SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 289

Primjena tematskih modela na analizu dokumenata na hrvatskom jeziku

Ivan Kusalić

Zagreb, lipanj 2011.

SADRŽAJ

1. Uvod

Kao posljedica popularizacije interneta i napretka tehnologije općenito, u današnje vrijeme javno su dostupne enormne količine podataka. Te je podatke potrebno obraditi i pretvoriti u upotrebljive informacije. Najveći dio dostupnih podataka tekstovnog su tipa. Da bi se uopće moglo početi obrađivati, obilje tekstovnih podataka većinom treba filtrirati ili razvrstati u nekakve kategorije dokumenata. Također, često je iz dokumenata potrebno izvući nekakve semantičke podatke.

Mogući odgovor na gore iznesene probleme nudi primjena tematskih modela. Ovim se statističkim modelima pokušavaju otkriti semantičke teme koje se isprepliću u samoj srži dokumenta i koje su motivirale nastanak samog dokumenta. Prilikom primjene tematskih modela, dokument se promatra kao mješavina nekoliko tema, a upotrebom modela otkriva se koje su konkretne teme prisutne u promatranom dokumentu. Pri tome su teme u suštini vjerojatnosne distribucije nad riječima iz odabranog vokabulara.

U ovom su radu proučavani mehanizmi rada dva popularna tematska modela: probabilističke semantičke analize (pLSA) i latentne Dirichletove alokacije (LDA). Model latentne Dirichletove alokacije primijenjen je na dokumente hrvatskog jezika. Demonstrirana je i perspektivnost primjene tematskih modela kao sredstva za redukciju dimenzionalnosti reprezentacije dokumenta.

U nastavku rada prvo se daje uvod u statističke modele te se opisuju generativni i diskriminativni modeli. Potom se detaljnije izučavaju grafički modeli s posebnim naglaskom na usmjerene grafičke modele. Nakon toga se daje uvod u modele s latentnim varijablama te se konačno izučavaju tematski model na primjeru model probabilističke latentne semantičke analize. Sljedeće je poglavlje posvećeno latentnoj Dirichletovoj alokaciji i ono predstavlja okosnicu teoretskog dijela rada. Potom se iznosi primjena Latentne Dirichletove alokacije na probleme obrade hrvatskog jezika. Prvo se eksperimentalno utvrde generativna svojstva LDA modela, a potom se pomoću istog modeliraju dokumenti na hrvatskom jeziku. Konačno se pokazuju mogućnosti primjene LDA modela za potrebe klasifikacije dokumenata.

2. Statistički modeli

Ponekad je potrebno simulirati ili reproducirati ciljani proces. Ukoliko unutrašnji mehanizmi procesa nisu u potpunosti poznati, pribjegava se pojednostavljenom modeliranju procesa. Statistički se model procesa izrađuje na temelju prikupljenih podataka i znanju o procesu, kao i na temelju pretpostavki o procesu koje ne moraju nužno biti istinite. Statistički model procesa obično se temelji na postojanju i vrijednostima određenih slučajnih varijabli, kao i na njihovim međusobnim vezama. Da bi se mogla napraviti zadovoljavajuća procjena vrijednosti koje varijable procesa poprimaju, potrebno je prikupiti dovoljno relevantnih podataka o samom procesu, obično na temelju opetovanog promatranja procesa.

Statistički se modeli dijele u dvije velike grupe: na diskriminativne modele i na generativne modele. Iznimno, konkretni statistički model može imati osobine svojstvene za obje grupe.

Tematski modeli (engl. *topic models*) su vrsta generativnih modela s latentnim varijablama i spadaju u skupinu usmjerenih grafičkih modela, koji su podvrsta statističkih modela. U nastavku poglavlja, prije razmatranja samih tematskih modela, razmatrat će se osnovna svojstva većih skupina statističkih modela kojima tematskih modeli pripadaju.

2.1. Generativni i diskriminativni modeli

Ukoliko se ne zna dovoljno o unutrašnjim mehanizmima ciljanog procesa, isti se može zadati na temelju skupa ulaza u proces i pripadnih izlaza/odziva. Stoga, neka je x ulaz u promatrani proces, a y pripadni izlaz.

Generativnim se modelima pokušavaju modelirati i unutarnji mehanizmi procesa, to jest modelira se zajednička distribucija $p(x, y)$. Generativni model je potpuni vjerojatnosni model svih varijabli procesa. Vrijednost izlaza y procjenjuje se pomoću

uvjetne vjerojatnosti $p(y|x)$ koja se dobije upotrebom Bayesovog pravila:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \tag{2.1}$$

Za razliku od generativnog modela, diskriminativnim modelom promatranog procesa direktno se modelira preslikavanje iz x u y, odnosno u slučaju vjerojatnosnih diskriminativnih modela, uvjetna razdioba $p(y|x)$. Ovim se modelima stoga ne modeliraju unutarnji mehanizmi procesa, samo se pokušava dobiti veza između ulaza x i izlaza y. Diskriminativni pogled na proces je shematski prikazan na slici 2.1. Kod diskriminativnih modela ne postoji vjerojatnosna distribucija pridružena ulazu x u proces.

Slika 2.1: Diskriminativan pogled na proces. O unutarnjim mehanizmima procesa se ne mora puno znati. Dovoljno je promatrati ulaze i izlaze iz procesa.

Izbor generativnog ili diskriminativnog procesa ima dalekosežne posljedice. Upotrebom generativnog modela, moguće je simulirati promatrani proces i dobiti vrijednosti za proizvoljne varijable modela. To znači da je generativnim modelom moguće generirati novi, smisleni skup ulaza i pripadnih izlaza, što nije moguće napraviti upotrebom diskriminativnog modela. S druge strane, ukoliko je zadatak samo odrediti izlaz y za dani ulaz x, primjena diskriminativnih modela generalno daje bolje rezultate, pod uvjetom da je skup uzoraka na temelju kojeg je model naučen dovoljno velik (Vapnik, 1998). Ovaj je rezultat intuitivno jasan, pošto se učenjem generativnog modela prvo rješava kompleksniji problem procjene načina rada cijelog procesa, te se na temelju tog rješenja konačno rješava i jednostavniji problem procjene odziva procesa y na dani ulaz x, dok diskriminativni model rješava samo potonji, jednostavniji problem. Donedavno je među znanstvenicima vladalo nepodijeljeno mišljenje da su diskriminativni modeli uvijek superiorniji od generativnih modela. Nedavna istraživanja (Andrew Y. Ng, 2001) pokazuju da, iako generativni modeli uistinu imaju veću asimptotsku pogrešku od diskriminativnih, za približavanje stvarne pogreška modela asimptotskoj često je potreban manji skup za učenje. Točnije, diskriminativni modeli uobičajeno trebaju broj uzoraka za učenje koji je linearan u odnosu na broj parametara modela, dok generativni modeli često trebaju broj uzoraka koji je samo logaritamski u odnosu na broj parametara modela. Posljedično, za generativne i diskriminativne modele slične složenosti

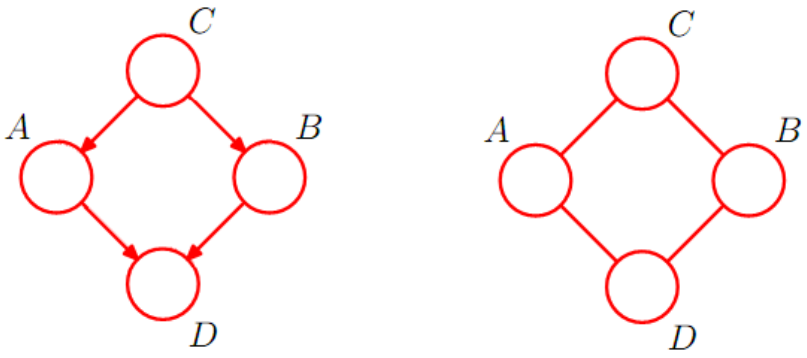
(broja parametara), na manjim skupovima za učenje bolje rezultate daju generativni modeli, a kako broj uzoraka u skupu za učenje raste, na koncu diskriminativni modeli postižu bolje rezultate.

2.2. Grafički modeli

Grafički modeli (engl. *Graphical models*) su vjerojatnosni modeli kojima se odnosi između internih varijabli modela predstavljaju pomoću grafa. Upotrebom grafa omogućen jednostavan uvid u internu strukturu modela i olakšava se razumjevanje principa na kojima se funkcionalnost model temelji. Dio slika korištenih u nastavku preuzete su iz (Bishop, 2006).

Graf je apstraktni prikaz skupa objekata objekata pri čemu su neki parovi objekata spojeni vezama. Objekti se nazivaju čvorovima ili vrhovima, a veze između njih bridovima ili lukovima. Grafovi se dijele u dvije velike skupine: na usmjerene grafove i na neusmjerene grafove. Ako su dva čvora a i b povezani bridom koji ide od a do b, tada se čvor a naziva još i roditeljem čvora b, a za čvor b se kaže da je dijete čvora a.

Na slici 2.2 prikazani su jednostavan usmjereni i neusmjereni graf. Grafovi se sastoje od četiri čvora, označeni s A, B, C i D. Primjer usmjerenog brida je brid od A do D na lijevo grafu, pri tome je čvor A roditelj čvoru D i D je dijete čvora A. Brid koji povezuje čvorove A i D na desnom grafu je neusmjeren.



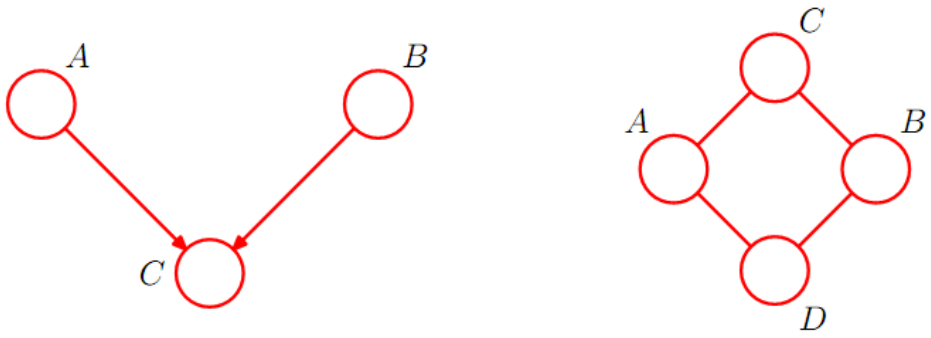
Slika 2.2: Primjeri usmjerenog grafa (lijevo) i neusmjerenog grafa (desno).

U vjerojatnosnim grafičkim modelima svaki čvor predstavlja slučajnu varijablu ili više njih, dok bridovi izražavaju vjerojatnosne veze među varijablama. Grafički model obuhvaća način na koji se zajednička distribucija nad svim varijablama može rastaviti na umnožak uvjetnih vjerojatnosti od kojih svaka ovisi o podskupu varijabli.

Po uzoru na grafove, grafički modeli dijele se na dvije velike skupine:

1. usmjereni grafički modeli odnosno Bayesove mreže (engl. *Bayesian network, belief network ili directed acyclic graphical model*),
2. neusmjerene grafičke modele odnosno Markovljeva slučajna polja (engl. *Markov random field, Markov network ili undirected graphical model*).

Usmjerenim grafovima jednostavnije se izražavaju kauzalne veze između varijabli, dok su neusmjereni grafovi bolje prilagođeni za izražavanje mekih ograničenja. Ove dvije skupine grafičkih modela različite su ekspresivnost, te se neki modeli mogu izraziti samo pomoću jednog ili samo pomoću drugog tipa grafičkog modela. Jednostavni primjeri takvih grafova prikazani su na slici 2.3.



Slika 2.3: Primjeri grafičkih modela koji se ne mogu izraziti u oba tipa grafičkih modela. Lijevi model nije moguće izraziti kao neusmjereni grafički model, dok se desni model ne može izraziti kao usmjereni grafički model.

Obje skupine grafičkih modela mogu se prevesti u poseban oblik poznat kao faktorizacijski graf. U nastavku će se razmatrati samo Bayesove mreže, odnosno neusmjereni grafički modeli.

2.2.1. Bayesove mreže

Neka je dana zajednička distribucija tri varijable a, b i c. Uzastopnim korištenjem svojstva uvjetne vjerojatnosti moguće je dobiti:

$$p(a, b, c) = p(c|a, b)p(b|a)p(a) \tag{2.2}$$

ili općenitije:

$$p(x_1, \ldots, x_n) = p(x_n|x_1, \ldots, x_{n-1})p(x_{n-1}|x_1, \ldots, x_{n-2}) \cdots p(x_2|x_1)p(x1) \tag{2.3}$$

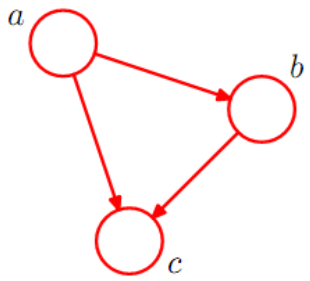
Prethodna jednadžba poznata je kao pravilo umnoška za vjerojatnosti i vrijedi za bilo koju zajedničku distribuciju n varijabli, pri čemu nije bitno da li su varijable x_i diskretne ili kontinuirane.

