Laboratorij za analizu teksta i inženjerstvo znanja – TakeLab

Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
Zavod za elektroniku, mikrolektroniku, računalne i inteligentne sustave
Unska 3, 10000 Zagreb, Hrvatska

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 631

Primjena modela djelomične pripadnosti za ekstrakciju ključnih fraza iz dokumenata

Filip Petkovski

Zagreb, listopad 2013.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA
ODBOR ZA DIPLOMSKI RAD PROFILA

Zagreb, 12. ožujka 2013.

DIPLOMSKI ZADATAK br. 631

Pristupnik: **Filip Petkovski**
Studij: Računarstvo
Profil: Računarska znanost

Zadatak: **Primjena modela djelomične pripadnosti za ekstrakciju ključnih fraza iz dokumenata**

Opis zadatka:

Učinkovito dohvaćanje informacija često podrazumijeva prethodno označavanje dokumenta ključnim riječima ili frazama koje najbolje opisuju sadržaj dokumenta. Zadatak automatske ekstrakcije ključnih fraza jedan je od osnovnih zadataka ekstrakcije informacija. Suvremeni postupci esktrakcije ključnih fraza temelje se na statističkim metodama i metodama strojnog učenja. Metode temeljene na nenadziranom strojnom učenju posebno su prikladne jer ne iziskuju označene podatke.
U okviru diplomskog rada potrebno je proučiti probabilističke modele grupiranja podataka, s naglaskom na bayesovski model djelomične pripadnosti, te načiniti usporedbu tih modela. Proučiti postupke za ekstrakciju ključnih fraza s naglaskom na nenadzirane postupke. Razraditi postupak za modeliranje dokumenata korištenjem modela djelomične pripadnosti te postupak nenadzirane ekstrakcije ključnih fraza temeljen na tom modelu. Načiniti programsku implementaciju postupka i primijeniti ga na zbirku dokumenata na hrvatskome jeziku. Provesti temeljito vrednovanje razvijenog postupka na ispitnoj zbirci Hine, usporediti postupak s referentnim postupcima te provesti analizu značajki i pogrešaka. Radu priložiti izvorni programski kod, programsku dokumentaciju i korištene skupove podataka.

Zadatak uručen pristupniku: 15. ožujka 2013.
Rok za predaju rada: 28. lipnja 2013.

Mentor:

Doc.dr.sc. Jan Šnajder

Djelovođa:

Doc.dr.sc. Tomislav Hrkać

Predsjednik odbora za diplomski rad profila:

Prof.dr.sc. Siniša Srbljić

Zahvaljujem svom mentoru doc. dr. sc. Janu Šnajderu strpljenju, pomoći i vodstvu pri izradi ovog diplomskog rada

SADRŽAJ

1. Uvod

Danas su dostupne goleme količine pisanog teksta koje sadrže različite vrste informacija, počevši od autora teksta, pa sve do činjenica, događaja i entiteta sadržanih u tekstu. Problem automatskog izvlačenja ključnih riječi iz teksta bavi se određivanjem riječi i izraza sadržanih u dokumentu koji najbolje opisuju dani dokument. Poznavanje takvih riječi od velike je važnosti jer nam govori je li nam određeni dokument bitan ili nebitan, odnosno daje li nam određeni dokument bitne informacije za zadanu temu.

Većina metoda za automatsku ekstrakciju ključnih izraza zahtijeva označavanje skupa podataka za učenje, zadatak koji je vrlo subjektivan, vremenski zahtjevan i skup. U ovom su radu korištene jedino nenadzirane metode strojnog učenja, dok je mali skup označenih dokumenata korišten isključivo za vrednovanje. Vrednovanje, kao učenje, je provedeno nad podacima hrvatskog jezika.

U okviru rada primijenjen je model djelomične pripadnosti za izvlačenje ključnih riječi iz dokumenata. Obzirom da se ovaj postupak pokazao neuspješnim, u radu je nastavljeno s primjenom modela Latentne Dirichletove alokacije, te su razvijene dvije metode koje koriste parametre procijenjene ovim modelom. Nadalje, provedena je evaluacija razvijenih metoda te analiza dobivenih rezultata.

Ovaj je rad strukturiran na sljedeći način: u drugom je poglavlju opisan model djelomične pripadnosti, počevši od standardnog modela mješavine konačnog broja gustoća. U trećem je poglavlju opisan model LDA te je napravljena usporedba s modelom djelomične pripadnosti. Četvrto poglavlje opisuje razvijene metode te način kako se one mogu iskoristiti za ekstrakciju ključnih riječi. Konačno, u zadnjem se poglavlju opisuje provedeno vrednovanje te dobiveni rezultati.

2. Srodni radovi

Witten et al. (1999) su među prvima razvili sustav za ekstrakciju ključnih riječi koji su nazvali KEA (Keyword Extracton Algorithm). KEA koristi statistički model naučen pomoću označenih podataka specifični za zadanu domenu. Autori su u njihovom sustavu koristili naivni Bayesov klasifikator dok su podatke predstavili vektorom TF-IDF (citesalton1988term). KEA je ograničen na ekstrakciju izraza koji se sastoje od najviše tri riječi, te se broj ključnih izraza treba zadati na početku algoritma. KEA je danas, zbog svoje jednostavnosti, jedna od popularnijih metoda, te se njezina implementacija može naći u obliku XML-RPC servisa[1]

Za razliku od Witten et al. (1999) koji su koristili statistički model, van der Plas et al. (2004) su pokušali iskoristiti leksičke resurse: riječnik grupe za istraživanje i razvoj elektroničkih riječnika (EDR riječnik) [2] te WordNet Sveučilišta Princeton [3]. Oba resursa definiraju slične relacije između entiteta (X je Y, X je do Y, itd.) pa su autori odlučili usporediti uspješnost korištenja jednog i drugog resursa. Vrednovanje je pokazalo da je WordNet prikladniji alat za rješavanje problema ovakvog tipa, dok su oba leksička resursa dala bolje rezultate od obične TF-IDF mjere.

Nadzirane metode za ekstrakciju ključni riječi koje se zasnivaju na strojnom učenju često uključuju korištenje stabala odluke (engl. *decision trees*), naivnog Bayesovog klasifikatora (engl. *naive Bayes classifier*) ili stroja s potpornim vektorima (engl. *support vector machine*). S druge strane, nenadzirane metode obično predstavljaju podatke u obliku TF-IDF vektora te koriste algoritme za grupiranje podataka kako bi otkrili semantički povezane riječi, rezultat koji se nakon toga koristi za ekstrakciju ključnih riječi. Pregled nenadziranih metoda za ekstrakciju ključnih riječi može se naći u Hasan i Ng (2010).

Ovaj se rad nadovezuje na metode opisane u (Saratlija et al., 2011) gdje se riječi dokumenata grupiraju čime se dobivaju grupe tematski povezanih riječi. Nadalje se po-

[1] http://metaoptimize.com/blog/2010/08/18/kea-keyphrase-extraction-as-an-xml-rpc-service/

[2] http://www.edrdg.org/

[3] http://wordnet.princeton.edu/

moću tih grupa iz svakog dokumenta izvlače ključne riječi te se iste na kraju proširuju u ključne izraze. Ostale metode koje su razvijene za Hrvatski jezik uključuju rad (Ahel et al. (2009)), metoda temeljena na TF-IDF zapisu dokumenata, te rad (Mijic et al. (2010)) u kojem je opisana metoda koja koristi Bayesov klasifikator.

3. Model djelomične pripadnosti

Model djelomične pripadnosti (engl. *partial membership model*) (Heller et al., 2008) je statistički generativni model koji svaki podatak generira pomoću više podatkovnih izvora. Za razliku od modela konačnog broja gustoća kod kojeg svaki podatak proizlazi iz točno jednog izvora, kod modela djelomične pripadnosti svaki izvor utječe na konačnu vrijednost generiranog podatka. Kao posljedica toga, ovaj nam model daje znatno veću fleksibilnost u odnosu na standardne modele mješavine konačnog broja gustoća.

U ovom ćemo poglavlju najprije spomenuti modele konačnog broja gustoća, a nakon toga ćemo neke njihove pretpostavke poopćiti kako bismo došli do modela djelomične pripadnosti. Također, objasnit ćemo osnovnu strukturu te način procjene parametara kod ovih modela.

