, naglasak smo stavili na proučavanje teorije, tj. matematičke pozadine samih algoritama. Smatramo kako je vrlo važno razumjeti zašto algoritmi (uspješno) rade te imati dobru intuiciju o njima (npr. kako odabrati parametre modela). Ipak, teorija ne bi trebala sama sebi biti svrha, pa smo tako isprobali neke od proučenih algoritama te se uvjerili da oni doista daju upotrebljive rezultate.

II. Razvoj modeliranja tema

Cilj modeliranja tema je nalaženje skraćenog opisa dokumenata u kolekciji iz kojeg bi se izvukle bitne značajke kao što su tema dokumenta i struktura kolekcije dokumenata.

Prva predložena metoda za rješavanje tog problema je tf-idf shema (*term frequency–inverse document frequency*) koja određuje važnost riječi za pojedini dokument u odnosu na ostale dokumente u kolekciji. *tf-idf vrijednost* riječi je omjer broja njenog pojavljivanja u dokumentu s brojem pojavljivanja u kolekciji (čime se rješava problem čestih riječi). Metoda tako reducira dokument na vektor realnih brojeva u kojem svaki broj predstavlja tf-idf vrijednost neke riječi iz unaprijed utvrđenog rječnika. Završni rezultat je matrica riječi po dokumentu, čiji stupci sadrže tf-idf vrijednosti za svaki dokument u kolekciji.

Nedostaci ovog pristupa su relativno malo smanjenje duljine opisa dokumenta i slabo otkrivanje statističke strukture[1]

[1] ne otkriva se semantička povezanost između dokumenata, kao niti između pojmova u rječniku

unutar i među dokumentima. Bolje rješenje, između ostalih, predloženo je LSI metodom (*latent semantic indexing*) koja se "nastavlja" na tf-idf. Zove se latentna zbog sposobnosti otkrivanja veza između semantički povezanih pojmova koji su skriveni(latentni) u kolekciji dokumenata. Pretraživanje dokumenata nad kojima je izvršena LSI metoda vraća rezultate sličnog značenja kriteriju pretraživanja čak i kad oni ne sadrže specifične riječi zahtijevane kriterijem. To se postiže korištenjem SVD dekompozicije na opisanoj tf-idf matrici kako bi se izdvojio potprostor tf-idf značajki s najvećom varijancom u kolekciji. Rezultat je uspostavljanje veze između riječi koje se koriste u sličnim kontekstima.

Iz LSI je proizašla metoda pLSI (*probabilistic LSI*) koja tretira svaku riječ u dokumentu kao uzorak iz *mixture* modela[2], gdje su komponente mixture modela slučajne varijable koje se mogu tumačiti kao reprezentacije tema. Generativni postupak je biranje dokumenta d iz kolekcije s vjerojatnošću $P(d)$, biranje skrivene teme t s vjerojatnošću $P(t|d)$ i generiranje riječi iz te teme s vjerojatnošću $P(w|t)$. Svaka riječ je tako generirana iz jedne teme, a različite riječi u dokumentu mogu biti generirane iz različitih tema. Svaki dokument je predstavljen kao distribucija na fiksiranom skupu tema koja predstavlja skraćeni opis dokumenta.

Problemi pLSI metode su sljedeći:

- broj parametara linearno raste s veličinom kolekcije,
- nije jasno kako dodijeliti distribuciju dokumentima izvan trening skupa.

Kako bi se riješili navedeni problemi predložen je LDA model kojeg detaljnije izlažemo u idućem odjeljku.

III. LDA

Najčešće korišteni model modeliranja tema je latentna Dirichletova alokacija (engl. *latent Dirichlet allocation, LDA*) koja je zapravo nenadzirani (engl. *unsupervised*) generativni, vjerojatnosni, grafički (engl. *probabilistic graphical*) model.

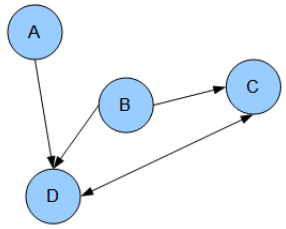
Ipak, prije upuštanja u detalje samog LDA modela, objasnit ćemo navedene atribute i uvesti potrebne definicije.

Za razliku od nadziranog učenja u kojem pojedini podatak za učenje ima jednu ili više značajki, te posebnu *oznaku* (na temelju koje radimo klasifikaciju ili regresiju), u nenadziranom učenju imamo samo značajke podataka među kojima pokušavamo pronaći sakrivenu strukturu.