Da bi se na temelju jednadžbe (2.3) dobio model Bayesove mreže, potrebno je provesti sljedeće korake:

1. za svaku varijablu x_i uvede se po jedan čvor i tom se čvoru pridruži uvjetna vjerojatnost za varijablu x_i;

2. za svaku uvjetnu vjerojatnost u jednadžbi (2.3) u graf se doda usmjereni brid od varijable po kojoj je promatrana uvjetna vjerojatnost uvjetovana do čvora uvjetovane varijable.

U slučaju tri varijable i jednadžbe (2.2) dobiva se graf prikazan na slici 2.4.



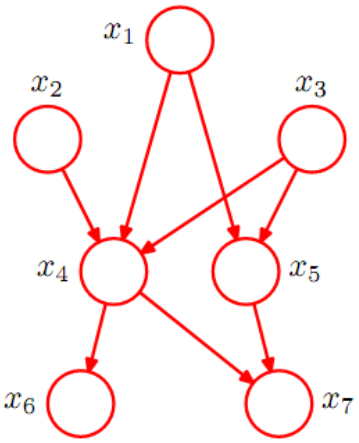
Slika 2.4: Jednostavni grafički model 3 varijable koji odgovara jednadžbi (2.2).

Tako su primjerice za uvjetnu vjerojatnost $p(c|a,b)$ dodani bridovi od čvorova a i b k čvoru c.

Treba primijetiti da je lijeva strana jednadžbe (2.3) simetrična u varijablama x_i dok desna strana nije. Tako je dekompozicijom desne strane jednadžbe implicitno odabran konkretan poredak varijabli x_i, a time i konkretan oblik grafa prikazan na slici 2.4. Odabirom drugačije dekompozicije zajedničke distribucije nad varijablama x_i dobio bi se i drugačiji graf.

Upotrebom jednadžbe (2.3) dobiveni grafički model je potpuno povezan i za svaki par čvorova postoji brid između njih, jer u svaki konkretni čvor x_i ulaze bridovi iz svih čvorova $x_j, j < i$, to jest čvorova koji mu prethode u odabranom poretku varijabli. Ovakva je dekompozicija moguća za svaku zajedničku distribuciju n varijabli.

Zanimljiviji su grafovi koji nisu potpuno povezani jer takvi grafovi otkrivaju zanimljive informacije i svojstva pripadnih distribucija. Graf prikazan na slici 2.5 nije potpuno povezan jer primjerice ne postoji brid između čvorova x_1 i x_2 ili primjerice x_4 i x_5.



Slika 2.5: Primjer grafa koji nije potpuno povezan.

Jednom kada je dostupan graf koji odgovara usmjerenom grafičkom modelu, veoma je jednostavno iz grafa očitati dekompoziciju zajedničke distribucije svih varijabli na umnoške uvjetnih vjerojatnosti. Svaka konkretna uvjetna vjerojatnost uvjetovana je samo s roditeljima pripadnog čvora u grafu. Za graf dan na slici 2.5, čvor x_5 ima roditelje x_1 i x_3, te pripadna uvjetna vjerojatnost glasi $p(x_5|x_1, x_3)$. Iz cijelog grafa za zajedničku distribuciju svih varijabli x_i vrijedi:

$$p(x_1, \ldots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5) \tag{2.4}$$

Općenita veza između danog usmjerenog grafa i pripadnog grafičkog modela može se izraziti sljedećom jednadžbom:

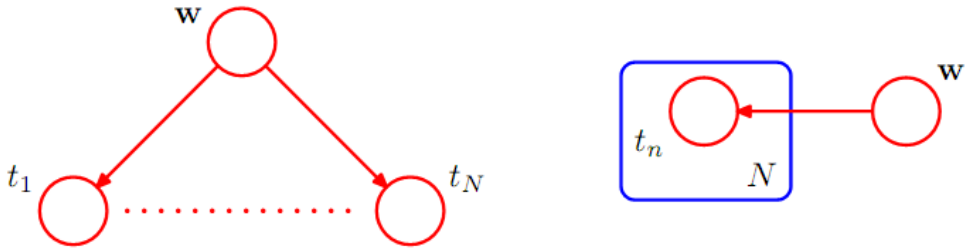
$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|rod_i) \tag{2.5}$$

gdje je rod_i skup svih roditelja od x_i. Dakle, zajednička distribucija svih varijabli usmjerenog grafičkog modela zadana je umnoškom uvjetnih vjerojatnosti za svaki čvor, uvjetovanih po svim varijablama koje odgovaraju čvorovima roditeljima u danom grafu.

Do sada je svakom čvoru bila pridružena po jedna varijabla, no moguće je jednom čvoru pridružiti i skup varijabli ili vektorsku varijablu.

Grafički modeli koji su razmatrani nemaju usmjerenih ciklusa, te se nazivaju usmjerenim acikličkim grafovima. Zahtjev da je usmjeren graf acikličan odgovara tvrdnji da postoji poredak čvorova pri kojemu svaki čvor ima samo roditelje koji su u odabranom poretku ispred promatranog čvora. Odavde sljedi korisnost usmjerenog acikličkog grafa za usmjerene grafičke modele: dvije varijable u dekompoziciji zajedničke distribucije ne mogu biti u isto vrijeme uvjetovana jedna o drugoj.

Za opis složenijih modela postaje nepraktično u graf upisivati eksplicitno sve varijable iz nekog skupa. Umjesto velikog broja srodnih varijabli može se prikazati samo predstavnik skupine uokviren u pravokutnik koji se naziva pladnjem. Uz predstavnik skupine, pladanj se označava i s brojem varijabli u skupini. Na slici 2.6 je prikazan grafički model koji sadrži skupinu varijabli $t_1, \ldots, t_N$ uvjetovanih po varijabli w. Na lijevom grafu je grafički model prikazan sa svim varijablama prisutnim u modelu, a na desnom je grafu isti taj model prikazan u sažetom obliku, pomoću pladnja umjesto skupine varijabli t_i.



Slika 2.6: Usmjereni grafički model prikazan na dva različita načina. Na lijevom je grafu model prikazan sa svim varijablama, a na desnom je grafu prikazan samo predstavnik skupine varijabli pomoću pladnja.

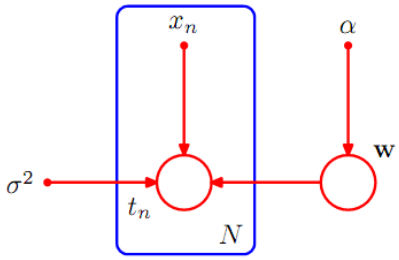
Kao primjer realnijeg grafičkog modela, neka je zadan model polinomijalne regresije, sa sljedećom razdiobom:

$$p(\mathbf{t}, \mathbf{w}|\mathbf{x}, \alpha, \sigma^2) = p(\mathbf{w}, \alpha) \prod_{n=1}^{N} p(t_n|\mathbf{w}, x_n, \sigma^2) \tag{2.6}$$

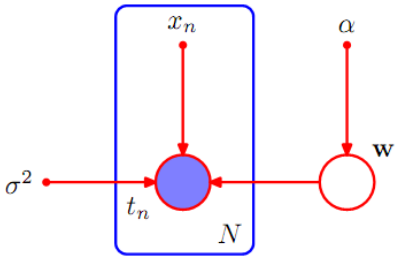
gdje je $\mathbf{w}$ vektor polinomijalnih koeficjenata, a $\mathbf{t} = (t_1, \ldots, t_N)^T$ ciljani izlazni vektor za ulazni vektor $\mathbf{x} = (x_1, \ldots, n_N)^T$.

U ovom su modelu parametri $\mathbf{x}$ i α deterministički, a neka je i varijanca σ^2 poznata i fiksirana, pa ujedno i deterministička. U prikazu grafičkih modela mogu se eksplicitno prikazati i deterministički parametri. Oni se prikazuju točkama, za razliku od slučajnih varijabli koje se prikazuju praznim kružnicama. Tako za polinomijalnu regresiju dobivamo grafički model prikazan na slici 2.7.

Ponekad je zgodno zbog jasnoće modela naglasiti koje su varijable vidljive a koje su latentne. Tada se vidljive varijable prikazuju zasjenčanim čvorovima, za razliku od latentnih varijabli kod kojih su čvorovi prazni. Na slici 2.8 je prikazan model polinomijalne regresije s naglašenim vidljivim varijablama, u ovom slučaju vidljive su samo varijable t_i.



Slika 2.7: Grafički model polinomijalne regresije s prikazanim deterministčkim parametrima



Slika 2.8: Grafički model polinomijalne regresije s prikazanim deterministčkim parametrima i označenim varijablama koje su vidljive.

Za grafičke modele veoma je bitan pojam uvjetne nezavisnosti. Neka za tri slučajne varijable a, b i c vrijedi:

$$p(c|a,b) = p(c|a) \tag{2.7}$$

U tom se slučaju kaže da je c uvjetno nezavisna od b ako je dana a. Jednadžba (2.7) se može upotrijebiti na izrazu za $p(b,c|a)$ i tada se dobije:

$$p(b,c|a) = p(c|a,b)p(b|a) = p(c|a)p(b|a) \tag{2.8}$$

Iz čega je vidljivo da se zajednička distribucija varijabli b i c, uvjetovana na varijabli a, faktorizira na umnožak nezavisnih faktora. Ovo pak znači da su varijable b i c statistički nezavisne ako je dana varijabla a.

Pojam uvjetne nezavisnosti je značajan za grafičke modele jer pojednostavljuje strukturu modela te smanjuje količinu računanja potrebnog za učenje modela.

2.3. Generativni modeli s latentnim varijablama

Latentne varijable su slučajne varijable procesa čije stanje nije moguće direktno očitati. Stanja ovih varijabli su skrivena, a o njihovim se vrijednostima zaključuje na temelju vrijednosti ostalih varijabli.

Modeli s latentnim varijablama (engl. *Latent variable models*) pokušavaju objasniti vrijednosti vidljivih varijabli na temelju vrijednosti latentnih varijabli. Pri tome stanja latentnih varijabli mogu odgovarati nekom svojstvu fizičke stvarnosti koje je u principu mjerljivo, no zbog praktičnih razloga se mjerenje ne provodi. Tada se latentne varijable nazivaju još i skrivenim varijablama (engl. *hidden variable*). Druga mogućnost je da se latentne varijable koriste za apstraktne koncepte poput kategorija, struktura podataka ili mentalnih stanja. Takve se varijable nazivaju još i hipotetskim varijablama (engl. *hypothetical variables ili hypothetical constructs*).

Modeli s latentnim varijablama pretpostavljaju da:

1. odziv vidljivih varijabli je uvjetovan stanjem latentnih varijabli,

2. vidljive varijable zadovoljavaju svojstvo lokalne nezavisnosti.

Pretpostavka lokalne nezavisnosti znači da su vidljive varijable međusobno nezavisne jednom kada su latentne varijable konkretizirane, to jest vidljive varijable su uvjetno nezavisne jedna o drugoj uz dane latentne varijable.

Pojam lokalne nezavisnosti uveli su 1968. godine Lazarsfeld i Henry (P.F. Lazarsfeld, 1968), te su ga demonstrirali na sljedećem primjeru. Neka je u anketi 1000 ljudi pitano čitaju li časopise A i B. Njihovi odgovori su sumirani u tablici 2.1.

Tablica 2.1: Odgovori ispitanika u anketi.

	čita A	ne čita A	ukupno
čita B	260	140	400
ne čita B	240	360	600
ukupno	500	500	1000

Iz tablice 2.1 vidi se da su dvije slučajne varijable (čitanje časopisa A i čitanje časopisa B) zavisne, $P(A,B) = \frac{260}{1000} \neq \frac{500}{1000}\frac{400}{1000} = P(A)P(B)$.

Ukoliko se u sklopu analize ankete promatra i edukacija ispitanika, dobiva se tablica 2.2.

Da bi čitanje časopisa A i čitanje časopisa B bio nezavisno za zadani nivo obrazovanosti, mora vrijediti pojedinačno za oba nivoa obrazovanja $P(A,B) = P(A)P(B)$. Uistinu: $\frac{240}{500} = \frac{300}{500}\frac{400}{500}$ i $\frac{20}{500} = \frac{100}{500}\frac{100}{500}$. Dakle, ukoliko se napravi podjela ispitanika po obrazovanju, nema ovisnosti između čitanja časopisa A i čitanja časopisa B. To jest, čitanje časopisa A i B nezavisni su ukoliko se obrazovanje uzme u obzir. Obrazovanje objašnjava razliku u čitanju A i B. Iako nivo obrazovanja ne mora biti poznat (pokriven pitanjima u anketi), i dalje se može pojaviti kao latentna varijabla u modelu.