3.1. Modeli mješavine konačnog broja gustoća

Modeli mješavine konačnog broja gustoća (engl. *Finite Mixture Models* (FMM)) (McLachlan i Peel, 2004) su generativni statistički modeli koji predstavljaju moćan alat za modeliranje podataka jedne ili više varijabli te je njihova primjena prisutna u gotovo svim područjima gdje je potrebna neka vrsta modeliranja podataka. U statističkoj obradi podataka, modeli mješavine konačnog broja gustoća predstavljaju formalan pristup nenadziranom učenju. U takvim je modelima svaki primjer iz skupa podataka modeliran kao uzorak jednog od više alternativnih izvora podataka. Određivanje izvora svakog podatkovnog primjera i procjena parametara svakog izvora predstavlja grupiranje skupa podataka u više nezavisnih grupa. Zbog formalnog pristupa grupiranju podataka, kod ovih se modela određivanje broja grupa te evaluacija dobivene particije može također provesti na formalan način. Ovi modeli nisu jedino korisni za grupiranje, već se ujedno mogu primijeniti i na problem procjene gustoće vjerojatnosne distribucije. Činjenica da broj mješavina može biti proizvoljan omogućava modeliranje vrlo složenih podatkovnih skupova.

Vjerojatnost generiranja točke *y* modelom mješavine konačnog broja gustoća određena je jednadžbom:

$$p(y|\theta) = \sum_{k=1}^{K} \pi_i \cdot p(y|\theta_k) \tag{3.1}$$

gdje

$$\pi_i \geq 0, \sum_{k=1}^{K} \pi_k = 1 \tag{3.2}$$

Vektor $\theta = [\theta_1, ..., \theta_K]$ sadrži parametre svake pojedine grupe, dok vektor $\Pi = [\pi_1, ..., \pi_K]$ određuje s kojom težinom svaki izvor (grupa) sudjeluje u vjerojatnosnoj distribuciji $p(y|\theta)$. Funkcija $p(y|\theta)$ zove se još i funkcija izglednosti (engl. *likelihood function*) skupa podataka.

Najčešći način procjene parametara nekog modela jest metoda najveće izglednosti (engl. *maximum likelihood*), koja pretpostavlja da je dani skup podataka najvjerojatniji skup koji se može dobiti. U tom su slučaju traženi parametri oni parametri koji maksimiziraju vrijednost funkcije izglednosti. U većini je slučajeva nemoguće odrediti derivaciju funkcije izglednosti, pa se zbog toga optimira logaritam funkcije izglednosti, odnosno log-izglednost skupa podataka (engl. *log-likelihood*).

Log-izglednost N nezavisnih i identično raspodijeljenih (engl. *identically and independently distributed (iid)*) primjera dana je formulom:

$$\log p(Y|\theta) = \log \prod_{i=1}^{N} p(y_i|\theta) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} \pi_i \times p(y_i|\theta_k) \tag{3.3}$$

Poznato je da se vrijednost θ koja maksimizira ovu log-izglednost ne može odrediti analitički, pa se zbog toga moraju koristiti aproksimativne metode.

Vjerojatnosna funkcija $p(y|\theta)$ može se zapisati i na drugi način koristeći skup latentnih varijabli. Latentne, ili skrivljene varijable su varijable koje se ne mogu izravno procjeniti iz skupa podataka, već se njihova vrijednost određuje pomoću vrijednosti ostalih, *vidljivih* varijabli. U ovom se slučaju mogu uvesti latentne varijable $\vec{Z} = [z_1, ...z_N]$ gdje je svaki z_i binarni vektor dimenzionalnosti K s točno jednom jedinicom dok su ostali njegovi članovi jednaki nuli. Jedinica na k-tom mjestu u binarnom vektoru označava da je primjer generiran k-tim izvorom. Svakom se primjeru iz

skupa podataka pridružuje jedna latentna varijabla koja određuje kojoj grupi primjer pripada. Formalno, vjerojatnost generiranja primjera y_i može se zapisati kao:

$$p(y|\theta) = \sum_z p(z) \cdot \prod_{k=1}^{K} p(y|\theta_k)^{z^{(k)}} \tag{3.4}$$

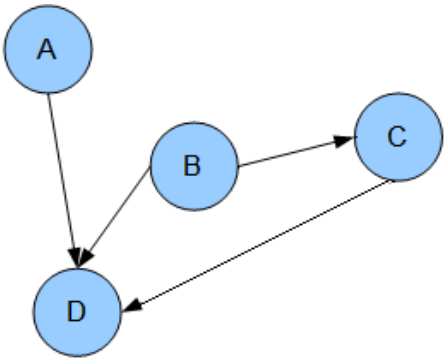
pri čemu smo iskoristili relaciju:

$$p(z^{(k)} = 1) = \pi_k \tag{3.5}$$

Standardni način procjene parametara modela mješavine konačnog broja gustoća jest algoritam *Expectation - Maximization* (EM-algoritam) (Moon, 1996), odnosno algoritam maksimizacije očekivanja. EM-algoritam je iterativni algoritam te predstavlja aproksimaciju procjene najveće izglednosti. Sastoji se od dva koraka koja uključuju izračun očekivanih pripadnosti koristeći trenutnih parametara te maksimizaciju parametara s procjenjenim pripadnosnim varijablama. Unatoč tome što je ovo najčešće korišten algoritam za učenje ovih modela, EM-algoritam jest po svojoj prirodi pohlepan te konvergira prema lokalnim ekstremima. Ovaj se problem u praksi pokušava ublažiti višekratnom nasumičnom inicijalizacijom parametara, slično kao kod heurističkih metoda poput grupiranja k-srednjih vrijednosti (engl. *k-means clustering*).

Najčešća alternativa EM-algoritmu su metode Markovljev lanac Monte Carlo, (*Markov Chain Monte Carlo* - MCMC) (Gilks et al., 1996; MacKay, 2003). Ove metode pokušavaju procjeniti aposteriornu razdiobu modela na način da uzimaju nasumične uzorke tvoreći Markovljev lanac. Najpopularnija MCMC-metoda jest Gibbsovo uzorkovanje, opisano u sljedećem poglavlju.

Slično kao kod ostalih parametarskih modela, i kod ovih modela postoji mogućnost prenaučenosti odnosno podnaučenosti, problem koji se u praksi rješava optimizacijom hiperparameta. Međutim, iako je kod nadziranih metoda taj postupak relativno jednostavan zbog prisutnosti oznaka, kod nenadziranih metoda je znatno teže odabrati *prave* hiperparametre. Primjerice, vrlo čest problem kod modela za grupiranje podataka jest odabir broja grupa jer sam problem nije dobro definiran. U rijetkom slučaju kada je poznata konačna particija, odnosno grupiranje podataka koje se želi dobiti, broj grupa je moguće optimirati unakrsnom provjerom (engl. cross-validation). Međutim, u praksi se najčešće želi istražiti skup podataka, pa se od nenadziranih metoda očekuje rezultat koji čovjeku nije unaprijed poznat. U tim se slučajevima primjenjuje model s više različitih hiperparametara te se analiziraju dobiveni rezultati.



Slika 3.1: Primjer Bayesove mreže

3.2. Bayesove mreže

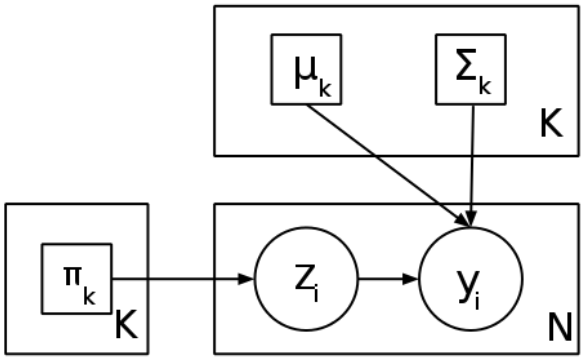
Svaki se model mješavine konačnog broja gustoća može prikazati grafički u obliku Bayesove mreže (Jensen, 1996). Bayesova mreža je aciklički usmjereni graf kojim je kodirana jedna jedinstvena vjerojatnosna razdioba. Čvorovi grafa predstavljaju varijable, dok bridovi označavaju veze između varijabli. Usmjeren brid od varijable A prema varijabli B označava uvjetnu zavisnost varijable B o varijabli A, dok nepostojanje brida označava uvjetnu nezavisnost. Primjer Bayesove mreže s četiri varijabli dan je na slici 3.1 Iz slike se može vidjeti da varijabla D uvjetno ovisi o svim ostalim varijablama, da varijabla C uvjetno ovisi o varijabli B, dok su varijable A i B uvjetno nezavisne od ostalih varijabli. Drugim riječima, vjerojatnosna distribucija sa slike 3.1 dana je formulom:

$$P(A, B, C, D) = P(A) \cdot P(B) \cdot P(C|B) \cdot P(D|A, B, C) \tag{3.6}$$

3.3. Model mješavine Gaussovih gustoća

Model mješavine Gaussovih gustoća (*Gaussian Mixture Model* (GMM)) (Reynolds, 2009) dobija se kada se svaki izvor podataka modelira Gaussovom razdiobom. Drugim riječima, za se svaki $p(y|\theta_k), i \in [1, .., K]$ uvrštava:

$$p(y|\theta_k) = g(y|\vec{\mu_k}, \Sigma_k) \tag{3.7}$$



Slika 3.2: Bayesova mreža mješavine konačnog broja Gaussovih gustoća

gdje je funkcija $g(y|\vec{\mu_k}, \Sigma_k)$ definira kao:

$$g(y|\vec{\mu_k}, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_k|}} exp(-\frac{1}{2}(y - \vec{\mu_k})^T \Sigma_k^{-1}(y - \vec{\mu_k})) \tag{3.8}$$

Vektor $\vec{\mu_k}$ je vektor srednjih vrijednosti k-tog izvora i definira oko koje će točke podaci biti generirani, dok je Σ_k kovarijacijska matrica i predstavlja raspršenost podataka koji će biti generirani tim izvorom.