Generativni model je model koji nasumično generira podatke na osnovu pretpostavljenih sakrivenih parametara. Za razliku od diskriminativnog modela koji može uzorkovati samo varijable cilja uvjetno na opažene varijable, generativni model je u stanju simulirati vrijednosti bilo koje varijable u modelu te je stoga puno fleksibilniji. Definira združenu vjerojatnosnu razdiobu na opaženim i sakrivenim slučajnim varijablama kako bi izračunao uvjetnu distribuciju sakrivenih varijabli uz dane opažene varijable. Ta uvjetna distribucija se još naziva i *posteriori* distribucija.

[2]u mixture modelu teme su predstavljene kao distribucije na podskupu rječnika (svaka riječ "dolazi" iz točno jedne teme)

Grafički modeli su vjerojatnosni modeli u kojima grafom označavamo uvjetne zavisnosti između slučajnih varijabli. Kako se ne bi crtalo svako ponavljanje slučajne varijable,



Slika 1. Primjer grafičkog modela. Svaki usmjereni brid označava uvjetnu zavisnost (npr. D je uvjetno zavisan o A).

definira se pravokutna notacija (engl. *plate notation*) u kojoj se pravokutnik koristi za grupiranje varijabli u podgraf koji se zajednički ponavlja (unutar pravokutnika se dodatno označava broj ponavljanja).

A. Dirichletova distribucija

Dirichletova distribucija, oznaka $Dir(\boldsymbol{\alpha})$, je naziv za familiju kontinuiranih vjerojatnosnih distribucija parametriziranu vektorom $\boldsymbol{\alpha}$ pozitivnih realnih brojeva. Sasvim konkretno, funkcija gustoće vjerojatnosti Dirichletove distribucije reda $K > 2$ s parametrima $\alpha_1, \ldots, \alpha_K > 0$ je dana s

$$f(x_1, \ldots, x_K; \alpha_1, \ldots, \alpha_K) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}, \qquad (1)$$

na otvorenom $(K-1)$-dimenzionalnom simpleksu definiranom s $x_1, \ldots, x_{K-1} > 0$, $x_K = 1 - x_1 - \cdots - x_{K-1}$. U izrazu (**??**) se javlja normalizirajuća konstanta $B(\boldsymbol{\alpha})$ koja je multinomijalna beta funkcija, a može se izraziti preko gama funkcije (koju intuitivno možemo shvatiti kao realno proširenje faktorijela)

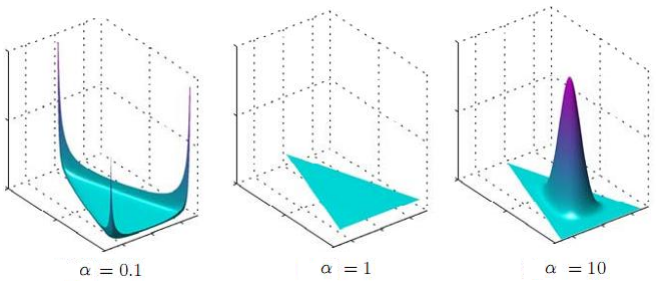
$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}.$$

Poseban slučaj Dirichletove distribucije koji se vrlo često javlja je tzv. *simetrična Dirichletova distribucija* u kojoj su svi elementi vektora $\boldsymbol{\alpha}$ međusobno jednaki. Stoga se ona može parametrizirati samo s jednom skalarnom vrijednošću α (parametar *koncentracije*), a funkcija gustoće se pojednostavljuje u

$$f(x_1, \ldots, x_K; \alpha) = \frac{\Gamma(\alpha K)}{\Gamma(\alpha)^K} \prod_{i=1}^{K} x_i^{\alpha - 1}.$$

Za LDA model će vrlo bitnu ulogu igrati upravo spomenuti parametar α te je zato od iznimne važnosti imati dobru intuiciju o njemu. Stoga ćemo dati nekoliko primjera kako α utječe na funkciju gustoće.

Za $\alpha = 1$, simetrična Dirichletova distribucija se zapravo svodi na uniformnu distribuciju na otvorenom $(K-1)$-dimenzionalnom simpleksu. Vrijednosti parametra koncentracije veće od 1 preferiraju guste, jednoliko distribuirane



Slika 2. Simetrična Dirichletova distribucija sa raznim vrijednostima parametra koncentracije.

razdiobe, drugim riječima, vrijednosti nekog primjerka će biti sve međusobno vrlo slične. Nasuprot tome, vrijednosti $\alpha < 1$ preferiraju rijetke distribucije, tj. one kod kojih će biti većina vrijednosti primjerka biti blizu nuli, a samo u nekoliko njih će biti glavnina "mase".