Tablica 2.2: Odgovori ispitanika u anketi s podatcima o obrazovanju istih.

visokoobrazovan				niskoobrazovan			
	čita A	ne čita A	ukupno		čita A	ne čita A	ukupno
čita B	240	60	300	čita B	20	80	100
ne čita B	160	40	200	ne čita B	80	320	400
ukupno	400	100	500	ukupno	100	400	500

Prilikom učenja generativnih modela s latentnim varijablama, traži se skup latentnih varijabli koje najbolje objašnjavaju dostupan skup za učenje. Pri tome se implicitno pretpostavlja da je model koji se uči dovoljno ekspresivan da je u stvarnosti mogao generirati skup za učenje.

Jedna od prednosti upotrebe latentnih varijabli je i ta što latentne varijable mogu značajno smanjiti dimenzionalnost podataka – velik se broj vidljivih varijabli može grupirati pomoću temeljnog koncepta koji odgovara latentnoj varijabli.

2.4. Tematski modeli

Tematski modeli (engl. *Topic models*) zasnivaju se na ideji da su dokumenti mješavine tema, odnosno da postoji nekoliko generalnih tema koje se isprepliču u dokumentu i koje su motivirale njegov nastanak. Pri tome se tema u suštini promatra kao vjerojatnosna distribucija nad riječima. Postoje dvije ključne pretpostavke na kojima se temelje tematski modeli:

1. semantičke informacije mogu se dobiti na temelju frekvencije pojeve riječi u dokumentu,

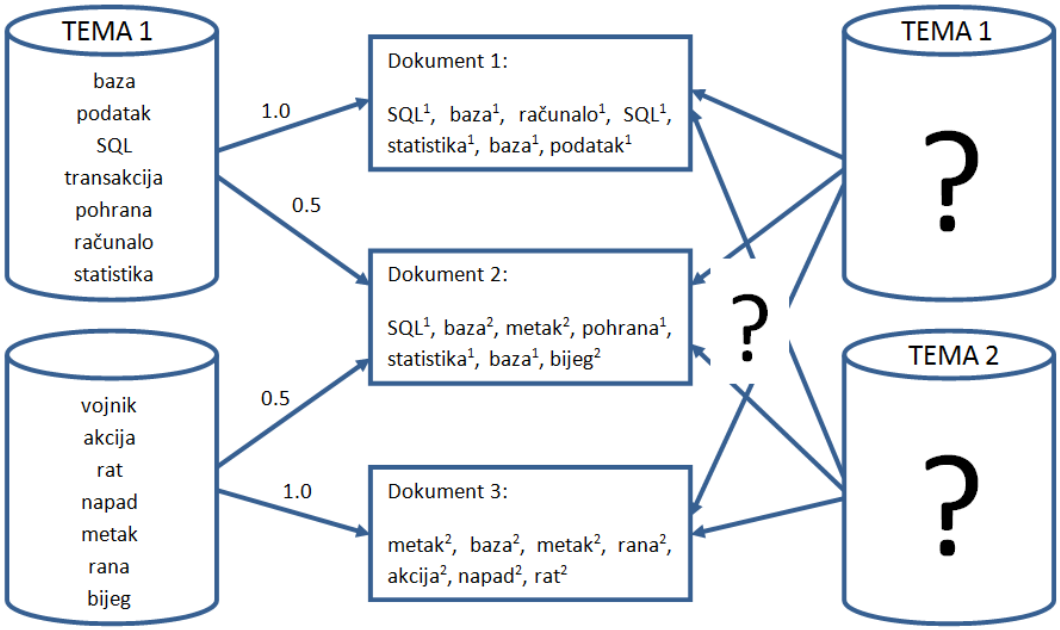
2. smanjenjem dimenzionalnosti moguće je očuvati semantičke informacije.

Tematski su modeli generativni modeli dokumenata, odnosno naučenim modelom moguće je generirati novi dokumenti. Konkretnije, da bi se generirali novi dokumenti, prvo se odabere distribucija tema za pojedini dokument. Nakon što je odabrana distribucija tema, da bi se generirala svaka pojedina riječ, slučajno se odabere tema u skladu s odabranom distribucijom, nakon čega se prema distribuciji nad riječi odabrane teme izabere nova riječ. Proces se ponavlja dokle god nije generiran dokument željene duljine.

Problem učenja tematskog modela je inverzan problemu generiranja novih dokumenata. Dan je korpus s konkretnim dokumentima, na temelju kojih se određuju teme koje su generirale dokumente u korpusu.

Upotreba vjerojatnosne distribucije nad riječima za modeliranje pojedine teme za posljedicu ima mogućnost interpretacije svake pojedine teme zasebno, što je velika prednost u odnosu na modele koji samo vrše projekcije da bi smanjili dimenzionalnost vektorskih prostora i čije koordinatne osi nemaju jednostavnu interpretaciju. Primjer potonjih je latentna semantička analiza (engl. *Latent semantic analysis*, LSA).

Bitno je naglasiti da generativni proces tematskih modela ne pravi nikakvu pretpostavku o poretku riječi u dokumentu, već se pretpostavlja da isti nije presudan i da je relevantne semantičke informacije moguće dobiti i bez znanja o poretka riječi u dokumentu. Ova pretpostavka se naziva pretpostavkom vreće riječi i često se pojavljuje u modelima prlikom pretraživanja informacija. Upotrebom tematskih modela svjesno se odbacuje velika količina zasigurno bitnih informacija da bi model bio dovoljno jednostavan, a u nadi da su zadržane informacije dovoljno bogate i kvalitetne.



Slika 2.9: Shematski prikaz generativnog procesa tematskog modela s dvije teme (lijevo) te prikaz problema statističkog zaključivanja, odnosno učenja tematskog modela (desno).

Slika 2.9 prikazuje tematski model s dvije teme u dvije različite situacije. Na lijevoj je polovici slike prikazana generativna primjena tematskog modela. Obje teme su već naučene te se sada pomoću naučenog modela generiraju tri dokumenta. Tema 1 se odnosi na proces spremanja podataka u bazu podataka, a tema 2 tiče se ratne akcije. Dokument 1 generira se samo na temelju teme 2 (označeno s težinskim faktorom 1.0),

dokument 2 generiram je primjenom obje teme (težinski faktori 0.5), a dokument 3 je generiran isključivo na temelju teme 2. Pripadnost riječi određenoj temi je označena brojem povrh iste. Riječi nisu ograničene na pripadnost samo jednoj temi. Riječ 'baza' je višeznačna te u ovom primjeru ima dvije različite primjene: baza podataka i vojna baza.

Desna strana slike 2.9 predstavlja problem učenja generativnog modela. Dane su frekvencije riječi prisutne u pojedinim dokumentima, a potrebno je odrediti temu koja je s najvećom vjerojatnošću generirala dokument, ili konkretnije – potrebno je odrediti vjerojatnosnu distribuciju nad riječima za svaku temu, distribuciju teme za svaki dokument, a ponekad i temu odgovornu za generiranje svake riječi.

Treba napomenuti da tematski modeli nisu primjenjivi samo za modeliranje tema u skupu dokumenata. Osim ove očigledne primjene, tematski se modeli mogu primijeniti na bilo kakvu zbirku skupova diskretnih podataka. Kod primjene na dokumente ta zbirka je korpus, skup diskretnih podataka je dokument, a diskretan podatak je riječ. Moguće je primijeniti tematske modele na kolekcije slika, pri čemu je slika analogna dokumentu, a pojedinačni slikovni elementi su analogni riječima. Primjer moguće primjene je i modeliranje navika korisnika, poznato još kao suradničko filtriranje (engl. *Collaborative filtering*), gdje se primjerice promatra korisnikov izbor filmova. Pri tome je dana kolekcija korisnika, gdje je korisnik analogan dokumentu, a film riječi.

2.4.1. Probabilistička latentna semantička analiza

Probabilistička latentna semantička analiza (engl. *Probabilistic latent semantic analysis*, pLSA), poznata još pod nazivima Probabilističko latentno semantičko indeksiranje (engl. *Probabilistic latent semantic indexing*, pLSI) i aspektni model (engl. *Aspect model*), je tematski model inspiriran latentnom semantičkom analizom. U osnovi, Latentna semantička analiza preslikava visokodimenzionalan vektor frekvencija riječi iz dokumenta u prostor manje dimenzionalnosti poznat kao latentni semantički prostor, gdje ostaju djelomično očuvane semantičke informacije o dokumentu. Iako je LSA u širokoj upotrebi, mnogi smatraju da je teoretska osnova manjkava, a koordinatne osi dobivenog latentnog semantičkog prostora nije jednostavno interpretirati. Za razliku od latentne semantičke analize koja se temelji na linearnoj algebri i vektorskim prostorima, probabilistička latentna semantička analiza temelji se na latentnim modelima.

Model probabilističke latentne semantičke analize je model latentne varijable koji pridružuje latentnu varijablu teme $z \in \mathcal{Z} = \{z_i, \ldots, z_K\}$ svakoj riječi u dokumentu. Neka je $\mathcal{D} = \{d_1, \ldots, d_N\}$ zbirka dokumenata, a $\mathcal{W} = \{w_1, \ldots, w_M\}$ skup riječi koje

čine vokabular. Model zajedničke vjerojatnosne distribucije nad $\mathcal{D} \times \mathcal{W}$, definiran je kao:

$$p(d,w) = p(w|d)p(d), \text{ gdje je } p(w|d) = \sum_{z \in \mathcal{Z}} p(w|z,d)p(z|d) = \sum_{z \in \mathcal{Z}} p(w|z)p(z|d) \tag{2.9}$$

odnosno:

$$p(d,w) = \sum_{z \in \mathcal{Z}} p(w|z)p(z|d)p(d) \tag{2.10}$$

Kao i ostali modeli latentne variable, i aspektni model pretpostavlja uvjetnu nezavisnost, za varijable d i w uvjetovane na stanju latentne varijable z, pri čemu je kardinalnost latentne varijable $|\mathcal{Z}| = K$ manja od broja riječi i dokumenata u korpusu. Aspektni se model može preoblikovati u oblik simetričan u dokumentima i riječima:

$$p(d,w) = \sum_{z \in \mathcal{Z}} p(w|z)p(d|z)p(z) \tag{2.11}$$

Obje formulacije aspektnog modela prikazane su na slici 2.10.

Slika 2.10: Na lijevom grafu je prikazan aspektni model izražen jednadžbom (2.10), a na desnom grafu je prikazan aspektni model izražen jednadžbom (2.11). Radi preglednosti napisane su i odgovarajuće uvjetne vjerojatnosti.

U modelu pLSA vjerojatnosti pojave riječi w uvjetovana je o temi z_i, pa $p(w|z_i)$ ima multinomijalnu distribuciju, pri ćemu se konkretne multinomijane distribucije $p(\cdot|z_i)$ nazivaju faktorima. Faktori se mogu predstaviti točkama na $(M-1)$-dimenzionalnom simpleksu[1], zato što se vokabular sastoji od M jedinstvenih riječi, a faktori su vjerojatnosne distribucije pa moraju zadovoljiti uvjet $\sum_{j=1}^{M} p(w_j|z_i) = 1$. Ukoliko se promatra konveksna ljuska svih K točaka (koje odgovaraju svakoj od K tema), ona čini L-dimenzionalni simpleks koji se nalazi unutar $(M-1)$-dimenzionalnog simpleksa, za čiju dimenzionalnost vrijedi $L \leq K - 1$, jer konveksna ljuska K točaka razapinje simpleks maksnimalne dimenzionalnosti $K - 1$. Točnije, iz temeljne jednadžbe pLSA modela (2.9) sljedi da model pretpostavlja da je $p(w|d)$ moguće aproksimirati s $\sum_{z \in \mathcal{Z}} p(w|z)p(z|d)$, to jest konveksnom kombinacijom faktora $p(w|z)$, gdje težinski faktori $0 \leq p(z|d) \leq 1$ jedinstveno definiraju točku na L-dimenzionalnom simpleksu.

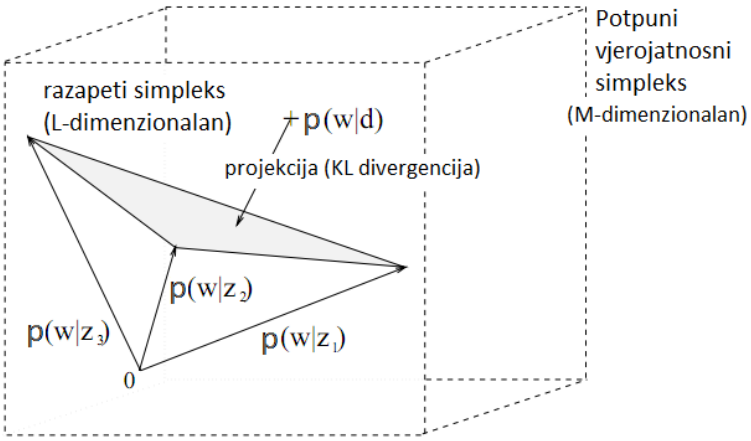
[1] $(n-1)$-dimenzionalni simpleks je ploha proizvoljnog oblika u n-dimenzionalnom prostoru

Kako je dimenzionalnost ovog simpleksa L manja od dimenzionalnosti $(M-1)$ potpunog vjerojatnosnog simpleksa, preslikavanjem u L-dimenzionalni simpleks vrši se redukcija dimenzionalnosti prostora multinomijalnih distribucija, te se rezultantni simpleks može poistovjetiti s probabilističkim latentnim semantičkim prostorom. Ukoliko temeljna jednadžba (2.9) ne vrijedi u potpunost, dolazi do pogreške jednakoj Kullback–Leiblerovoj divergenciji između stvarne i aproksimativne distribucije. Kullback-Leiblerova divergencija definirana je s:

$$D_{KL}(p,q) = \sum_i p_i \, log \frac{p_i}{q_i} \tag{2.12}$$

Slika 2.11 shematski prikazuje pLSA prostor.