Model Gaussovi mješavina može se pojednostaviti na više načina. Kovarijacijske matrice $\Sigma_{0,..K}$ mogu biti punog ranga ili dijagonalne. Nadalje, kovarijacijska matrica i ostali parametri mogu biti zajednički za sve izvore. Odabir složenijeg ili jednostavnijeg modela ponajviše ovisi o veličini skupa podataka, a procjena parametara se najčešće provodi EM-algoritmom.

Bayesova mreža ovog modela prikazana je na slici 3.2. Treba napomenuti da sve varijable unutar jednog pravokutnika ponavljaju se onoliko puta koliko je označeno u donjem desnom kutu pravokutnika, dok se sve konstante nalaze u kvadratićima. Iz bayesove slike se izravno može očitati vjerojatnosna razdioba skupa podataka $\mathcal{D} = \{y_1, ..., y_N\}$:

$$p(\mathcal{D}, \vec{z}|\vec{\mu}, \vec{\Sigma}, \vec{\pi}) = \prod_{i=1}^{N} \prod_{k=1}^{K} p(y_i|\mu_k, \Sigma_k, z_i) \cdot p(z_i|\pi_k) \tag{3.9}$$

U poslednjih se 10-ak godina sve češće koriste Bayesove inačice modela mješavina gustoća. Kod ovih inačica, parametri modela nisu konstante već vjerojatnosne

razdiobe. Nadalje, ovi modeli sadrže formalno ugrađen mehanizam zaglađivanja u obliku apriorne razdiobe nad parametrima $p(\theta_k)$. Zbog toga je u ove modele vrlo jednostavno dodati ekspertsko (*apriorno*) znanje o domeni, pa ih je stoga teže prenaučiti. S povećanjem broja podataka za učenje, utjecaj apriorne razdiobe se sve više smanjuje dok se utjecaj podataka za učenje sve više povećava.

Ideja Bayesovske procjene jest modeliranje svakog parametra u modelu kombiniranjem apriorne razdiobe s informacijama dobivenih pomoću skupa podataka, dok se generalni pristup procjene parametara može se sažeti na sljedeći način:

- Definirati apriornu razdiobu $p(\theta_k)$ nad parametrom θ_k
- Procijeniti parametre funkcije izlednosti $p(D|\theta_k)$ pomoću podataka za učenje $\mathcal{D}$
- Koristeći Bayesovog pravila, izračunati aposteriornu distribuciju $p(\theta_k|\mathcal{D})$

pri čemu je Bayesovo pravilo definirano kao:

$$p(\theta_k|\mathcal{D}) = \frac{p(\mathcal{D}|\theta) \cdot p(\theta_k)}{p(\mathcal{D})} \tag{3.10}$$

gdje je

$$p(\mathcal{D}) = \int_{\theta_k} p(\mathcal{D}|\theta_k) \cdot p(\theta_k) d\theta_k \tag{3.11}$$

Uvođenje dodatnih razdioba povećava složenost zaključivanja i procjenu parametara modela. Konkretnije, najveći problem kod Bayesovske procjene parametara jest izračun vjerojatnosti skupa podataka $p(\mathcal{D})$ jer je često integral umnoška apriorne razdiobe i funkcije izglednosti neizračunljiv. Ovaj se problem olakšava na način da se funkcija izglednosti $p(y|\theta_k)$ i apriorna razdioba $p(\theta_k)$ odabiru tako da imaju isti funkcijski oblik, odnosno da dolaze od iste porodice funkcija. Posljedica toga je da će njihov umnožak dati razdiobu koja će također imati isti funkcijski oblik kao i apriorna razdioba. Apriorna razdioba ovog tipa naziva se konjugirana aprioriorna razdioba (engl. *conjugate prior*).

Konkretno, kod Bayesovog modela mješavine Gaussovih gustoća, parametri se mogu modelirati na sljedeći način:

- Središnji vektor μ_k ima Gaussovu razdiobu s hiperparametrima $\vec{\mu_0}$ i $\vec{\sigma_0}$
- Kovarijacijska matrica Σ_k je uzorak iz Inverzne-Wishartove razdiobe s hiperparametrima Φ i η_0

– Vektor mješavina Π ima Dirichletovu razdiobu s hiperparametrom α

– Svaki uzorak y_i generiran je Gaussovom razdiobom s parametrima μ_k i Σ_k

Bitna razdioba u ovom modelu je Dirichletova razdioba, često označavana s $Dir(\vec{\alpha})$. Zanimljivo svojstvo ove razdiobe jest činjenica da je svaki njezini uzorak vektor koji također predstavlja vjerojatnosnu razdiobu. Stoga, logično je koristiti $Dir(\vec{\alpha})$ kao razdiobu parametra Π, jer njezinim uzorkovanjem izravno dobijemo težine pojedinih izvora podataka. Vektor $\vec{\alpha}$ izravno određuje očekivanu vrijednost svake komponente (dimenzije):

$$E[X_i] = \frac{\alpha_i}{\sum_k \alpha_k} \tag{3.12}$$

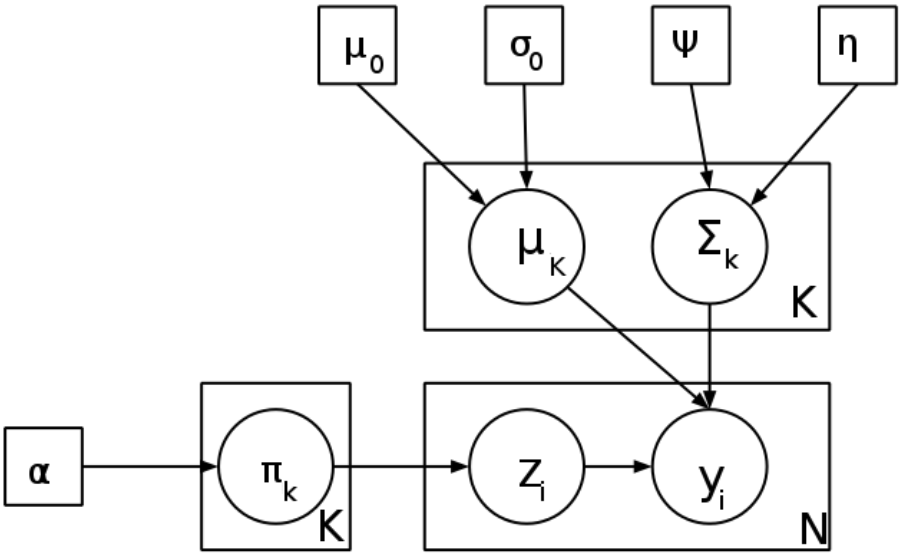
Kod Bayesovih se modela u većini slučajeva koristi simetrična Dirichletova razdioba kod koje su članovi vektora $\vec{\alpha}$ jednaki, odnosno $\vec{\alpha} = \{\alpha, ..., \alpha\}$ Parametar α se u tom slučaju naziva koncentracijski parametar jer određuje koliko će vjerojatnosna masa biti koncentrirana odnosno raspršena. Drugim riječim, ukoliko je vrijednost α-e puno manja od 1, mali broj komponenata će imati znatno veću vrijednost od ostalih, odnosno vjerojatnosna masa će biti koncentrirana u nekoliko komponenti. S druge strane, ukoliko je α puno veći od 1, komponente uzorka će imati približno iste vrijednosti, odnosno vjerojatnosna masa će biti raspršena po svim komponentama.

Bayesova mreža ovog modela prikazana je na slici 3.2, dok je vjerojatnost generiranja uzorka dana formulom:

$$\begin{aligned} p(\mathcal{D}, \vec{z}, \vec{\mu}, \Sigma, \vec{\pi} | \vec{\mu_0}, \vec{\sigma_0}, \Phi, \eta) = \prod_{i=1}^{N} \prod_{k=1}^{K} p(y_i | \vec{\mu_k}, \Sigma_k, \vec{z_i}) \\ \cdot p(\vec{\mu_k} | \vec{\mu_0}, \vec{\sigma_0}) \cdot p(\Sigma_k | \Phi, \eta) \cdot p(\vec{z_i} | \pi_k) \cdot p(\pi_k | \alpha) \end{aligned} \tag{3.13}$$

3.4. Model djelomične pripadnosti

Koncept djelomične pripadnosti je prilično intuitivan i praktičan. Primjerice, osoba može imati dvije različite etničke pozadine, pa možemo reći da ta osoba djelomično pripada jednoj i drugoj etničkoj grupi. Kada je u pitanju tekst, jedan dokument može



Slika 3.3: Bayesova mreža mješavine konačnog broja Gaussovih gustoća u Bayesovom okruženju

pripadati više kategorija ili tema, pa bi u tom slučaju model djelomične pripadnosti mogao biti koristan.