Dirichletova distribucija ima još jedno svojstvo koje će se pokazati veoma važnim za LDA model — ona je *konjugatna apriorna distribucija* (engl. conjugate prior) za multinomijalnu distribuciju.

Općenito, ako je aposteriori distribucija $p(\theta|x)$ u istoj familiji kao i apriori distribucija $p(\theta)$, tada kažemo da su one *konjugatne distribucije*, a apriori distribucija se naziva *konjugatna apriorna distribucija*. To svojstvo ponekad olakšava račun. Na primjer, zamislimo općeniti problem inferencije distribucije za parametara θ uz dane podatke x. Iz Bayesova teorema slijedi da je aposteriori distribucija jednaka

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta')p(\theta')d\theta'}. \tag{2}$$

Sada je jasno da će za različiti izbor apriorne distribucije $p(\theta)$ integral u (**??**) biti lakše ili teže izračunati. Konjugatna apriorna distribucija daje zatvorenu formu za aposteriornu distribuciju; u protivnom nužna bi bila (teška) numerička integracija.

U našem slučaju imamo $\theta \sim Dir(\alpha)$, $X \sim Multi(\theta)$ povlači

$$p(\theta|X) \sim Dir(\alpha + n).$$

Pri tome je $n = (n_1, \ldots, n_K)$, a n_k broj pojavljivanja broja k u uzorku. Intuitivno možemo shvatiti da aposteriori distribucija $p(\theta|X)$ ima veće vjerojatnosti na onim koordinatama koje odgovaraju većem broju opažanja.

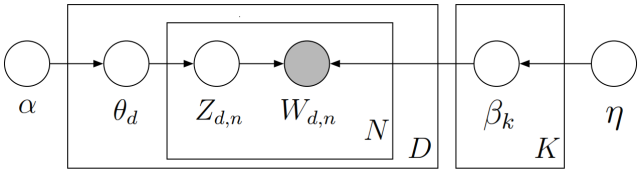
B. Hipoteze LDA modela

Sada možemo definirati LDA model. Osnovna pretpostavka modela je da se cijeli korpus dokumenata može opisati sa K tema. Pri tome *tema* se definira kao distribucija po svim riječima fiksnog rječnika, a svaka od njih se može javiti u svakom dokumentu sa nekim udjelom. Na temelju toga, dalje se pretpostavlja da je svaki dokument nastao kao generativni proces kojeg možemo opisati u koracima:

1) za svaki dokument nasumično izabire koje teme će se u njemu pojavljivati i s kolikom udjelom, drugim riječima odredi distribuciju po temama,
2) za svaku riječ dokumenta:
   a) nasumično izabire temu kojoj ta riječ pripada (na osnovi distribucije po temama),
   b) nasumično izabire riječ iz teme (na osnovi odgovarajuće distribucije po svim riječima rječnika).

Odmah možemo primijetiti kako model zapravo ne uzima u obzir poredak riječi u dokumentu (tzv. *bag of words* pretpostavka). Iako je dosta nerealistična, za ciljeve LDA modela ona je dovoljna (zamislimo članak u kojem su riječi permutirane; čak i tada bi mogli dobiti dojam o čemu on govori).

Kako su u stvarnosti jedini opaženi podaci riječi dokumenata, cilj LDA je na temelju njih izračunati sakrivenu strukturu dokumenata (teme, distribuciju tema po dokumentima, pripadnost riječi dokumenta pojedinim temama). Zapravo želimo preokrenuti imaginarni proces koji generira dokumente, te izračunati parametre procesa za koje vrijedi da su upravo oni najvjerojatniji kandidat za generator opaženih dokumenata.



Slika 3. Grafički model LDA metode.