Primjenom pLSA na dokument on se predstavlja kao mješavina tema, pri čemu je svakoj temi dodjeljuje ukupni udi koji joj pripada, to jest udio riječi u dokumentu koje su motivirane tom temom. Time se zapravo dobiva redukcija dimenzionalnosti dokumenta s M dimenzija na K dimenzija, jer se umjesto frekvencija svih riječi iz vokabulara pamte samo udjeli pojedinih tema. Zapravo redukcijom se efektivno dobiva dokument dimenzije $K-1$ jer je na temelju poznatih udjela $K-1$ tema moguće izračunati udio K-te teme. Ipak, radi jednostavnosti i preglednosti uobičajeno je promatrati dokument kao K-dimenzionalni vektor udjela svih K tema.



Slika 2.11: Prikaz probabilističkog latentnog semantičkog prostora koji je L-dimenzionalni simpleks ugrađen unutar potpunog $(M-1)$-dimenzionalnog vjerojatnosnog simpleksa. Projekcijom u prostor manje dimenzionalnosti napravljena pogreška jednaka Kullback–Leiblerovoj divergenciji koja je rezultat korištenja aproksimacije $p(w|d)$ prema temeljnoj jednadžbi pLSA modela (2.9). Slika je inspirirana prikazom iz originalnog članka, (Hofmann, 1999).

Pokazuje se da u praksi pLSA model generalno daje bolje rezultate od modela LSA (Hofmann, 1999). Dodatna prednost modela pLSA jest mogućnost modeliranja

višeznačnih riječi što nije u potpunosti moguće kod modela LSA, gdje se svaka riječ uvijek poistovjećuje s točno jednom točkom u vektorskom prostoru koji može imati samo jedno značenje.

Uz sva korisna svojstva model pLSA posjeduje i neka značajnija ograničenja:

1. broj parametara koje je potrebno odrediti u modelu pLSA, zbog jednadžbe (2.9), iznosi $K(M + N) = KM + KN$ što raste linearno s brojem dokumenata u korpusu N;

2. kao posljedica velikog broja parametara, model pLSA sklon je problemu prenaučenosti, što pokazuju i eksperimentalna istraživanja (David M. Blei, 2003);

3. uvjetne vjerojatnosti koje se modeliraju u modelu pLSA uvjetovane su dokumentom d, koji je najčešće predstavljen indeksom dokumenta u korpusu.

Posebno je ozbiljno zadnje navedeno ograničenje. Kako pLSA koristi uvjetne vjerojatnosti uvjetovane o indeksu dokumenta (d) u korpusu, pLSA nije pravi generativni model. Naime ne postoji dobar odabir vrijednosti indeksa d za neviđene dokumente jer se isti ne nalaze u korpusu i stoga nemaju svoj indeks. Iako postoje metode kojima se ovo ograničenje pokušava ublažiti (Hofmann, 1999), one se ne može potpuno otkloniti. Posljedično, model pLSA moguće je potpuno smisleno koristiti samo na korpusima koji se ne mijenjaju tokom vremena, primjerice za dohvat sličnih dokumenata. Da bi se problem fiksnog korpusa ublažio, moguće je periodički ponovno učiti model pLSA, kako se ciljani korpus mijenja.

U praksi se model pLSA ne koristi pretjerano često, djelom i kao posljedica formulacije modela Latentne Dirichletove alokacije (engl. *Latent Dirichlet allocation*, LDA) samo tri godine nakon modela pLSA. Ovaj se model može promatrati kao generalizacija modela pLSA i bit će detaljnije razmatran u nastavku rada. Uz razne druge prednosti, model LDA je pravi generativni model, jer se njegovom upotrebom mogu potpuno prirodno generirati novi skupovi podataka. Za tematske modele generativni smjer nema puno smisla, jer se dobivaju dokumenti bez strukture. Ipak, mogućnost generiranja novih dokumenata potvrđuje ispravnost pretpostavki modela i omogućava bolje razumjevanje način rada modeliranog procesa.

3. Latentna Dirichletova alokacija

Latentna Dirichletova alokacija (engl. *Latent Dirichlet allocation*, LDA) popularni je tematski model koji se može promatrati kao nadogradnja ili čak generalizacija modela pLSA (Mark Girolami, 2003). U ovom se poglavlju proučava model LDA te se isti uspoređuje s modelom pLSA. Dio slika korištenih u nastavku preuzete su iz (David M. Blei, 2003).

Prethodno opisani model probabilističke latentne semantičke analize uvodi vjerojatnosni model na nivou riječi, no ne i na nivou dokumenata. U pLSA je svaki dokument predstavljen vektorom brojeva koji predstavljaju udjele pojedinih tema u dokumentu, no kako je svaka tema uvjetovana o dokumentu (odnosno indeksu dokumenta u korpus), ovi udjeli ne daju ispravan generativni model. Posljedično, broj parametara modela raste linearno s brojem dokumenata što dovodi do značajnog problema prenaučenosti. Osim toga, nije jasno kako dodijeliti vjerojatnosti dokumentu koji ne pripada korpusu na kojem je model učen.

Osnovna pretpostavka modela LSA i pLSA je pretpostavka vreće riječi (engl. *Bag of words*), što znači da se dokument može predstaviti kao popis frekvencija riječi, a da se ne pazi na sam poredak riječi u dokumentu. Da bi se dobio ispravni generativni vjerojatnosni model, logično je dalje pretpostaviti i da predak dokumenata u korpusu nije važan, kao i poredak teme unutar dokumenta.

Finettijev reprezentacijski teorem (de Finetti, 1990) tvrdi da se bilo koja zbirka izmjenjivih slučajnih varijabli može predstaviti kao (u principu beskonačna) mješavinska distribucija (engl. *Mixture distribution*). Točnije, skup slučajnih varijabli je izmjenjiv ako je zajednička distribucija invarijantna na permutacije. Neka je π permutacija prirodnih brojeva $1, \ldots, N$. Tada vrijedi:

$$p(z_1, \ldots, z_N) = p(z_{\pi(1)}, \ldots, z_{\pi(N)}) \tag{3.1}$$

Model Latentne Dirichletove alokacije (LDA, (David M. Blei, 2003)) pretpostavlja da su teme koje su inspirirale nastanak dokumenta međusobno izmjenjive unutar dokumenta. Zato je potrebno proučavati mješane modele (engl. *Mixture models*).

Važno je naglasiti da pretpostavka izmjenjivosti nije ekvivalenta pretpostavci da su slučajne varijable nezavisne i identično distribuirane. Sve što se ovom pretpostavkom tvrdi je da su dokumenti uvjetno nezavisni i identično distribuirani. Pri tome se uvjetovanje vrši po latentnim varijablama – parametrima vjerojatnosnog modela. To pak znači da se zajednička distribucija varijabli faktorizira ukoliko se marginaliziraju latentne varijable.

Pretpostavka izmjenjivosti je u prvom redu motivirana potrebom za računalno ne prezahtjevnim modelima. Iako je ova pretpostavka pojednostavljenje stvarnog procesa, te ne mora nužno biti u potpunosti istinita, modeli kojima rezultira nisu nužno jednostavni i linearni.

Iako se u LDA pretpostavlja izmjenjivost riječi (vreća riječi) i izmjenjivost dokumenata, postoji nekoliko podvrsta izmjenjivosti, pa se pristup izgradnji LDA može primijeniti i na kompleksnije modele u kojima se mogu promatrati mješavine većih struktura, poput n-grama.

Kao i u slučaju modela pLSA, i kod modela LDA se koriste termini poput 'korpus', 'dokument' i 'riječ'. Ovi se termini koriste samo da bi model bio konkretniji, a moguće ga je primijeniti na razne kolekcije skupova diskretnih podataka.

U opisu modela LDA koristit će se sljedeča terminologija:

- riječ je osnovna jedinica diskretne informacije, definirana kao element vokabulara indeksiranog s $\{1, \ldots, V\}$. Riječi se predstavljaju kao jedinični V-dimenzionalni vektori, koji imaju samo jednu komponentu jednaku jedinici, a sve ostale komponente jednake nuli. Tako se v-ta riječ u vokabularu može označiti vektorom $w : w^v = 1$ te $w^u = 0$, za $u \neq v$, pri čemu gornji indeks označava komponentu vektora.
- dokument je niz od N riječi označen s $\mathbf{w} = \{w_1, w_2, \ldots, w_N\}$, gdje je w_n n-ta riječ u sekvenci
- korpus je kolekcija M dokumenata označena s $D = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$

Kao i u slučaju pLSA, LDA predstavlja dokumente kao slučajne mješavine nad latentnim temama, gdje je svaka tema opisana vjerojatnosnom distribucijom nad riječima. LDA pretpostavlja da se dokumenti generiraju sljedećim postupkom:

1. odabere se duljina dokumenta $N \sim Poisson(\xi)$
2. odabere se $\theta \sim Dirichlet(\alpha)$
3. za svaku od N riječi w_n:

(a) odabere se tema $z_n \sim Multinomial(\theta)$

(b) odabere se riječ w_n iz $p(w_n|z_n, \beta)$, multinomijalne distribucije uvjetovane po temi z_n

U gore navedenom generativnom postupku izbor Poissonove distribucije za modeliranje duljine dokumenta N nema direktni utjecaj na ostatak modela, pošto je generiranje duljine dokumenta nezavisno od svih ostali koraka. Zato se duljina dokumenta može modelirati bilo kojom drugom prikladnijom distribucijom, a u nastavku razmatranja će se tretirati kao da je N konstanta, a ne slučajna varijabla.

Pretpostavlja se da je dimenzionalnost k Dirichletove distribucije, a time i tematske varijable z, unaprijed određena i konstantna, to jest broj tema je unaprijed zadani parametar modela.

Multinomialna distribucija iz koje se uzorkuju pojedine riječi $w_i \sim p(w_n|z_n, \beta)$, je parametrizirana s $k \times V$ matricom β, gdje je $\beta_{ij} = p(w^j = 1|z^i = 1)$, vjerojatnost pojedine riječi pod odabranom temom. U osnovnom modelu ova je nepoznata matrica konstantna i potrebno ju je odrediti.

Dirichletova k-dimenzionalna slučajna varijabla θ može poprimiti vrijednost na $(k-1)$-dimenzionalnom simpleksu prema sljedećoj distribuciji:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1} \tag{3.2}$$

gdje je funkcija $\Gamma(\cdot)$ definirana s:

$$\Gamma(x) = \int_0^{\infty} u^{x-1} e^{-u} du \tag{3.3}$$

Dirichletova distribucija je konjugirana distribucija multinomijalnoj distribuciji. Ukoliko je q konjugirana distribucija distribuciji p, primjenom Bayesova teorema s apriornom distribucijom q rezultanta aposteriorna distribucija će biti istog oblika kao i q, jer se množenjem izglednosti razdiobe p i pripadne konjugirane apriorne distribucije q opet dobiva distribucija q s novim parametrima. Parametri Dirichletove distribucije imaju jasnu interpretaciju – parametar α_i predstavlja efektivan broj observacija događaja s rednim broje i, prije nego je bilo kakav pokus izveden. Stoga je upotrebom parametra α moguće zagladiti rezultantnu aposteriori distribuciju.

Osim što je Dirichletova distribucija konjugirana multinomijnalnoj distribucije, ona pripada eksponencijalnoj familiji distribucija i ima konačno-dimenzionalnu dovoljnu statistiku, te je zato posebno pogodna za korištenje u modelu poput ovog. Kao posljedica izbora Dirichletove distribucije, moguće je ostvariti efikasne postupke određivanje parametara modela LDA.

Multinominalna distribucija je definirana s:

$$p(m_1, m_2, \ldots, m_K|\boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \cdots m_K} \prod_{k=1}^{K} \mu_k^{m_k} \tag{3.4}$$

gdje je $\boldsymbol{\mu}$ vektor vjerojatnosti svakog pojedinog mogućeg ishoda, N je ukupni broj ponavljanja pokusa, a m_k je broj realizcija k-tog ishoda, za koje vrijedi $\sum_{k=1}^{K} m_k = N$.