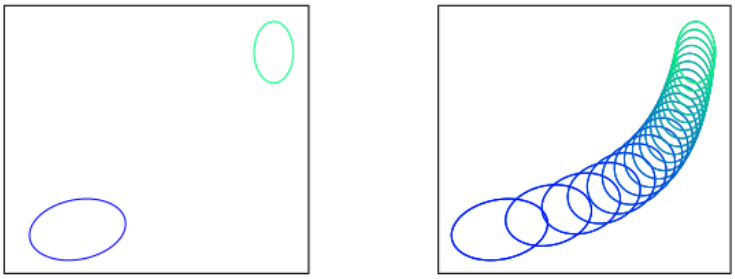
Djelomična pripadnost je konceptualno različita od nepoznate, odnosno nesigurne pripadnosti. Činjenica da je osoba X 60% Europljanin nije isto što i vjerojatnost od 60% da je osoba X Europljanin. Stoga, rezultati modela mješavine konačnog broja gustoća ne mogu se interpretirati na isti način kao i rezultati modela djelomične pripadnosti.

Model djelomične pripadnosti (engl *Partial Membership Model* (PMM)) (Heller et al., 2008) može se izvesti relaksiranjem modela mješavine konačnog broja gustoća. Krenimo od vjerojatnosne razdiobe za FMM:

$$p(y|\theta) = \sum_{i=1}^{K} \pi_k \cdot p(y|\theta_k) \tag{3.14}$$

koja se može zapisati pomoću latentnih varijabli kao:

$$p(y_i|\theta) = \sum_{z_i}^{K} p(z_i) \cdot \prod_{k=1}^{K} p(y_i|\theta_k)^{z_i^{(k)}} \tag{3.15}$$



Slika 3.4: Razlika između modela mješavine konačnog broja gustoća i modela djelomične pripadnosti (Heller et al., 2008)

Možemo primjetiti da ukoliko je $z_i^{(k)} = 1$, i-ti primjer pripada k-toj grupi. Također, bitno je napomenuti da $p(z_i)$ u ovom nije ništa drugo nego π_k, odnosno vjerojatnost da je $z_i^{(k)} = 1$ je jednaka p_k. Stoga varijabla z_i određuje pripadnost j-tog primjera pojedinim grupama. Ograničenje koje postoji u standardnom modelu mješavine gustoća jest to da su latentne varijable diskretne ($z_i^{(k)} \in \{0, 1\}$), pa je zbog toga moguća pripadnost samo jednoj grupi. Kako bismo dobili model djelomične pripadnosti, potrebno je relaksirati ovo ograničenje i dopustiti latentnim varijablama da poprime kontinuirane vrijednosti u rasponu $[0, 1]$. Međutim, kako bismo dobili valjanu formulu za vjerojatnost pojedinog primjera s ovim predpostavkama, potrebno je modificirati početni izraz na sljedeći način:

$$p(y_i|\theta) = \int_{z_i} p(z_i) \frac{1}{c} \prod_{i=1}^{K} p(y_i|\theta)^{z_i^{(k)}} dz_i \tag{3.16}$$

Navedena modifikacija uključuje integriranje preko cijelog područja mogućih vrijednosti z_i te dodatnu konstantu c čija je svrha normalizacija izraza kako bi se dobila valjana vjerojatnosna funkcija.

Razlika između standardnog i djelomičnog modela prikazana je na slici 3.4 Na slici su prikazane konture dviju mješavina Gaussovih gustoća. Na lijevoj su slici prikazane konture standardnog modela, dok su na desnoj slici prikazane konture djelomičnog modela te se oba modela sastoje od dvije Gaussove gustoće. Razlika između ova dva modela je u tome što model djelomične pripadnosti može generirati i bolje modelirati podatke koje se nalaze između grupa.

Bayesova mreža modela prikazana je na slici 3.5, dok je vjerojatnosna razdioba dana dana formulom:

$$p(\mathcal{D}, \vec{\mu}, \Sigma, \vec{z}, \vec{\pi} | \vec{\mu_0}, \vec{\sigma_0}, \Phi, \eta, a, b) = \prod_{i=1}^{N} \prod_{k=1}^{K} p(y_i | \vec{\mu_k}, \Sigma_k, \vec{z_i}) \\ \cdot p(\vec{\mu_k} | \vec{\mu_0}, \vec{\sigma_0}) \cdot p(\Sigma_k | \Phi, \eta) \cdot p(\vec{z_i} | \pi_k) \cdot p(\pi_k | \alpha, \rho) \cdot p(\alpha | b) \cdot p(\rho | a) \tag{3.17}$$

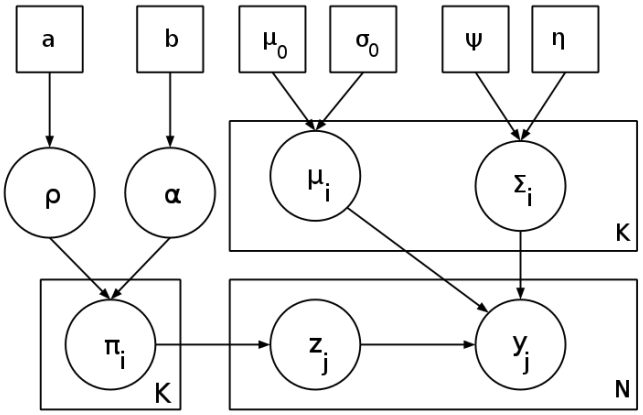
Jednom kada je definirana struktura modela, mogu se definirati razdiobe pojedinih varijabli. U ovom je radu korištena Bayesova inačica modela, što znači da su i sami parametri modela slučajne varijable, a pojedine su distribucije definirane na sljedeći način:

- Središnji vektor grupe μ_k ima Gaussovu razdiobu s hiperparametrima μ_0 i σ_0
- Kovarijacijska matrica Σ_k je uzorak iz Inverzne-Wishartove razdiobe s hiperparametrima Φ i η_0
- Vektor mješavina π ima Dirichletovu razdiobu s hiperparametrom α
- Parametar a je konstanta izvučena iz eksponencijalne distribucije be^{-ba}
- Vektor pripadnosti z_i ima Dirichletovu razdiobu s hiperparametrom $a \cdot \pi$
- Svaki uzorak y_i generiran je Gaussovom razdiobom s nekim parametrima μ_k i Σ_k

Dirichletova razdioba i kod ovog modela ima sličnu ulogu kao i kod Bayesove inačice modela konačnog broja gustoća.

Promotrimo kako model djelomične pripadnosti generira podatke. Neka je definiran model s K grupa te Y = $[y_1, ...y_N]$ podatkovnih primjera. Najprije se odabiru parametri pojedinih izvora kao uzorci Gaussove (μ_k) i Inverzne-Wishartove (Σ_k) razdiobe. Vektor mješavina π dobija se kao uzorak Dirichletove razdiobe. Nadalje, za svaki podatak y_i koji se želi generirati, odabire se vektor pripadnosti z_i kao uzorak Dirichletove razdiobe s parametrom π skaliran s konstantom a. Na kraju, podatak se generira pomoću svih K izvora s pripadnim težinama $\pi_1, ..., \pi_K$.

Intuitivno, parametri modela mogu se objasniti na sljedeći način. Parametri pojedinih izvora μ_k i Σ_k definiraju razdiobe značajki podatkovnih primjera. Vektor mješavina π definira makro kompoziciju cijelog skupa podataka (75% ljudi je iz A etničke pozadine, dok je 25% iz B etničke pozadine). Parametar skaliranja a govori koliko će pojedini uzorci biti slični generalnoj populaciji (hoće li svi ljudi biti 75% iz A i 25% iz B etničke pozadine?) Vektori pripadnosti $z_1, ..., z_N$ definiraju pripadnost pojedinih podatkovnih primjera različitim izvorima.