Uvedimo sada neke oznake kako bi mogli matematičkim jezikom precizno opisati model. Neka su teme $\beta_{1:K}$ gdje je β_i distribucija na cijelom rječniku. Neka je θ_d distribucija tema za d-ti dokument. Nadalje, sa z_{dn} ćemo označiti pripadnost n-te riječi nekoj temi u d-tom dokumentu. Konačno, w_{dn} će biti opažena n-ta riječ u d-tom dokumentu. Pri tome slučajne varijable $Z_{d,n}$ i $W_{d,n}$ potječu iz multinomijalne razdiobe, dok slučajne varijable β_k i θ_d potječu iz Dirichletove razdiobe. Tada opisani imaginarni generativni proces odgovara sljedećoj združenoj distribuciji sakrivenih i opaženih slučajnih varijabli

$$P(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} P(\beta_i) \prod_{d=1}^{D} P(\theta_d) \left(\prod_{n=1}^{N} P(z_{d,n}|\theta_D) P(w_{d,n}|\beta_{1:K}, z_{d,n}) \right).$$

Sada je posteriori distribucija definirana s

$$P(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D}) = \frac{P(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{P(w_{1:D})}. \tag{3}$$

Iako je odabir Dirichletove distribucije i njeno svojstvo konjugatnosti pogodno za izvode, egzaktno računanje nazivnika u izrazu (**??**) iznimno je težak zadatak, te su stoga razvijene razne aproksimativne metode za računanje posteriori

distribucije, a kao primjer navodimo algoritme zasnovane na uzorkovanju, te determinističke.

Jedan od primjera za prvu grupu je Gibbsovo uzorkovanje (engl. *Gibbs sampling*) koji sakuplja uzorke iz aposteriori distribucije kako bi ju aproksimirao sa empiričkom distribucijom. To postiže definirajući Markovljev lanac čija je stacionarna distribucija upravo posteriori distribucija.

Varijacijske metode su deterministički algoritmi koji definiraju parametriziranu familiju distribucija na sakrivenim slučajnim varijablama i zatim pokušavaju pronaći člana te familije koji je "najbliži" posteriori distribuciju. Time zapravo problem pretvaraju u optimizacijski.

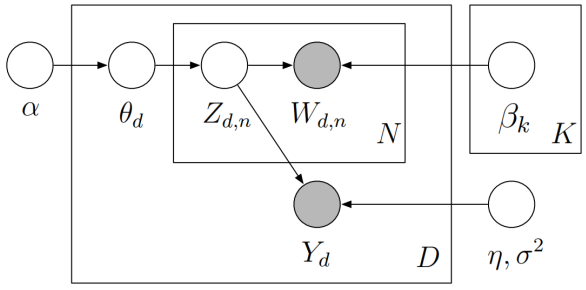
Postavalja se još i pitanje odabira parametara LDA modela. Prva mogućnost je korištenje neke od metoda pretraživanja prostora parametara (npr. traženje po rešetci). Tada pomoću kros-validacije i neke pogodne mjere kvalitete modela (npr. mjere *iznenađenosti*) možemo odabrati najbolje parametre. Ipak, pokazalo se da postoje i neki heuristički pristupi koji daju vrlo dobre rezultate u većini slučajeva. To su vrijednosti $\alpha = K/50$ te $\beta = 0.01$ koje smo i mi koristili u našim eksperimentima.

Prije samog učenja modela, potrebno je napraviti pretprocesiranje u smislu uklanjanja učestalih riječi bez posebnog značenja (veznici, čestice, prijedlozi i sl.). Osim toga, može se i riječi svesti na kanonski oblik postupkom lematizacije.

C. Varijacije LDA modela

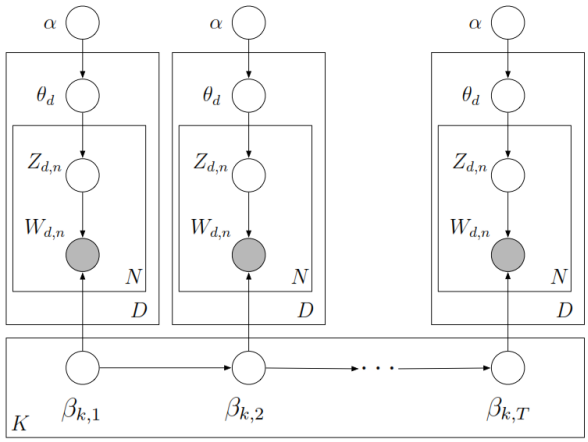
LDA model je izrazito modularan te stoga postoje njegove razne varijante. Osvrnut ćemo se samo na neke od njih.

Osim nenadziranog, postoji i nadzirana varijanta LDA modela (engl. *supervised LDA, sLDA*) u kojem svakom dokumentu dodajemo varijablu povratne informacije (engl. *response variable)*. Na primjer, povratna informacija može biti ocjena uz filmsku kritiku, ili pak broj korisnika koji su ocijenili članak korisnim. Modelom nalazimo latentne teme koje će najbolje predvidjeti povratne informacije.