Multionominalna distribucija prirodan je odabir za distribuciju kojom se bira konkretna od nekoliko mogućih realizacija. U ovom slučaju se multinomijnalnom distribucijom modelira izbor konkretne od k mougćih tema, a potom i izbor konkretne od V mogućih riječi, pri čemu je taj izbor uvjetovan po prethodno odabranoj temi.

Ukoliko su dani parametri α i β, te veličina dokumenta N, zajednička distribucija mješavine tema θ, skupa od N tema **z** i skupa od N riječi **w** dana je s:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \tag{3.5}$$

gdje je $p(z_n|\theta)$ isto što i θ_i za jedinstveni i takav da je $z_n^i = 1$.

Integriranjem po θ i sumacijom po z dobije se marginalna distribucija dokumenta:

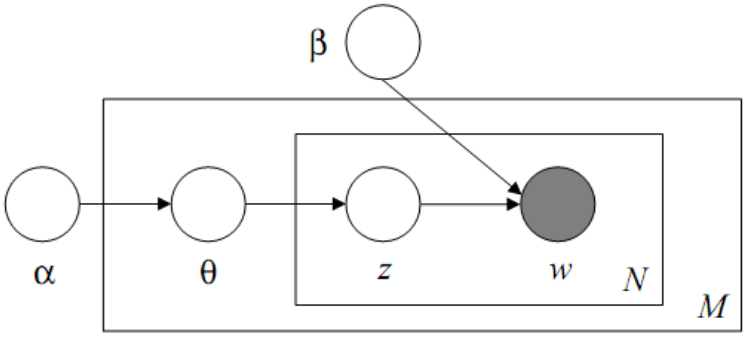
$$p(\mathbf{w}|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) p(w_n|z_n, \beta) \right) d\theta \tag{3.6}$$

Ukoliko se uzme produkt marginalnih distribucija svih dokumenata u korpusu, dobije se vjerojatnost korpusa:

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta) \right) d\theta_d \tag{3.7}$$

Na slici 3.1 je prikazan model LDA kao usmjereni grafički model. Iz grafa je vidljivo da je LDA hijerarhijski model koji se sastoji od tri nivoa. Parametri α i β su parametri na razini korpusa, što znači da su uzorkovani samo jednom i vrijede za cijeli korpus. Dirichletove slučajne varijable θ_d su varijable na razini dokumenta, uzorkovane po jednom za svaki dokument. Konačno, slučajne varijable z_{dn} i w_{dn} su varijable na razini pojedine riječi i uzorkuju se za svaku riječ posebno.

Bitno je primijetiti da se i tematske varijable z_{dn} uzorkuju prilikom izbora svake pojedine riječi. Posljedica ove činjenice je da svaki dokument može biti motiviran s nekoliko različitih tema, te je LDA uistinu ispravan tematski model.



Slika 3.1: LDA kao usmjereni grafički model. Prikaz koristi notaciju pladnjeva, pri čemu vanjski pladanj predstavlja izbor dokumenta, a unutrnji pladanj predstavlja izbor pojedine riječi u konkretnom dokumentu. Kako su samo izbori pojedinih riječi vidljivi, samo je čvor w zasjenčan.

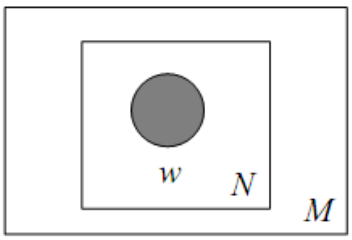
3.1. Usporedba LDA i ostalih modela s latentnim varijablama

Poučno je usporediti model LDA sa ostalim modelima s latentnim varijablama.

Upotrebom unigram modela se svaka riječ svakog dokumenta, nezavisno od ostalih, uzorkuje iz multinomijnalne distribucije, koja je distribucija zajednička za cijeli korpus:

$$p(w) = \prod_{n=1}^{N} p(w_n) \tag{3.8}$$

Slika 3.2 prikazuje unigram model kao grafički model. Treba napomenuti da unigram model nema latentnih varijabli.



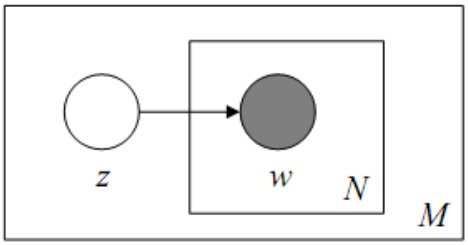
Slika 3.2: Unigram model kao grafički model.

Uvede li se latentna tematska varijabla z u unigram model, dobiva se model mješavine unigrama. Korištenjem ovog modela, svaki dokument se generira tako da se prvo uzorkuje tematska varijabla z i dobije tema dokumenta. Potom se nezavisno uzorkuje

N riječi iz uvjetne distribucije $p(w|z)$. Vjerojatnost dokumenta pod ovim modelom je:

$$p(w) = \sum_{z} p(z) \prod_{n=1}^{N} p(w_n|z) \tag{3.9}$$

Slika 3.3 prikazuje model mješavine unigrama kao grafički model. Bitno je naglasiti da je ovim modelom svaki dokument ima točno jednu temu. Model LDA može imati više tema za pojedini dokument, a zahtjeva samo jedan dodatan parametar, α.

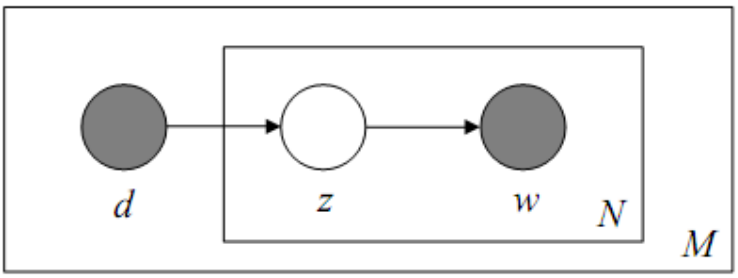


Slika 3.3: Model mješavine unigrama kao grafički model.

Konačno, slijedi prikaz prethodno razmatranog modela Probabilističke latentne semantičke analize odnosno aspektnog modela. Ovaj model je prethodno već opisan, a temelji se na sljedećoj vjerojatnosnoj distribuciji:

$$p(d, w_n) = p(d) \sum_{z} p(w_n|z)p(z|d) \tag{3.10}$$

Slika 3.4 prikazuje pLSA kao grafički model.



Slika 3.4: Aspektni model ili pLSA kao grafički model.

Aspektni model čuva velik broj parametara direktno povezanih sa svakim dokumentom (preko indeksa dokumenta d) koji mu omogućavaju pridruživanje više tema svakom pojedinom dokumentu. Zbog te povezanosti sa skupom za učenje pLSA nije pravi generativni model, a zbog velikog broja parametara ($kV + kM$) koji rastu linearno s brojem dokumenata u korpusu, model pLSA ima problema s prenaučenošću.

LDA izbjegava probleme modela pLSA tako što ne čuva velik broj parametar za svaki dokument, nego umjesto toga mješavine tema tretira kao slučajnu varijablu (s k parametara). Tako je LDA pravi generativni model, a ima samo $k + kV$ parametara čiji broj ne raste s veličinom korpusa. Za razliku od modela pLSA, model LDA ne pati od problema prenaučenosti.

Razlike između LDA i ostalih prethodno navedenih modela mogu se uočiti ukoliko se promatra geometrijska interpretacija latentnog prostora. Sva četiri prethodno opisana modela: unigram model, model mješavine unigrama, modeli pLSA i LDA koriste prostor distribucija nad riječima. Stoga se svaka distribucija može poistovjetiti s točkom na $(V - 1)$-dimenzionalnom simpleksu ili simpleksu riječi.

Unigram model odredi samo jednu točku na simpleksu riječi i dalje pretpostavlja da su sve riječi iz korpusa nastale uzorkovanjem te distribucije.

Ostali modeli imaju latentne varijable i koriste k točaka na simpleksu riječi. Ove točke tvore $(k-1)$-dimenzionalni ili tematski simpleks. Cijeli tematski simpleks nalazi se na simpleksu riječi, pa je tako i svaka točka na tematskom simpleksu ujedno i točka na simpleksu riječi.

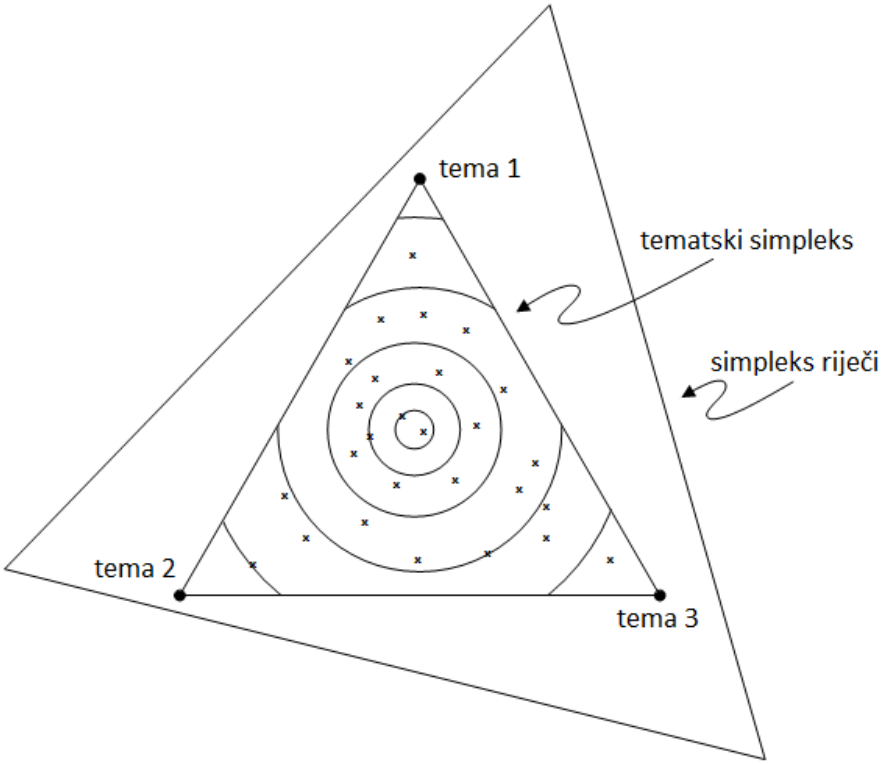
Različiti latentni model se razlikuju u načinu upotrebe tematskog simpleksa da bi generirali dokumente.

Model mješavine unigrama pretpostavlja da je svaki dokument generiran korištenjem jednog od k vrhova tematskog modela. Pojedini vrh (ujedno i točka na simpleksu riječi) je slučajno odabran i sve riječi jednog dokumenta dolaze iz tog vrha.

Aspektni model (pLSA) pretpostavlja da svaka riječ svakog dokumenta *iz skupa za učenje* dolazi iz slučajno odabrane teme. Teme se uzorkuju iz distribucije koja odgovara jednoj točki na tematskom simpleksu, a po jedna takva točka postoji za svaki dokument. Skup ovih točaka čini empirijsku distribuciju na tematskom simpleksu.

Model LDA pretpostavlja da je svaka riječ dokumenta, bilo iz skupa za učenje, bilo novog dokumenta, generirana pomoću slučajno odabrane teme uzorkovane iz distribucije parametrizirane slučajnim parametrom, koji je uzorkovan po jednom za svaki dokument iz zaglađene distribucije na tematskom simpleksu.

Slika 3.5 shematski prikazuje simpleks riječi i tematski simpleks te ilustrira razlike između prethodno navedenih metoda.



Slika 3.5: Shematski prikaz simpleksa riječi na kojem leži tematski simpleks u slučaju vokabulara od tri riječi. Vrhovi simpleksa riječi odgovaraju situaciji kada pripadna riječ ima vjerojatnost odabira jednaku 1. Tri istaknute točke na simpleksu riječi razapinju tematski simpleks i odgovaraju konkretnim distribucijama nad riječima. Model mješavine unigrama stavlja distribuciju svakog pojedinog dokumenta u jedan od ova tri vrha. Aspektni model uvodi empirijsku distribuciju nad tematskim simpleksom prikazanu točkama označenim 'x' znakom. Model LDA koristi zaglađenu distribuciju nad tematskim simpleksom čije su konture prikazane na slici. Ovaj je prikaz preuzet iz (David M. Blei, 2003).