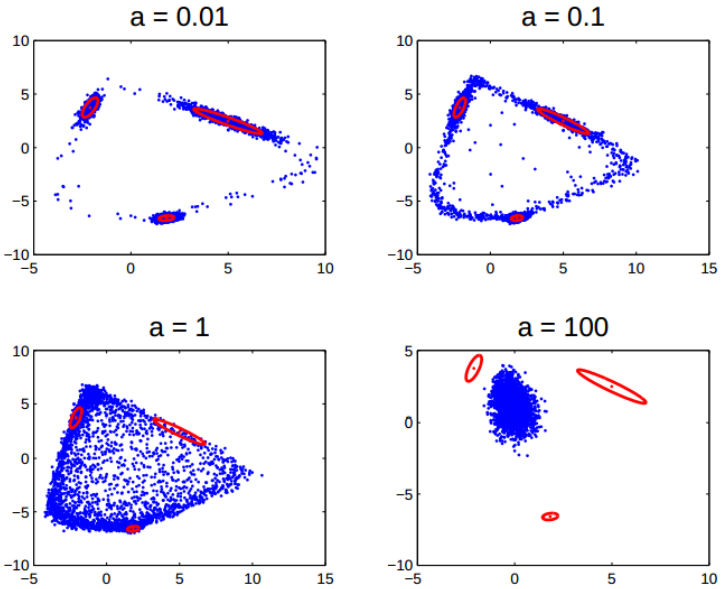
Slika 3.5: Bayesova mreža modela djelomične pripadnosti

Primjeri sintetičkih skupova podataka generirani korištenjem različitih parametara prikazani su na slici 3.6. Iz generiranih primjera se može primjetiti da podaci postaju sve više grupirani kako se vrijednost parametra a smanjuje. Kada je vrijednost parametra a velika, pojedini podaci postaju vrlo slični makro kompoziciji. Ovo je ponašanje izravna posljedica korištenja simetrične Dirichletove razdiobe kao razdiobu nad parametrom π. Nadalje, može se dokazati da u graničnom slučaju kada $a \to 0$, model djelomične pripadnosti je ekvivalentan standardnom modelu mješavine konačnog broja gustoća.

3.5. Procjena parametara modela djelomične pripadnosti

Označit ćemo skup podataka $\mathcal{D}$ kao matricu $\mathbf{Y}$ dimenzija $N \times D$. Neka je $\mathbf{\Omega} = \{\Pi, \theta, \rho, a\}$ skup svih nepoznatih parametara a $\Phi = \{\mu_0, \sigma_0, \eta_0, b\}$ skup svih hiperparametara. U tom je slučaju cilj izračunati $p(\mathbf{\Omega}|\mathbf{Y}, \mathbf{\Phi})$, za što su autori originalnog rada primijenili MCMC-algoritam.

Općenito, MCMC-metode najprije definiraju Markovljev lanac čija je stacionarna razdioba jednaka razdiobi koja se želi izračunati. Iteriranjem, lanac konvergira prema svojoj stacionarnoj razdiobi, a svako je stanje lanca jedan uzorak stacionarne razdiobe.



Slika 3.6: Primjeri sintetičkih skupova generirani modelom djelomične pripadnosti (Heller et al., 2008)

Iako je Gibbsovo uzorkovanje (engl. *Gibbs sampling*) (Casella i George, 1992) zbog njegove jednostavnosti najpopularniji MCMC-algoritam, ova metoda ne može iskoristiti činjenicu da su sve nepoznate varijable modela djelomične pripadnosti zapravo kontinuirane. Nadalje, moguće je odrediti derivaciju logaritma zajedničke vjerojatnosne funkcije svih nepoznatih parametara, pa su zbog toga autori odlučili primijeniti Hibridno Monte Carlo uzorkovanje (engl. *Hybrid Monte Carlo*, *Hamiltonian Monte Carlo*) (MacKay, 2003) za procjenu svih nepoznatih parametara.

Hibridno Monte Carlo uzorkovanje je algoritam koji koristi informaciju o derivaciji logaritma uzajamne vjerojatnosti nepoznatih parametara kako bi lakše pronašao stanja s većim vjerojatnostima. Kada se radi s podacima visoke dimenzionalnosti, informacija o derivaciji često dovodi do znatno brže konvergencije, slično kao što je gradijentni spust gotovo uvijek brži od pretraživanja putem slučajnog odabira. Drugim riječima, algoritam hibridnog MC uzorkovanja simulira dinamiku sustava u kontinuiranom području Ω s energetskom funkcijom $\mathcal{E} = -\log p(\mathbf{X}, \Omega|\Psi)$, Derivacija energetske funkcije daje informaciju o područjima visoke vjerojatnosti te vuče nova stanja prema tim područjima. Detaljan opis algoritma može se naći u (MacKay, 2003)

4. Latentna Dirichletova alokacija

Ovo poglavlje sadrži opis modela Latentne Dirichletove alokacije, njegove strukture, pretpostavki te njegovo učenje. Nakon opisa samog modela (3.1) i opisa algoritma učenja parametara (3.2) slijedi usporedba ovog modela s modelom djelomične pripadnosti.

4.1. Model

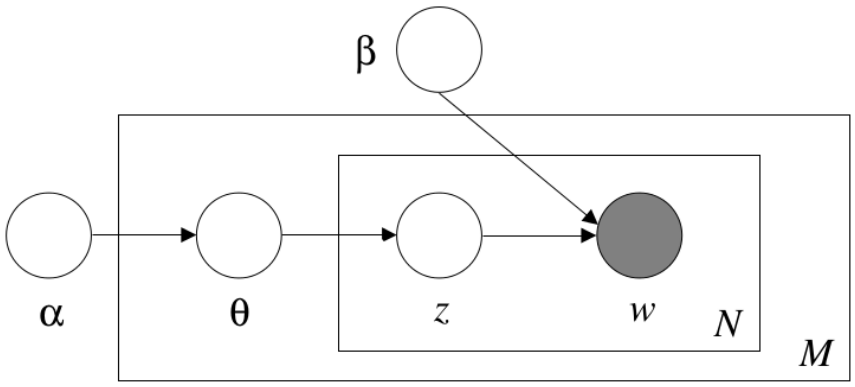
Latentna Dirichletova alokacija (engl. *Latent Dirichlet Allocation*) (Blei et al., 2003) je generativni probabilistički model koji svaki dokument (podatkovni primjer) **w** u zbirci podataka $\mathcal{D}$ modelira kao mješavinu tema z_k, gdje svaka tema predstavlja raspodjelu riječi vokabulara. Obzirom da je ovaj model razvijen najprije za modeliranje tekstnih podataka, često se umjesto termina primjer koristi termin *dokument*, a umjesto termina *atribut* koristi se termin *riječ*.

LDA pokušava modelirati zbirku podataka na način da:

- svakom podatku (dokumentu) u zbirci podataka pridruži visoku vjerojatnost
- dokumentima koji nisu u zbirci, a slični su dokumentima koji jesu unutar zbirke, također pridruži visoku vjerojatnost

Model predpostavlja da je svaki primjer, odnosno dokument, generiran na sljedeći način:

1. Odabere se duljina dokumenta d_i kao uzorak Poissonove razdiobe $(N\ Poisson(\zeta))$
2. Odabere se raspodjela tema u dokumentu $\grave{}_{\mathbf{i}}\ Dir(\alpha)$ kao uzorak Dirichletove razdiobe
3. za svaku od N riječi $w_{j,i}$ u dokumentu d_i:
   (a) Odabere se tema $z_{i,j}\ Multinomial(\theta_i)$ iz koje $w_{i,j}$ dolazi.



Slika 4.1: Bayesova mreža LDA modela (Blei et al., 2003)

(b) Odabere se $w_{i,j}$ kao uzorak iz z_i, pri čemu se $p(w_i|z_i, \beta)$ modelira kao multinomijalna razdioba gdje β sadrži parametre razdiobe.

Slično kao i model mješavine konačnog broja gustoća, standardni model LDA pretpostavlja konačan broj tema k, što znači da je dimenzionalnost Dirichletove razdiobe unaprijed određena. Teme se modeliraju kao matrica β dimenzija $K \times |V|$, gdje svaki redak predstavlja jednu razdiobu nad riječima vokabulara V. Formalno, $\beta_{ij} = p(w^j = 1|z^i = 1)$. Bayesova mreža modela LDA prikazana je na slici 4.1

Zajednička vjerojatnost mješavine tema, skupa K tema te dokumenta d duljine N dana je sljedećom formulom:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{i=1}^{N} p(z_i|\theta) p(w_n|z_n, \beta) \tag{4.1}$$

gdje je α hiperparametar Dirichletove razdiobe te govori do koje će mjere teme biti raspršene odnosno koncentrirane. Velik α dovodi do slične teme, dok mala vrijednost α-e rezultira vrlo različitim temama. Razdioba $p(z_n|\grave{})$ označava vjerojatnost teme z_n, dok $p(w_n|z_n, \beta)$ označava vjerojatnost riječi w_n unutar teme z_n. Marginalizacijom slučajnih varijabli dobiva se sljedeća formula za vjerojatnost pojedinog dokumenta d:

$$p(d|\alpha, \beta) = \int_{\theta} (\prod_{i=1}^{N} \sum_{z_n} p(z_n|\theta) \cdot p(w_{i,j}|z_n, \beta)) d\theta \tag{4.2}$$

dok je vjerojatnost korpusa veličine M dokumenata definirana kao:

$$p(\mathcal{D}|\alpha,\beta) = \prod_{i=1}^{M} p(d_i|\alpha,\beta) \tag{4.3}$$

Na slici 4.1 prikazana Bayesova mreža modela LDA. Može vidjeti da su sve razdiobe osim pojedinih riječi označene praznim kružićima. To znači da su samo riječi dokumenata poznati, dok su sve ostale varijable latentne. Nadalje, LDA model se sastoji od tri razine. Na prvoj se razini nalaze hiperparametri koji su unaprijed određene konstante. Druga razina sadrži parametre $\theta_{1,...k}$ koje se generiraju za svaki dokument d_i unutar zbirke dokumenata. Na kraju, na trećoj se razini nalaze varijable $z_{i,j}$ i $w_{i,j}$ koje se generiraju za svaku riječ unutar svakog dokumenta d_i. Modeli ovakvog tipa nazivaju se Hierarhijski Bayesovi modeli.