Slika 4. Grafički model sLDA metode. Varijabla Y_d predstavlja varijablu povratne informacije.

U osnovnom LDA modelu pretpostavka je da poredak dokumenata unutar kolekcije nije bitan. Tu pretpostavku ćemo odbaciti kako bi mogli modelirati vremensku zavisnost tematske strukture dokumenata. Tako dolazimo do dinamičkog modeliranja tema (engl. *dynamic topic modeling, DTM*) kojim možemo proučavati vremenski razvoj tema u kolekciji. Pod time smatramo kako se u nekoj temi frekvencija pojedinih riječi mijenja (zbog novih otkrića, razvoja jezika i sl.).



Slika 5. Grafički model DTM metode. Vertikalni odsječci zapravo predstavljaju osnovnu LDA metodu.

IV. Rezultati

A. Osnovni LDA

Za osnovni model LDA preuzeli smo skup podataka agencije Associated Press koji sadrži 2246 dokumenata sa web adrese[3]. Koristili smo LDA implementaciju D. Bleia napisanu u programskom jeziku C dostupnu na istoj adresi.

Modelu smo postavili sljedeće parametre:

$$\alpha = 0.5,$$
$$K = 100.$$

Pri tome α nije fiksiran već ga uči sam model. Teme smo inicijalizirali nasumično, a model je konvergirao nakon 48 iteracija.

Prikaz deset riječi s najvećom vjerojatnosti u nekoliko tema nakon 5 iteracija algoritma:

police	south	government	stock
drug	africa	rebels	market
government	apartheid	contra	index
states	african	united	exchange
united	president	military	stocks
troops	year	contras	trading
city	white	opposition	million
estate	business	states	points
honduran	years	soviet	shares
people	film	sandinista	board

Prikaz tih istih tema nakon konvergencije algoritma(48 iteracija):

[3]http://www.cs.princeton.edu/~blei/lda-c/index.html

drug	south	rebels	stock
police	africa	government	market
states	african	contra	index
united	apartheid	contras	stocks
government	president	military	million
city	government	sandinistas	points
arrested	black	sandinista	trading
cocaine	national	aid	exchange
honduran	years	ortega	shares
medellin	white	talks	rose

Vidimo da su teme već nakon nekoliko iteracija postale prepoznatljive i zadržale većinu najčešćih riječi do konvergencije.

B. Nadzirani LDA

Za nadzirani LDA model preuzeli smo skup označenih slika sa stranice [4], a isprobat ćemo ga kroz sLDA implementaciju u C++ autora C. Wanga sa stranice [5]. Skup podataka koji smo koristili je unaprijed pripremljen i sadrži 1600 anotiranih slika podijeljenih u 8 klasa. Od toga smo polovicu koristili za treniranje i drugu polovicu na testiranje.

Istrenirali smo tri različita modela.

- 1.model s 10 tema s fiksnim parametrom $\alpha = 0.5$
  Dobivena konfuzijska matrica je sljedećeg oblika:

klasa	1	2	3	4	5	6	7	8
1	69	1	2	4	0	8	7	9
2	5	69	13	10	0	2	0	1
3	1	8	74	13	1	0	2	1
4	10	27	9	50	0	1	2	1
5	0	0	0	0	79	0	11	10
6	15	0	0	0	0	73	2	10
7	7	0	2	2	5	2	68	14
8	7	0	1	4	4	22	14	48

  Naglasimo da stupci te matrice predstavljaju stvarne klase, dok retci predstavljaju modelom predviđene klase. Prosječna točnost iznosi 0.662.

- 2. model s 10 tema koji uči parametar α.
  U ovom slučaju konfuzijska matrica je:

klasa	1	2	3	4	5	6	7	8
1	73	3	0	7	0	5	8	4
2	3	62	15	18	0	1	0	1
3	1	5	76	13	2	0	2	1
4	7	23	3	62	1	1	1	2
5	0	0	0	1	79	0	11	9
6	14	0	0	1	0	74	1	10
7	11	0	3	4	2	0	68	12
8	10	0	2	3	4	20	14	47

  Prosječna točnost iznosi 0.676.

- 3. model s 20 tema i fiksnim $\alpha = 0.4$
  Konfuzijska matrica je:

klasa	1	2	3	4	5	6	7	8
1	76	3	2	4	0	4	7	4
2	4	73	9	13	0	0	0	1
3	0	4	82	9	1	1	2	1
4	8	25	5	60	1	0	0	1
5	0	0	0	0	82	0	12	6
6	12	0	0	0	0	79	1	8
7	5	0	4	2	2	0	78	9
8	9	1	1	0	5	23	11	50

  Prosječna točnost iznosi 0.725.