3.2. Učenje modela LDA

Glavni problem koji je potrebno riješiti prilikom učenja modela LDA je izračun aposteriorne distribucije nad latentnim varijablama uz dani dokument za učenje:

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \tag{3.11}$$

Da bi se normalizirala ova distribucija, potrebno je marginalizirati po latentnim varijablama, i dobiva se:

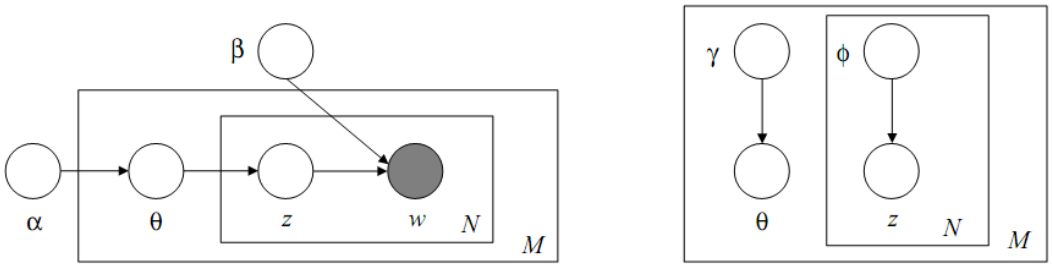
$$p(\mathbf{w}|\alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} \right) d\theta \tag{3.12}$$

što nije moguće eksplicitno riješiti zbog veze između θ i β prilikom sumiranja po latentnim temama.

Iako prethodni problem nije moguće eksplicitno riješiti, postoje razne metode kojim se ovaj problem može riješiti aproksimativno. Dva najčešća pristupa rješavanju ovog problema su primjena stohastičkih metoda poput Gibbsovog uzorkovanja (engl. *Gibbs sampling*) ili primjena determinističkog aproksimativnog algoritma varijacijskog zaključivanja (engl. *Variational inference*). Iako različiti algoritmi u principu rezultiraju različitim procjenama parametara (pogotovo u slučaju primjene nedeterminističkih algoritama), te su procjene svejedno dovoljno dobre i primjenjive u praksi.

Kao primjer algoritma za rješavanje prethodnog problema, slijedi kratak opis ideje algoritma koji predlažu autori modela LDA, algoritma varijacijskog zaključivanja. Osnovna je ideja koristiti familiju donjih granica za log-izglednost, određenih pomoću skupa varijacijskih parametara. Ovi se parametri biraju tako da se dobije što je moguće tješnja donja granica. Da bi se dobila familija prikladnih donjih granica, u grafičkom modelu se uklanjaju problematične veze između θ i β koja nastaje zbog bridova između θ, $\mathbf{z}$ i $\mathbf{w}$. Uklanjanjem ovih bridova i $\mathbf{w}$ čvorova dobiva se jednostavniji model s slobodnim varijacijskim parametrima. Slika 3.6 prikazuje stari i novi model. Dobivena familija distribucija nad latentnim varijablama je zadana sa sljedećom varijacijskom distribucijom:

$$q(\theta, \mathbf{z}|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n) \tag{3.13}$$



Slika 3.6: Lijevi graf prikazuje standardni model LDA. Desni graf prikazuje model LDA prilagođen za primjenu varijacijskog zaključivanja da bi se odredila aposteriori distribucija modela LDA.

gdje su Dirichletov parametar γ i multionmijalni parametri $(\phi_1, \ldots, \phi_N)$ slobodni varijacijski parametri koje je potrebno odrediti. Problem određivanja varijacijskih parametara dovodi do sljedećeg optimizacijskog problema:

$$(\gamma^*, \phi^*) = argmin_{(\gamma,\phi)} D_{KL}(q(\theta, \mathbf{z}|\gamma, \phi) \mid\mid p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)) \tag{3.14}$$

to jest, optimalne vrijednosti varijacijskih parametara nalaze se minimiziranjem Kullback-Leibler (KL) divergencije između varijacijske distribucije i stvarne aposteriori vjerojatnosti, iz čega slijede korigacijske jednadžbe:

$$\phi_{ni} \propto \beta_{iw_n} exp\{E_q[log(\theta_i)|\gamma]\} \tag{3.15}$$

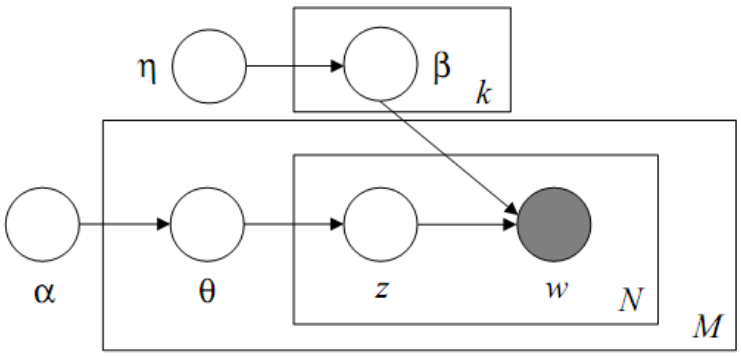
$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni} \tag{3.16}$$

Više informacija o varijacijskom pristupu učenja LDA je moguće naći u (David M. Blei, 2003).

3.3. Zaglađeni model LDA

Ukoliko se LDA primjenjuje na korpus s velikim vokabularom, tada se javlja ozbiljan problem rijetkosti podataka. Novi dokumenti koji je potrebno obraditi često sadrže riječi koje se prethodno nisu pojavile u korpusu. Procjena maksimalne izglednosti za multinomijnalne parametre dodjeljuje vjerojatnost jednaku nuli novim riječima, a posljedično i cijelom novom dokumentu. Standardan pristup ovom problemu uključuje zaglađivanje multinomijnalnih parametara, tako da se svim riječima dodjeli strogo pozitivna vjerojatnost bez obzira da li su riječi prethodno već viđene ili ne. Problem s ovim pristupom je da uobičajene teoretske postavke standardnih metoda zaglađivanja, poput Laplaceovog zaglađivanja, više nisu ispunjene jer LDA koristi mješavinu distribucija. Ipak ove se metode u praksi i dalje često koriste.

Bolja alternativa je primijeniti modifikaciju modela LDA, poznatu pod nazivom zaglađeni model LDA (engl. *Smoothed LDA model*). Ovaj je model prikazan na slici 3.7.



Slika 3.7: Grafički prikaz zaglađenog modela LDA.

U slučaju zaglađenog LDA, matrica β se tretira kao $k \times V$ slučajna matrica. Ova matrica ima po jedan redak za svaku komponentu mješavine distribucija. Pretpostavlja nezavisnost ovih redaka i da je svaki redak nezavisno uzorkovan iz izmjenjive Dirichletove distribucije, odnosno Dirichletove distribucije kod koje je $\alpha_i = \eta, \forall i$, koja ima samo jedan parametar, η.

4. Primjene

Latentnu Dirichletovu alokaciju moguće je primijeniti na širok spektar problema uključujući modeliranje dokumenata, klasifikacija dokumenata, modeliranje korisničkog ponašanja, itd. Općenitije, model LDA je primjenjiv na bilo koji problem gdje je dana kolekcija skupova diskretnih podataka, pri čemu su skupovi unutar kolekcije međusobno izmjenjivi i gdje su diskretni podatci unutar pojedinog skupa također međusobno izmjenjivi.

U nastavku poglavlja su dane neke primjene modela LDA na razne probleme obrade hrvatskog jezika.

4.1. Korišteni korpus

Za sva ispitivanja iznesena u nastavku korišten je zaglađeni model LDA nad korpusom na hrvatskom jeziku. Korpus koji je korišten za učenje zaglađenog modela LDA se sastoji od dvije rubrike Vjesnikovih novinskih članaka iz razdoblja od 1999. od 2009. godine. Kategorije koje su se koristile su "Crna kronika" i "Sport". U korpusu za učenje postoje 17564 članka iz kategorije "Crna kronika" i 27898 članaka iz kategorije "Sport". Korpus za učenje sadrži 80% ukupno dostupnih dokumenata. Preostalih 20% dokumenata čini skup za ispitivanje.

Na temelju korpusa za učenje, izgrađen je vokabular od 38349 riječi. Prilikom izgradnje ovog vokabulara korištena je lematizacija za hrvatski jezik (Snajder et al., 2008). Također, u vokabularu se ne nalaze brojevi, a ne izbacene su i riječi prisutne u manje od pet dokumenata, kao i riječi koje se nalaze u više od 50% dokumenata.

Na temelju skupa za učenje naučeno je nekoliko zaglađenih modela LDA. Učeni su modeli sa 2, 5, 15, 35 i 100 tema. Osim na skupu za učenje, napravljena je i usporedba modela sa 2, 5, 15, 35 i 100 tema pomoću metode unakrsne provjere (engl. *cross validation*). Provedena je unakrsna provjera od sedam koraka, gdje se pri svakom koraku $\frac{1}{7}$ inicijalnog skupa za učenje nije koristila za učenje već za provjeru modela. Usporedba modela provedena je temeljem postignute log-izglednosti na promatranom

korpusu:

$$log\, p(D|\alpha,\beta) = log\left\{ \prod_{d=1}^{M} \int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right) d\theta_d \right\} \quad (4.1)$$

Tablica 4.1 prikazuje dobivene rezultate. U oba su slučaja postignuti rezultati slični. Najbolji rezultati se postižu upotrebom modela LDA s pet tema, a što je broj tema dalji od ovog broja to su rezultati lošiji. Dobri rezultati LDA modela s relativno malo tema posljedica su korištenja malog korpusa za učenje. Kako je ponašanje na skupu za učenje sukladno ponašanju prilikom unakrsne provjere, slijedi da korišteni modeli LDA nisu skloni problemu prenaučenosti. Ovo je stoga eksperimentalna potvrda iste tvrdnje autora, (David M. Blei, 2003).

Tablica 4.1: Usporedba zaglađenih modela LDA s različitim brojem tema. Prikazni su rezultati usporedbe modela na cijelom skupu za učenje kao i rezultati dobiveni pomoću metode unakrsne provjere od sedam koraka. Rezultati su prikazani kao log-izglednosti dobivene na promatranim korpusima.

Broj tema	Skup za učenje	Unakrsna provjera
2	$-6.9907 \cdot 10^7$	$-1.0282 \cdot 10^7$
5	$-6.8693 \cdot 10^7$	$-1.0109 \cdot 10^7$
15	$-6.9241 \cdot 10^7$	$-1.0316 \cdot 10^7$
35	$-6.9854 \cdot 10^7$	$-1.0513 \cdot 10^7$
100	$-7.1308 \cdot 10^7$	$-1.0846 \cdot 10^7$

Ovi modeli LDA primjenjuju se u nastavku na izabrane probleme obrade hrvatskog jezika.

4.2. Provjera modela na generiranim dokumentima

U prethodnim je poglavljima nekoliko puta stavljan naglasak na činjenicu da je LDA generativan model. Stoga sada slijedi detaljniji empirijski pogled na ovo svojstvo modela LDA.

Kako je LDA usmjereni grafički model, moguće je primijeniti metodu generiranja uzoraka poznatu kao metoda roditeljskog uzorkovanja (engl. *ancestral sampling*) ili kao unaprijedno uzorkovanje (engl. *forward sampling*). Ovo je jednostavna metoda generiranja uzoraka, primjenjiva na uzorkovanje usmjerenih grafičkih modela. Usmjereni grafički modeli imaju svojstvo da svaki čvor ovisi samo o svojim roditeljima, a

kako u grafu nema ciklusa, postoji jasna podjela na roditelje i djecu. Stoga se metodom roditeljskog uzorkovanja čvorovi uzorkuju tako da se prvo uzorkuju roditelji, a tek onda djeca. Kad je potrebno uzorkovati djecu, već su poznate vrijednosti roditelja, te su tako svi potrebni podatci uvijek dostupni.

Konkretno, primjenom roditeljskog uzorkovanja na model LDA prilikom generiranja novog dokumenta prvo se uzorkuje distribucija tema θ za novi dokument. Potom se, prilikom generiranja novih riječi, za svaku riječ prvo uzorkuje tema z, a potom i sama riječ w. Nakon što je izgenerirano dovoljno riječi, proces generiranja novog dokumenta je završen.

U empirijskoj provjeri generativnog svojstva zaglađenog modela LDA navedenoj u nastavku korišteni su naučeni modeli LDA opisani u prethodnom odjeljku. Prvo se naučeni modeli LDA koriste da bi se generirali novi dokumenti na temelju postojećih, a potom se modeli LDA primjene na generirane dokumente, te se uspoređuju udjeli pojedinih tema između originalinih i generiranih dokumenata. Ako je LDA ispravan generativan model, mora na generiranim dokumentima dobiti distribucije tema veoma slične distribucijama tema originalnih dokumenata.

Točnije, na temelju originalnih dokumenata odredi se dužina novih dokumenata i udjeli pojedinih tema. U provedenoj provjeri je duljina generiranih dokumenata jednaka duljini originalnih dokumenata. Nakon što se odrede udjeli pojedinih tema θ originalnog dokumenta, isti se odbacuje. Potom se na temelju preuzete distribucije tema θ generira potreban broji riječi tako da se uzorkuje tema z koja je odgovorna za nastanak nove riječi, te se konačno odredi novo-generirana riječ w.