4.2. Procjena parametara LDA modela

Za procjenu parametara, autori modela su koristili algoritam varijacijskog zaključivanja (engl. *Variational Inference*) (Wainwright i Jordan, 2008). Međutim, implementacija korištena u ovom radu koristi Gibbsovo uzorkovanje za učenje modela, pa ćemo zbog toga opisati ovu metodu procjene parametara.

Općenito, u slučajevima kada je izračun

Gibbsovo uzorkovanje spada u porodicu Markovljev lanac Monte Carlo (Markov chain Monte Carlo) metoda te je, zbog njegove jednostavnosti, najčešće korištena metoda iz ove porodice. Ideja metode jest integriranje podataka u proces uzorkovanja na način da se za svaki podatkovni primjer definira nova varijabla čija se vrijednost fiksira. Razdioba ostalih varijabli je u tom slučaju podacima uvjetovana posteriorna razdioba koja se treba odrediti. Kod Gibbsovog uzorkovanja se u svakom koraku uzorkuje vrijednost jedne varijable uvjetovana ostalim varijablama u modelu. Na taj način, nakon velikog broja koraka, Markovljev lanac konvergira prema traženoj razdiobi.

Podaci se u našem slučaju sastoje od riječi $\mathbf{w} = \{w_1, ...w_{|V|}\}$, pri čemu je svaka riječ w_i pridružena jednom ili više dokumenata d_j. Vjerojatnost pojavljivanja svake riječi w_i unutar dokumenta d_j se može zapisati na sljedeći način:

$$p(w_i) = \sum_{j=1}^{N} p(z_j = j)p(w_i|z_i = j)$$

Za primjenu algoritma Gibbsovog uzorkovanja, potrebno je odrediti uvjetne razdiobe

svih varijabli čije su vrijednosti nepoznate. Uvjetna razdioba varijable z_i dana je sljedećom formulom:

$$p(z_i|\mathbf{z_{-i}}, \mathbf{w}) \propto p(w_i|z_i = j, \mathbf{z_{-i}}, \mathbf{w}_{-i})p(z_i|\mathbf{z_{-i}}) \tag{4.4}$$

pri čemu $\mathbf{z_{-i}}$ sadrži vrijednosti svih z varijabli osim z_i, dok $\mathbf{w_{-i}}$ sadrži vrijednosti svih w varijabli, osim varijable w_i. Parametar ` (raspodjela tema po dokumentima) se ne pojavljuje u gornjem izrazu jer je moguće odrediti uvjetne razdiobe latentnih varijabli koje ovise samo o $\mathbf{w_i}$ i $\mathbf{z_i}$. Prvu razdiobu možemo zapisati kao:

$$p(w_i|z_i = j, \mathbf{z_{-i}}, \mathbf{w}_{-i}) = \int p(w_i|z_{-1} = j, \phi^{(j)})p(\phi^{(j)}|\mathbf{z_{-i}w_{-i}})d\phi^{(j)} \tag{4.5}$$

gdje parametar $\phi^{(j)}$ sadrži raspodjelu riječi j-te teme, a njegova je uvjetna razdioba dana formulom:

$$p(\phi^{(j)}|\mathbf{z_{-i}}, \mathbf{w}_{-i}) \propto p(w_{-i}|\phi^{(j)}, z_{-i})p(\phi^{(j)}) \tag{4.6}$$

Koristeći do sad definirane relacije te uzevši u obzir sve razdiobe LDA modela, može se doći do sljedećeg izraza za uvjetne razdiobe latentnih varijabli:

$$p(z_i = j|z_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{w_i} + \beta}{n_{-i,j}^{(\cdot)} + |V|} \frac{n^{d_i} + \alpha}{n_{-i}^{di} + K\alpha} \tag{4.7}$$

gdje je $n_{-i,j}^{(\cdot)}$ broj riječi, osim trenutne, pridružene temi j, dok $n_{-i,j}^{(w_i)}$ označava koliko je puta riječ w_i pridružena temi j, izuzev trenutne riječi. Parametar n^{d_i} predstavlja broj riječi dokumenta d_i, dok je $n_{-i}^{d_i}$ broj riječi dokumenta d_i, isključujući trenutnu riječ.

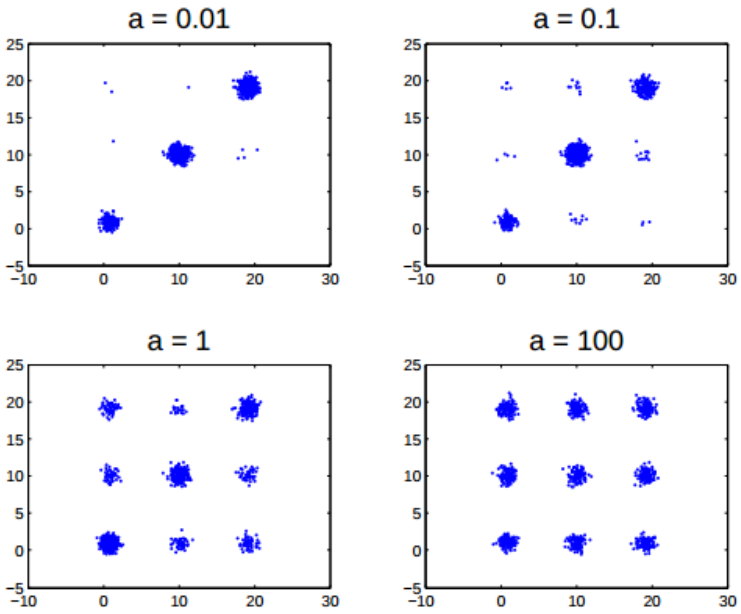
4.3. Usporedba modela latentne Dirichletove alokacije i modela djelomične pripadnosti

Iako su model latentne Dirichletove alokacije i model djelomične pripadnosti prilično slični u svojoj strukturi, postoje neke ključne razlike koje je potrebno spomenuti.

Oba modela sadrže Dirichletovi varijablu θ odnosno π koja definira makro zastupljenost tema po dokumentima. Međutim, kod LDA modela, ova se varijabla interpretira kao vjerojatnosna razdioba nad temama, dok kod modela djelomične pripadnosti, ova varijabla definira težinske faktore s kojima svaki izvor sudjeluje u generiranje podataka. Nadalje, svaka se riječ (atribut) u LDA modelu generira zasebno, dok se kod modela djelomične pripadnosti dokumenti (primjeri) uzorkuju izravno iz svakog izvora. Drugim riječima, prilikom generiranja podataka, LDA model pretpostavlja da je svaki atribut uzorkovan zasebno i neovisno o ostalim atributima, dok model djelomične pripadnosti izravno generira vektore atributa.

Na slici 4.2 su prikazani primjeri sintetičkih skupova generiranih LDA modelom koristeći tri izvora s Gaussovom umjesto multinomijalnom razdiobom. Kod modela LDA, svaki je atribut (riječ) pojedinog primjera (dokumenta) generiran jednim izvorom (temom). Obzirom da je se raspodjela riječi po temama uzorkuje iz Dirichletove razdiobe, mala vrijednost α-e znači da će atributi najčešće dolaziti iz malog broja izvora (raspodjela po temama će biti rijetka (engl. *sparse*), dok velika vrijednost α-e daje model kod kojeg su riječi približno uniformno raspodjeleni po temama. Iz rezultata dobivenih za $\alpha = 0.01$ se može zaključiti da su središta izvora u točkama $(0, 0)$, $(10, 10)$ i $(20, 20)$. Kako se vrijednost parametra α povećava, tako točke postaju sve raspršenije jer se svaka od njih počne generirati pomoću više izvora podataka.