Modele s više od 20 tema nismo isprobavali zbog tehničkih poteškoća i vremenskog ograničenja.

Također, dinamički LDA nažalost nismo uspjeli isprobati zbog tehničkih poteškoća.

Općenito, uspješnost nenadziranog učenja je nešto teže odrediti u usporedbi sa nadziranim iz razloga što najčešće nemamo neku prirodnu metriku. Za nadzirani LDA rezultati su bili bolji kada smo modelu dopustili da nauči i parametar α. Točnost je skočila s 62.2% na 67.6%. Još se bolje pokazalo povećanje broja tema na 20 koje je rezultiralo povećanjem točnosti na 72.5%.

Navedimo još i preciznosti (izražene u postocima) za svaku pojedinu klasu:

model	1	2	3	4	5	6	7	8
10, α fix.	60	65	73	60	88	67	64	51
10 tema	61	66	76	56	89	73	64	54
20 tema	66	68	79	68	90	73	70	62

te osjetljivosti (također u postocima):

model	1	2	3	4	5	6	7	8
10, α fix	69	69	74	50	79	73	68	48
10 tema	73	62	76	62	79	74	68	47
20 tema	76	73	82	60	82	79	78	50

V. ZAKLJUČAK

U ovom radu pokušali smo dati kratki pregled problematike modeliranja tema. Razvoj rješavanja tih problema doveo je do pojave LDA modela kojeg smo detaljnije objasnili. Ipak, nismo ulazili u preduboku teoriju nego nam je prvenstveni cilj bio razviti dobru intuiciju o modelu.

Osim osnovne inačice, model, zbog svoje velike modularnosti, dolazi i u raznim varijantama. Odlučili smo se isprobati, naravno, osnovni LDA model, ali isto tako i njegovu nadziranu varijantu. Pored njih, objasnili smo još i dinamičku varijantu koja modelira promjenu tema kroz vrijeme.

Jedan od većih problema osnovnog modela, koji je po svojoj prirodi oblik nenadziranog učenja, je evaluacija njegovih rezultata. Iako su razvijene brojne mjere kvalitete, ljudska procjena smislenosti rezultata može isto tako biti jedan način evaluacije. (Mogli bi se složiti da je bolje imati kakvu-takvu strukturu među velikim korpusom dokumenata, nego nikakvu.) Kod nadzirane varijante nemamo takvih problema, a rezultati koje smo dobili su vrlo dobri.

[4] http://labelme.csail.mit.edu/Release3.0/
[5] http://www.cs.cmu.edu/~chongw/slda/

LITERATURA

[1] Blei, Jordan, Ng, *Latent Dirichlet Allocation*[6], Journal of Machine Learning Research 3, 2003

[2] Blei, Lafferty, *Dynamic Topic Models*[7], 2008

[3] Blei, Lafferty, *Topic models*[8], 2009

[4] Blei, *Probabilistic topic models*[9], 2012

[5] Blei, McAuliffe, *Supervised topic models*[10], 2007

[6] Blei, video lectures, Machine Learning Summer School (MLSS)[11], Cambridge, 2009

[7] Wikipedia: Machine learning portal (28.4.2014). FL: Wikimedia Foundation[12]

[8] Blei, Wang, Gerrish, Chang, Boyd-Grabber, *Reading Tea Leaves: How Humans Interpret Topic Models*[13], 2009

[9] Korenčić, *Latentna Dirichletova alokacija*[14], seminarski rad, Zagreb, 2014

[6] http://machinelearning.wustl.edu/mlpapers/paper_files/BleiNJ03.pdf
[7] http://www.cs.cmu.edu/~lafferty/pub/dtm.pdf
[8] http://www.cs.princeton.edu/~blei/papers/BleiLafferty2009.pdf
[9] http://www.cs.princeton.edu/~blei/papers/Blei2012.pdf
[10] https://www.cs.princeton.edu/~blei/papers/BleiMcAuliffe2007.pdf
[11] http://videolectures.net/mlss09uk_blei_tm/
[12] http://en.wikipedia.org/wiki/Machine_learning
[13] http://www.umiacs.umd.edu/~jbg/docs/nips2009-rtl.pdf
[14] http://www.umiacs.umd.edu/~jbg/docs/nips2009-rtl.pdf