Usporedba distribucija tema originalnog i generiranog dokumenta se vrši primjenom simetrične Kullback-Leiblerove divergencije, koja je uobičajene mjera sličnosti za tematske modele, (Mark Steyvers, 2007). Simetrična Kullback-Leiblerova divergencija je definirane s:

$$\widehat{D}_{KL}(p,q) = \frac{1}{2}\big[D_{KL}(p,q) + D_{KL}(q,p)\big] \tag{4.2}$$

pri čemu je normalna Kullback-Leiblerova divergencija definirana s:

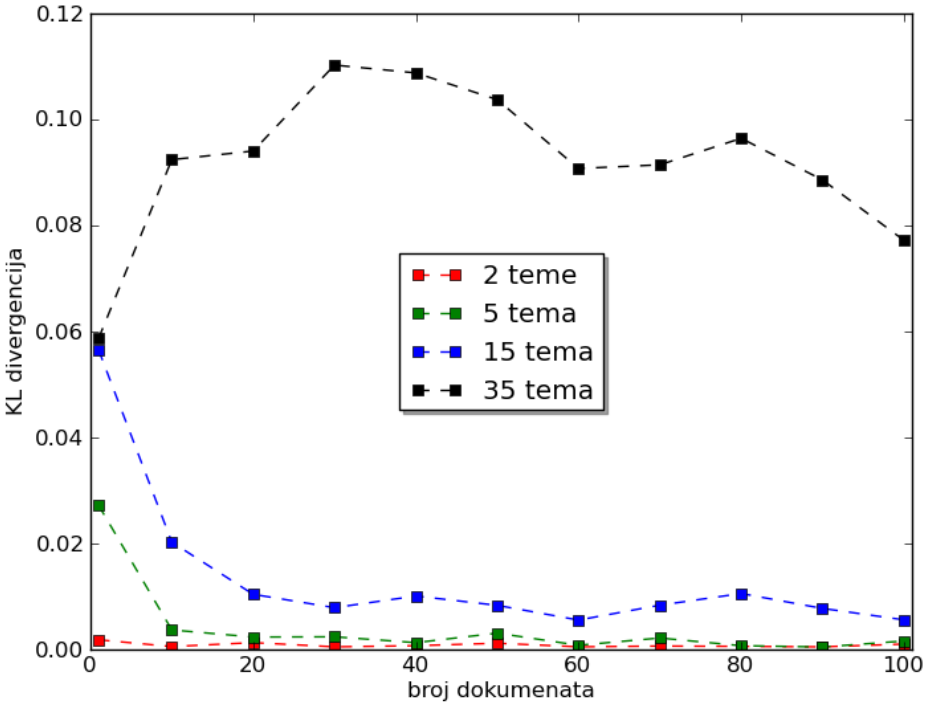
$$D_{KL}(p,q) = \sum_i p_i \, log\frac{p_i}{q_i} \tag{4.3}$$

U gornjim se izrazima pretpostavlja da vrijedi $0 \; log \; 0 = 0$. Ukoliko je promatrani i-ti element samo jedne od distribucija p i q jednak nuli, Kullback-Leiblerova divergencija nije definirana.

Novinski su članci obično veoma kratki, a kako pouzdanost postavljene distribucije tema ovisi o broju riječi u članku, u testu se umjesto originalnog članka slučajno

(uniformno, iz zadanog korpusa) odabere zadani broj članaka te se oni spoje u jedan dokument.

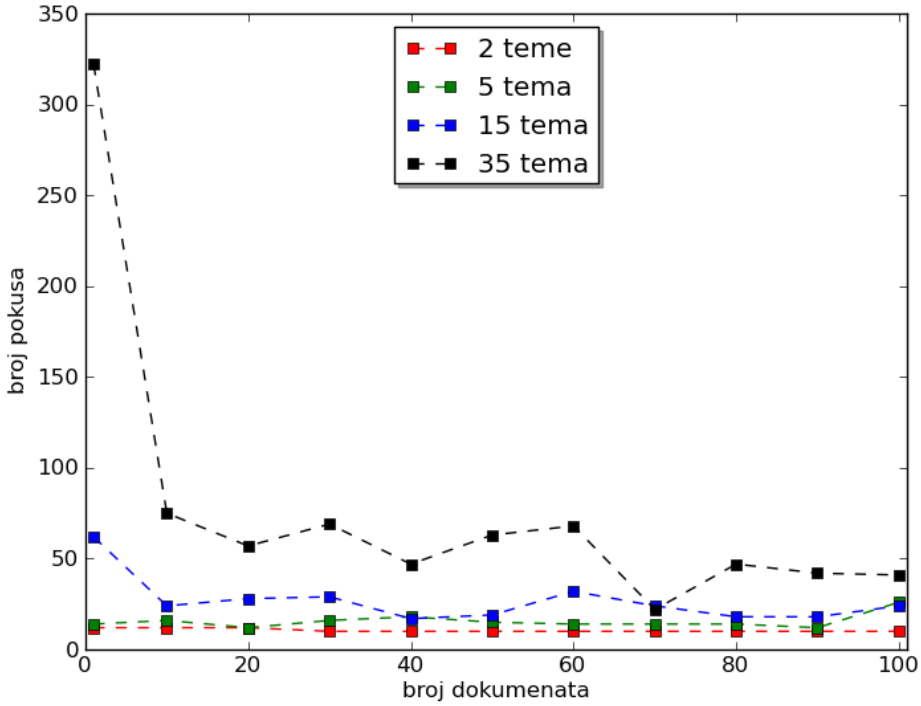
Slika 4.1 prikazuje rezultate ispitivanja generativnog ponašanja zaglađenog modela LDA. Prikazani su rezultati za modele LDA sa 2, 5, 15 i 35 tema. Kullback-Leiblerove divergencije za model s 35 tema su, očekivano, značajno veće od ostalih, pošto je korišteni korpus za učenje premalen za toliki broj tema, a isto se proizlazi i iz log-izglednosti tog modela (tablica 4.1).



Slika 4.1: Graf prikazuje Kullback-Leiblerovu divergenciju između distribucije tema izvornog dokumenta i generiranog dokumenta u ovisnosti o broju novinskih članaka koji čine izvorni dokument. Krivulje predstavljaju ponašanje naučenih modela LDA s različitim brojem tema.

Kao mjera udaljenosti distribucija u ovom testu izabrana je simetrična Kullback-Leiblerova divergencija. KL-divergencija nije definirana u slučaju da jedna distribucija nekom događaju pridružuje vrijednost nula, a druga ne i tada su distribucije neusporedive. Zato je na slici 4.2 prikazan broj pokusa potreban da se napravi 10 uspješnih mjerenja KL divergencije u ovisnosti o broju članaka koji čine jedan izvorni dokument za generativni proces. Ove je rezultate potrebno uzeti u obzir prilikom interpretacije grafa na slici 4.1. Broj pokusa je značajno veći od 10 samo u slučaju modela LDA s 35 tema, što je sukladno s prethodno iznesenom pretpostavkom da je upotrebljeni korpus

za učenje premalen za 35 tema.



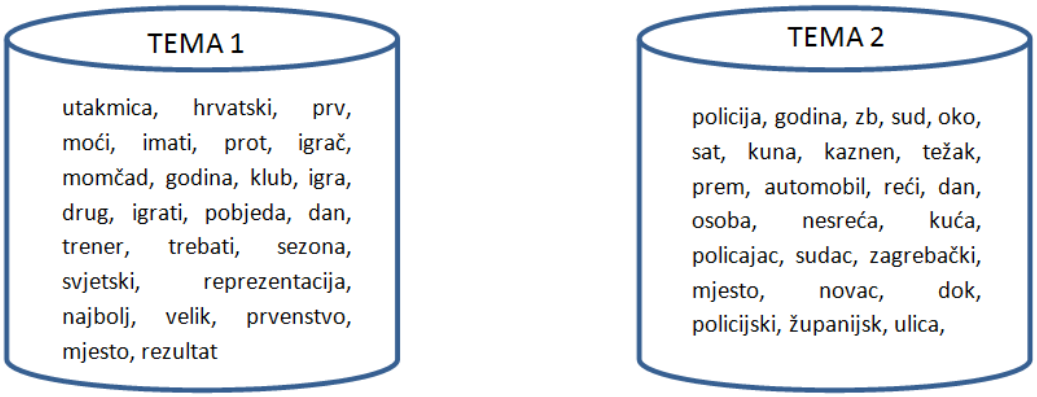
Slika 4.2: Graf prikazuje broj pokusa potreban da se napravi 10 uspješnih mjerena KL divergencije u ovisnosti o broju članaka koji čine izvorni dokument.

Kao što se može vidjeti na slici 4.1, zaglađeni modeli LDA postižu veoma malu Kullback-Leiblerovu divergenciju, što znači da uspješno rekonstruiraju početnu distribuciju iz generiranih dokumenata. Treba primijetiti da pouzdanost rekonstrukcije distribucije tema ovisi o veličini dokumenata korištenih u generativnom procesu, kao i o prikladnosti odabranog broja tema. Ovim je testom potvrđena ispravnost i smislenost modela LDA kao i njegova primjenjivost u generativnom smjeru.

4.3. Modeliranje dokumenata

Model LDA je tematski model čijom se primjenom na dokument dobije distribucija tema koje su inspirirale nastanak istog. Pri tome su teme zapravo vjerojatnosne distribucije nad riječima iz odabranog vokabulara. Posljedično, teme koje model LDA odabere za vrijeme učenja nemaju pripadne semantičke oznake koje bi sažele što pojedina tema zapravo izražava. Interpretacija semantičkih informacija koje teme nose u potpunosti je prepuštena korisniku.

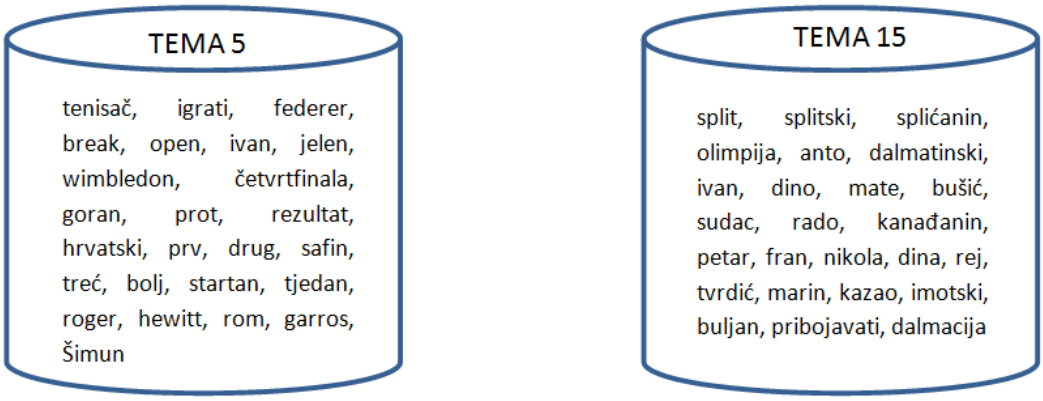
Na slici 4.3 prikazane su obje teme modela LDA s dvije teme. Kako su teme distribucije nad svim riječima vokabulara, prikazano je samo 25 najvjerojatnijih riječi iz svake teme. Već je prethodno prethodno rečeno da ove teme nemaju pridružena semantička tumačenja. Ipak, uzevši u obzir da je pripadni model LDA učen na člancima iz kategorija “Crna kronika” i “Sport”, nakon uvida u sadržaje tema nameće se pripisivanje jedne teme prvenstveno kategoriji “Sport”, a druge kategoriji “Crna kronika”. Tako primjerice prva tema daje veliku vjerojatnost riječima: utakmica, igrač, momčad i klub; dok druga tema daje veliku vjerojatnost riječima: policija, sud, nesreća i novac. Zanimljivo je primijetiti da obje teme daju veliku vjerojatnost riječi "mjesto". Ova činjenica potkrjepjuje već iznesenu tvrdnju da su teme vjerojatnosne distribucije nad riječima, a ne puka podjela riječi u kategorije. Iako samo spekulacija, razumno je pretpostaviti da se riječ "mjesto" u kontekstu kategorije “Sport” prvenstveno odnosi na ostvareni rank sudionika, dok se ta ista riječ u kontekstu kategorije “Crna kronika” odnosi na lokacije događaja. Pod ovom pretpostavkom slijedi da model LDA može donekle uhvatiti različita značenja iste riječi.



Slika 4.3: Slika prikazuje obje teme modela LDA s dvije teme. Svaka tema je predstavljena s 25 najvjerojatnijih riječi iz pripadnih distribucija. Prikazane riječi su lematizirane.

Tumačenje semantike tema postaje značajno kompliciranije kako broj tema raste. Na slici 4.4 prikazane su dvije teme modela LDA sa 100 tema. Uvidom u najvjerojatnije riječi svake teme, moguće je uočiti neke semantičke pravilnosti u temama. Primjerice tema 5 daje veliku vjerojatnost riječima: tenisač, break, wimbledon i četvrtfinala. Također spominju se osobna imena poznatih tenisača, poput: goran, roger, federer i safin. Stoga je prirodno pomisliti da se tema 5 bavi tenisom. Situacija je ponešto kompleksnija kod teme 15. Ova tema daje veliku vjerojatnost riječima poput: split, splitski, splićanin, dalmatinski, imotski i dalmacija. Također spominju se neka pretežno dalmatinska osobna imena: anto, dino i mate. Tako se nameće interpretacija

teme 15 kao teme koja se bavi Dalmacijom. No što je sa riječima: sudac, kanađanin, kazao i pribojavati? Veliku vjerojatnost ovih riječi pod temom koja se navodno bavi Dalmacijom nije jednostavno objasniti.