Iz dobivenih rezultata se može vidjeti da je LDA model na neki način jednostavniji jer nije u mogućnosti generirati sve moguće točke između odabranih izvora podataka. Međutim, ova činjenica ne ukazuje na to kako je ovo lošiji model, već govori da LDA ima čvršće definiranu strukturu te da je pogodniji za neke vrste problema.



Slika 4.2: Skupovi podataka generirani LDA modelom (Blei et al., 2003)

5. Metode za ekstrakciju ključnih riječi

Ovo poglavlje sadrži opis razvijenih metoda za izvlačenje ključnih riječi te je objašnjeno zašto smo se odlučili koristiti jedino LDA model za rješavanje ovog problema.

5.1. Ekstrakcija ključnih riječi pomoću modela djelomične pripadnosti

Neka je zadan skup podataka $\mathcal{D}$ u obliku matrice dimenzija $N \times D$ gdje svaki redak predstavlja jedan podatkovni primjer (dokument). Grupiranje podataka $\mathcal{D}$ modelom djelomične pripadnosti nastaje procjenom matrice $\Pi = \{\vec{\pi_1}, ..., \vec{\pi_n}\}$ kod koje je i-i redak indikatorska varijabla i-og primjera (varijabla koja određuje pripadnost i-og primjera pojedinim grupama), a j-ti stupac određuje pripadnost pojedinog primjera j-oj grupi.

Pretpostavika je da ukoliko transponiramo matricu $\mathcal{D}$ i umjesto dokumenata grupiramo same riječi, one riječi koje se pojavljuju u istim dokumentima završit će u istim grupama. Time bismo dobili grupe u kojima će riječi biti na neki način (semantički ili tematski) povezani. Prateći (Saratlija et al., 2011), računanjem preklapanja grupa s riječima dokumenata bismo pokušali odrediti ključne riječi svakog dokumenta.

Kako bismo odredili prvih k riječi kod kojih je utjecaj j-e grupe najveći, trebamo u j-om stupcu matrice Π pronaći indekse redaka k najvećih vrijednosti. Time bismo dobili male grupe usko povezanih riječi koje kasnije možemo koristiti za izvlačenje ključnih riječi iz dokumenata.

Međutim, u praksi se ponašanje modela pokazalo drugačije od inicijalne pretpostavke. U nastavku su prikazane riječi s najvećim pripadnostima prvih 3 grupa nakon primjene modela na skup od 60 dokumenata. Pored svake riječi nalazi se odgovarajuća vrijednost pripadnosti $\pi_i^{(j)}$.

Grupa 1	Grupa 2	Grupa 3
priopćenje:0.0932	cilj:0.0875	niz:0.0876
sporazum:0.0887	njemački:0.0758	desni:0.0839
nadležan:0.0871	iva:0.0723	vrijeme:0.0837
sport:0.0840	izdanje:0.0722	zaposlenik:0.0818
droga:0.0835	glas:0.0712	sjeverni:0.0814

Iz samih se grupa može vidjeti da između najizglednijih riječi jedne grupe ne postoji silna semantička veza. Nadalje, primjenom metoda razvijenih za model LDA (podpoglavlja 5.2.1 i 5.2.2) postignuta je preciznost manja od 10% nakon čega je nastavljeno s istraživanjem mogućnosti latentne Dirichletove alokacije.

5.2. Ekstrakcija ključnih riječi pomoću modela LDA

LDA model je moguće primijeniti na veliki dio problema, a u ovom su poglavlju opisane dvije metode za ekstrakciju ključnih riječi iz dokumenata. Prva metoda pretpostavlja da su ključne riječi pojedinog dokumenta sadržane u jednoj temi, dok druga tema pokušava pronaći ključne riječi dokumenta u više tema modela LDA.

5.2.1. Ekstrakcija ključnih riječi iz jedne teme

Prva je metoda razvijena za obradu malih skupova podataka (skupova s manje od 100 dokumenata) te se zasniva na pretpostavci da jedna tema sadrži sve ključne riječi jednog dokumenta.

Motivacija iza ovog pristupa proizlazi iz intuicije da ukoliko je broj tema približno jednak broju dokumenata, LDA modelom dobili bismo takve teme u kojima će najvjerojatnije riječi jedne teme biti zastupljene u malom broju dokumenata. Drugim riječima, očekivali bismo da će najvjerojatnije riječi iz određene teme proizlaziti iz jednog jedinog dokumenta, a u idealnom slučaju bismo dobili jedan na jedan preslikavanje između LDA tema i dokumenata u skupu podataka. Iz pretpostavke da su najvjerojatnije riječi teme ujedno i karakteristične riječi pripadnih dokumenata slijedi da možemo ključne riječi određenog dokumenta izvući iz najvjerojatnijih riječi pripadne teme. S tim pretpostavkama na umu, problem se svodi na određivanje teme koja najbolje opisuje pojedini dokument, izvlačenje ključnih riječi dokumenta te proširivanje ključnih riječi u izraze koje tvore zatvorenu semantičku cjelinu.

Pridruživanje tema dokumentima

Za određivanje jedinstvene teme dokumenta mogu se iskoristiti parametri LDA modela. Nakon što je završen proces učenja modela, svakom dokumentu d_i je pridružen stohastički vektor θ_i koji definira raspodjelu riječi po dokumentima. Ovaj parametar još možemo interpretirali kao vjerojatnosti da dokument d_i pripada pojedinoj temi ϕ_k. Također, svaka tema ϕ_k je distribucija nad riječima vokabulara, odnosno predstavlja skup parova (*riječ*, p), $p \in (0, 1)$. Vektor θ_i je parametar koji određuje udio pojedinih tema u dokumentu d_i. Primjerice, ukoliko se radi o modelu s dvije teme, te ukoliko vektor θ_i ima vrijednost $(0.7, 0.3)$, možemo pretpostaviti da je 70% riječi dokumenta d_i generirano prema prvoj temi (prema distribuciji ϕ_1), te da je 30% riječi generirano prema distribuciji druge teme.

Nadalje, prilikom određivanja jedinstvene teme dokumenta želimo uzeti u obzir samo karakteristične riječi tog dokumenta (riječi s velikom vjerojatnošću u nekoj temi). Pridruživanje dokumentu d_i najzastupljeniju temu ϕ_k bi u ovom bi slučaju bilo nedovoljno dobro rješenje, jer nemamo garanciju da se najvjerojatnije riječi teme ϕ_k pojavljuju u dokumentu d_i.

Stoga smo definirali heurističku funkciju koja vrednuje pripadnost dokumenta d_i temi ϕ_k na sljedeći način:

$$S(d_i|\phi_k) = \sum_{j=1}^{len(d_i)} \{1|I(w_{i,j}, \phi_k, n) \wedge \theta_{i,j} > T_1 \wedge \theta_{i,j} \cdot \phi_{k,ij} > T_2\} \tag{5.1}$$

gdje hiperparametar $\theta_{i,k}$ označava udio teme ϕ_k u dokumentu d_i, dok hiperparametar $\phi_{k,ij}$ označava vjerojatnost generiranja j-te riječ i-tog dokumenta temom ϕ_k. Funkcija $len(x)$ vraća duljinu dokumenta x dok je $I(w_{ij}, \phi_k, n)$ funkcija koja vraća 1 ako i samo ako je j-ta riječ i-tog dokumenta sadržana u n najvjerojatnijih riječi teme ϕ_k. Izrazi $\theta_{i,j} > T_1$ i $\theta_{i,j} \cdot \phi_{k,ij} > T_2$ koriste se kako bi se iz sumacije isključili članovi čije su odgovarajuće vrijednosti manje od pragova T_1 i T_2. Bitno je napomenuti da T_1 i T_2

Izračunavanje pripadnosti dokumenta d_i temi ϕ_k može se opisati sljedećim pseudokodom:

1. Suma = 0;
2. Za svaku riječ $w_{i,j}$ u dokumentu d_i:
    2.1. Ako $P(w_{i,j}, \phi_k, n) = 1 \wedge \phi_{k,ij} > T1 \wedge \phi_{k,ij} \cdot \theta_{i,k} > T2$:
        2.1.1. Suma += $\phi_{k,ij} \cdot \theta_{i,k}$

3. Vrati Suma;

Korištenjem samo prvih (najvjerojatnijih) n riječi dobivenih tema (korištenjem *reduciranih tema*) izbacuje se utjecaj onih riječi koje nemaju veliku vjerojatnost pojavljivanja u nekoj od naučenih tema. Prema našoj intuiciji su takve riječi nekarakteristične za bilo koji od dokumenata, te smanjuju uspješnost sustava prilikom ekstrakcije ključnih riječi. Nadalje, kako bi se smanjio utjecaj riječi koje imaju veliku vjerojatnost u više tema, nakon fiksiranja hiperparametra n se također izbacuju riječi koje imaju rang $< n$ (najvjerojatnija riječ ima rang 1) u više od jedne teme.