Slika 4.4: Slika prikazuje teme modela LDA sa 100 tema. Svaka tema je predstavljena s 25 najvjerojatnijih riječi iz pripadnih distribucija. Prikazane riječi su lematizirane.

Prethodna analiza tema koje koristi model LDA sugerira da treba biti oprezan prilikom pokušaja tumačenja semantike istih. Nepažljivom analizom korisnik bi mogao pomisliti da se Dalmtinci izrazito vole parničiti, da su bojažljivi i da svi imaju rodbinu u Kanadi. Neispravnost ovakvih generalizacija je očigledna.

Ukoliko se udjeli tema koriste za uspoređivanje sličnosti dokumenata javlja se nekoliko problema. Prvo, velik broj dokumenata postaje neusporediv jer ne djele iste teme, odnosno jedan dokument pridruži vjerojatnost 0 promatranoj temi, dok drugi dokument pridruži pozitivnu vjerojatnost istoj temi. Dokle god dokumenti dijele barem jednu zajendiču teme, donekle su usporedivi, no kako broj tema raste, velika većina dokumenata koristi samo manji podskup tema. Posljedično sve veći udio dokumenata postaje potpuno neusporediv.

Sljedeći problem koji se javlja je veoma teško semantičko tumačenje velike sličnosti dokumenata koje dijele iste teme. Primjerice, jedan (slučajno odabrani) članak govori o pronalasku trupla u kanalu. U ovisnosti o upotrebljenom modelu LDA s različitim brojem tema, najsličniji članci odabranom govore o ranjavanju osobe u eksploziji benzinske crpe, o dojavi o podmetnutoj bombi ili o ranjavanju vatrenim oružjem. Uzme li se u obzir da su teme samo vjerojatnosne distribucije nad riječima, a da se velikom broju riječi dodjeljuju slične vjerojatnosti, ovakvi rezultati nisu tako neobični.

Uzevši u obzir prethodno iznesene rezultate, nameće se zaključak da, iako model LDA (kao predstavnik tematskih modela) može donekle pružiti semantičke informacije o dokumentima, prilikom tumačenja ovakvih informacija i donošenja zaključaka na

temelju istih, treba biti veoma oprezan.

4.4. Klasifikacija dokumenata

Model LDA nije klasifikator, te stoga nije odmah očito gdje bi se mogao primjenjivati u procesu klasifikacije dokumenata. Odgovor leži u reprezentaciji dokumenata putem distribucije tema θ.

Najjednostavniji izbor značajki za potrebe klasifikacije dokumenata je direktna primjena vreće riječi – dokument se predstavlja kao vektor frekvencija riječi, to jest za svaku riječ iz vokabulara bilježi se koliko se puta pojavila u dokumentu. Popularna alternativa direktnoj primjeni vreće riječi kao značajki za klasifikacijski proces su tf-idf (engl. *term frequency–inverse document frequency*) značajke. Ove značajke veću težinu daju diskriminatnim riječima te su stoga najčešće bolji izbor od vreće riječi. Ipak, i ovaj model značajki, kao i vreća riječi koristi vektor značajki s po jednim elementom za svaku riječ iz vokabulara i ukoliko se koriste veliki vokabulari postaje izrazito nezgrapan.

S druge strane, broj tema koje se primjenjuju u modelu LDA je uvijek značajno manji od veličine korištenog vokabulara, a svaki se dokument može predstaviti pripadnom distribucijom tema. Tako LDA reducira dimenzionalnost i vrši selekciju značajki.

Veličina vokabulara u ozbiljnijim primjena kreće se od nekoliko desetaka tisuća do nekoliko stotina tisuća ili čak milijuna riječi. Nasuprot tome, broj tema u tematskim modelima ne prelazi nekoliko stotina. LDA stoga postiže značajnu redukciju dimenzionalnosti vektora značajki, no postavlja se pitanje koliko se informacija takvom redukcijom gubi, odnosno čini li distribucija tema za pojedini dokument kvalitetan skup značajki. Da bi se odgovorilo na ovo piranje u nastavku su iznesen rezultati eksperimenta u kojem se vrši klasifikacija dokumenata, a za značajke klasifikatora koriste se LDA značajke (distribucija tema u dokumentu), značajke vreće riječi i tf-idf značajke.

Za potrebe klasifikacije koristi se stroj s potpornim vektorima (engl. *Support vector machine*, SVM). SVM je veoma popularan klasifikator opće namjene, a posebno je pogodan kao klasifikator u ovom eksperimentu jer je konzistentan i uvijek pronalazi najbolje globalno rješenje pripadnog optimizacijskog problema. Primjenom klasifikatora SVM s različitim značajkama moći će se dobiti kvalitetna usporedba ekspresivnosti istih za potrebe klasifikacije. Klasifikator SVM može reproducirati i nelinearne granice, no u ovom eksperimentu se to svojstvo neće koristi jer zahtjeva podešavanje dodatnih parametara o čijem odabiru ovise i performanse klasifikatora te otežava uspoređivanje kvalitete značajki.

U tablici 4.2 dani su rezultati primjene klasifikatora SVM na problem binarne klasifikacije članaka u kategorije "Crna kronika" odnosno "Sport".

Tablica 4.2: Rezultati klasifikacije strojem s potpornim vektorima uz korištenje različitih vektora značajki.

Značajke	Skup za učenje	Skup za ispitivanje
vreća riječi	100%	99.8944%
tf-idf	99.9626%	99.912%
LDA, 2 teme	99.7272%	99.78%
LDA, 5 tema	99.5205%	99.6657%
LDA, 15 tema	99.4347%	99.5865%
LDA, 35 tema	99.3665%	99.4809%
LDA, 100 tema	99.4567%	99.5249%

Očekivano, tf-idf značajke postižu najbolje rezultate na skupu za ispitivanje, a najbolje rezultate na skupu za učenje postižu značajke vreće riječi. Oba tipa značajki imaju neznatno slabije rezultate na skup za ispitivanje nego na skupu za učenje. Zanimljivo je primijetiti da u slučaju LDA značajki (distribucija tema), rezultati na skupu za ispitivanje ne opadaju. Upotrebom LDA značajki dobivenih primjenom svih pet naučenih modela LDA postižu se lošiji rezultati nego primjenom značajki tf-idf ili značajki vreće riječi. Ipak, razlika nije drastična, a redukcija u dimenzionalnosti je veoma velika. Zanimljivo je primijetiti da izbor broja tema u modelima LDA utječe na performanse samo u manjoj mjeri. Posebno je zanimljiv slučaj upotrebe značajki modela LDA sa samo dvije teme. U ovom slučaju vektor značajki sastoji se samo od dva broja (udjela svake od dvije teme u dokumentu), a postignute performanse su usporedive s rezultatima klasifikatora s tf-idf značajkama ili značajkama vreće riječi koje koriste vektor značajki od 38349 elemenata.

Efektivna redukcija dimenzionalnosti je još veća, jer se vektor značajki za model LDA sa K tema može predstaviti sa $K - 1$ brojem, udjelom svih tema osim jedne, primjerice zadnje, teme jer zbroj udjela svih tema mora biti jedan. Tako se u slučaju modela LDA s dvije teme dimenzionalnost vektora značajki reducira sa 38349 na 1. Ova nevjeronatna redukcija dimenzionalnosti može, ovisno o vrsti klasifikatora, imati izrazito značajne posljedice na peformanse procesa klasifikacije. Proces klasifikacije se može ubrzati za isti faktor za koji je reducirana veličina vektora značajki. Tako se u slučaju modela LDA s dvije teme može postići ubrzanje od četiri reda velićine.

Na temelju gore navedenog eksperimenta može se zaključiti da je model LDA

veoma perspektivna metoda redukcije dimenzionalnosti značajki za potrebe klasifikacije dokumenata. Ipak je potrebno naglasiti da je klasifikacijski zadatak opisan u ovom poglavlju veoma jednostavan i potrebno je provesti daljnja istraživanja.

5. Zaključak

U današnje vrijeme, sve češće je potrebno obraditi izrazito velike količine podataka. Najveći dio dostupnih podataka tekstovnog su tipa. Da bi se uopće moglo početi obrađivati, obilje tekstovnih podataka većinom treba filtrirati ili razvrstati u nekakve kategorije dokumenata. Također, često je iz dokumenata potrebno izvući određene semantičke podatke. Jedan pristup rješavanju ovih problema jest upotreba tematskih modela.

Tematski modeli su vrsta generativnih modela s latentnim varijablama i spadaju u skupinu usmjerenih grafičkih modela, koji su podvrsta statističkih modela. Ovi se modeli temelje na pretpostavkama da se semantičke informacije mogu dobiti na temelju frekvencije pojave riječi u dokumentu i da je smanjenjem dimenzionalnosti moguće očuvati semantičke informacije.

Centralni pojam u tematskim modelima je tema. Tema je vjerojatnosna distribucija nad riječima iz odabranog vokabulara. Da bi se primijenili tematski modeli, dokument se promatra kao mješavina nekoliko tema, a upotrebom modela otkriva se koje su konkretne teme prisutne u promatranom dokumentu.

U ovom su radu proučavani tematski modeli i njihova primjena na dokumente hrvatskog jezika. Kao primjeri tematskih modela proučavani su probabilistička semantička analiza (pLSA) i latentna Dirichletova alokacija (LDA). Model latentne Dirichletove alokacije primijenjen je na dokumente hrvatskog jezika. Model LDA upotrebljen je i kao sredstvo za redukciju dimenzionalnosti reprezentacije dokumenta, pri čemu su dobiveni dobri rezultati koji upućuju na perspektivnost upotrebe tematskih modela u postupku pretprocesiranja i filtriranja tekstovnih dokumenata.

Sljedeći korak bio bi testiranje modela LDA na značajno većem korpusu, a posljedično i većem vokabularu. Tako bi se dobio uvid u ponašanju modela u realnijim uvjetima. Potrebno je provesti iscrpno testiranje performansi modela, posebice brzine računanja udjela tema za nove dokumente. Upotrebu modela kao sredstva za redukciju dimenzionalnosti dokumenata potrebno detaljnije ispitati. Konačno, potrebno je automatizirati proces određivanja optimalnog broja tema.

LITERATURA

Michael I. Jordan Andrew Y. Ng. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, 2001.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Michael I. Jordan David M. Blei, Andrew Y. Ng. Latent dirichlet allocation. *Journal of Machine Learnigng Research*, 2003.

B. de Finetti. *Theory of probability Vol. 1-2*. John Wiley and Sons, 1990.

Thomas Hofmann. Probabilistic latent semantic analysis, 1999.

Ata Kaban Mark Girolami. On an equivalence between plsi and lda, 2003.

Tom Griffths Mark Steyvers. Probabilistic topic models, 2007.

N.W. Henry P.F. Lazarsfeld. *Latent structure analysis*. Houghton Mifflin, 1968.

J. Snajder, B. Dalbelo Basic, and M. Tadic. Automatic acquisition of inflectional lexica for morphological normalisation. *Information Processing and Management*, 44(5): 1720 – 1731, 2008.

V.N. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, 1998.

Primjena tematskih modela na analizu dokumenata na hrvatskom jeziku

Sažetak

Generativni modeli s latentnim varijablama statistički su modeli podataka koji podatke opisuju temeljem njihovih skrivenih odnosno latentnih svojstava. Tematski modeli (engl. *topic models*) vrsta su generativnih modela s latentnim varijablama koji omogućavaju modeliranje apstraktnih tema sadržanih u tekstu dokumenta. Dana je teorijska podloga tematskih modela kao i njihov smještaj unutar većih grupa statističkih modela. Proučavane su teoretske osnove modela probabilističke semantičke analize (pLSA) i latentne Dirichletove alokacije (LDA). U eksperimentalnom dijelu pokazana je ispravnost generativnog smjera modela LDA i rezultati primjene istog na modeliranje dokumenata hrvatskog jezika. Na kraju je demonstrirana perspektivnost modela LDA za redukciju dimenzionalnosti reprezentacije dokumenata.

Ključne riječi: tematski modeli, hrvatski jezik, LDA, pLSA

Application of Topic Models to Analysis of Croatian Documents

Abstract

A latent variable model is a generative statistical model that relates a set of observable variables to a set of latent variables. A topic model is a type of latent variable model for discovering the abstract topics that occur in a collection of documents. Description of topic models is given. Theoretical foundations of Probabilistic latent semantic analysis model (pLSA) and Latent Dirichlet allocation model (LDA) are presented. Generative ability of LDA model is demonstrated. To model documents written in Croatian language, various LDA models are used. Demonstration of LDA model's applicability to reduction of document's dimensionality is given.

Keywords: topic models, Croatian language, LDA, pLSA