Cijeli postupak može se opisati sljedećim pseudokodom:

1. Pripremiti skup podataka
2. Primijeniti LDA nad skupom podataka
3. Izabrati vrijednost hiperparametra n
4. Iz svake teme izbaciti riječi koje imaju rang $> n$, i tako dobiti reducirane teme.
5. Iz svake teme izbaciti riječi koje se pojavljuju u više tema.
6. Za svaki dokument d_i:
   6.1. Za svaku temu ϕ_k izračunati vrijednost funkcije $S(d_i|\phi_k)$
7. dokumentu d_i pridružiti temu ϕ_k s najvećom vrijednošću $S(d_i|\phi_k)$.

Prilikom provođenja postupka, velika je vjerojatnost da će više dokumenata dobiti istu temu, što znači da se iz različitih dokumenata mogu izvući iste ključne riječi.

Ekstrakcija ključnih riječi

Ekstrakcija ključne riječi provodi se nakon određivanja (reduciranih) tema dokumenata. Za svaki dokument d_i, kao ključne riječi odabiru se one riječi koje se ujedno nalaze i u njegovoj reduciranoj temi. Odnosno, ključne riječi jednog dokumenta tvore presjek tog dokumenta s njegovom reduciranom temom. Stoga čak i oni dokumenti kojima je pridružena ista reducirana tema neće nužno imati identičan skup ključnih riječi.

Generiranje ključnih izraza

Nakon što su ključne riječi izvučene iz dokumenata, slijedi krajnji postupak generiranja ključnih izraza. Ključne se riječi proširuju u ključne izraze prema postupku opisanom u (Saratlija et al., 2011). Najprije se u dokumentu označi prvo pojavljivanje svake ključne riječi, jer se pretpostavlja da se tada koncepti opisuju u punoj formi. Nakon toga se ključne riječi proširuju u izraze prema sljedećem algoritmu:

1. Za svaku označenu riječ koja je imenica ili pridjev,
   označiti susjedne imenice i pridjeve ukoliko su u istom padežu, rodu i broju.
2. Označiti svaki neoznačeni prijedlog ukoliko se nalazi između dvije označene riječi, te ukoliko padež druge riječi gramatički odgovara prijedlogu.
3. Ponavljati sve dok postoje promjene u označavanju.
4. Sve označene, neprekidne sekvence predstavljaju ključne izraze.

Primjerice, neka je zadana rečenica:

predsjednik republike i premijer prikrivaju pravu istinu te zataškavaju kršenje zakona

u kojoj su sve ključne riječi potcrtane.

U prvoj će se iteraciji algoritma označiti riječi *pravu*, *republike*, i *kršenje*, te će nakon toga algoritam ekspanzije stati. Dobiveni ključni izrazi će biti: *predsjednik republike*, *pravu istinu*, *kršenje zakona*.

5.2.2. Ekstrakcija ključnih riječi iz više tema

Druga metoda obrađena u ovom radu također koristi reducirane teme LDA modela opisane u poglavlju 5.2.1, ali je vođena pretpostavkom da ključne riječi mogu dolaziti iz više reduciranih tema. Stoga, ova metoda ne radi pridruživanje jedinstvenih tema dokumentima, već svaki dokument modelira kao njihovu mješavinu. Postupak određivanja reduciranih tema identičan je onom u prethodnoj metodi, dok se izvlačenje ključnih riječi dokumenta d_i radi na način da se kao ključne riječi označavaju one riječi koje imaju rang $< n$ u barem jednoj od naučenih tema. Drugim riječima, riječ d_{ij} će biti proglašena ključnom ukoliko ima rang $< n$ u jednoj od tema modela, odnosno ukoliko se pojavljuje u jednoj od reduciranih tema.

Postupak izvlačenja ključnih riječi iz dokumenta d_i može se opisati sljedećim pseudokodom:

1. Za svaku reduciranu temu ϕ'_k:
   1.1. Proglasi d_{ij} ključnom ukoliko se d_{ij} pojavljuje u reduciranoj temi ϕ'_k.
2. Vrati listu ključnih riječi

Nakon izvlačenja ključnih riječi provodi se postupak njihovog proširivanja koji je također identičan postupku korištenom u prethodnoj metodi.

6. Vrednovanje

U ovom su poglavlju opisani provedeni eksperimenti te rezultati empirijskog vrednovanja opisanih metoda.

6.0.3. Skup podataka

Za vrednovanje predloženih metoda korišteni su označeni novinski članci na Hrvatskom jeziku (Mijic et al., 2010). Cijeli skup podataka sadrži 1020 članaka hrvatske novinske agencija HINA a označavanje je radilo osam označivača. Označivači su označavali na način da su najprije definirali skup pravila označavanja, te su nakon toga samostalno odredili skup ključnih riječi svog dijela skupa članaka. Jedan je podskup od 60 članaka dodijeljen svakom od osam označivača kako bi se moglo odrediti međusobno slaganje. Ostalih 960 dokumenata je podijeljeno u 8 isključivih podskupova od 120 članaka, te je svaki skup bio dodijeljen točno jednom označivaču. U ovom je radu u svrhu evaluacije bio korišten zajednički podskup od 60 članaka, dok je preostali dio podataka iskorišten za procjenu parametara modela.

6.0.4. Način vrednovanja

Kako bi se detaljnije ispitalo ponašanje sustava, proveli smo eksperimente u tri različita okruženja.

U okruženju *60x2* za učenje parametara modela i za evaluaciju postupaka korišten je isti evaluacijski skup od 60 dokumenata. Unatoč tome što je postupak ovog tipa neprikladan za realne primjene, htjeli smo procijeniti gornju granicu uspješnosti sustava. U okruženju *960-60* je za procjenu LDA parametara korišten skup od 960 dokumenata, dok je evaluacija provedena nad evaluacijskim skupom od 60 dokumenata. Na kraju smo htjeli vidjeti hoće li povećanje skupa podataka rezultirati poboljšanjem u odnosu na okruženje *60x2*, pa smo za učenje koristili cijeli skup od 1020 dokumenata, dok smo sustav evaluirali nad istim skupom kao i u prethodnim okruženjima.

Čest problem koji se javlja kod evaluacije sustava za izvlačenje ključnije riječi jest sama subjektivnost zadatka. Stoga, podudaranje izlaza sustava s oznakama označivača često uključuje razne aproksimacije. U našem slučaju, definirali smo da su dva izraza jednaka ukoliko je ispunjen barem jedan od navedenih uvjeta:

1. Izrazi su identični
2. Predviđen izraz sadrži izraz označivača
3. Preklapanje izraza veće je od 50% prosječne duljine tih dvaju izraza

Ovaj je mjera jednakosti također korištena u slučaju kada se označivači nisu međusobno slagali oko ključnih izraza određenog dokumenta. Ključan se izraz I izbacuje iz skupa oznaka za dokument d_i ukoliko je manje od 50% označivača odabralo I kao ključni izraz dokumenta d_i.

6.0.5. Mjere evaluacije

Prilikom evaluacije korištene su dvije mjere, F_1 mjera (Sokolova et al., 2006) te Kendall Tau udaljenost (Noether, 1981).

F_1 je mjera koja je kombinacija *preciznosti* i *odziva* sustava. Preciznost se u ovom kontekstu definira kao udio točnih izraza sustava u odnosu na sve izraze sustava. Odziv je pak definiran kao udio točnih ključnih izraza sustava u odnosu na broj izraza koji je sustav trebao vratiti. Drugim riječima, preciznost je dana formulom:

$$P = \frac{|\{\text{točni izrazi}\} \cap \{\text{izrazi sustava}\}|}{|\{\text{izrazi sustava}\}|} \tag{6.1}$$

dok je odziv jednak:

$$R = \frac{|\{\text{točni izrazi}\} \cap \{\text{izrazi sustava}\}|}{|\{\text{točni izrazi}\}|} \tag{6.2}$$

U tom je slučaju, F_1-mjera harmonička sredina preciznosti i odziva:

$$R = \frac{2}{\frac{1}{P} + \frac{1}{R}} = 2\frac{PR}{P+R} \tag{6.3}$$

U ovom je radu korištena asimetrična F_1 mjera. Prilikom mjerenja preciznosti kao zlatni standard je uzeta unija ključnih izraza svih označivača, jer smo htjeli da se ključni izraz ispravnim ukoliko ga je barem jedan označivač označio. S druge strane za izračun odziva kao zlatni standard uzet je podskup ključnih izraza označenih od više od 5 označivača, jer smo htjeli uzeti u obzir samo one izraze koje je označila većina označivača.

Jedna od negativnih strana F mjere u evaluaciji sustava ovog tipa je loša skalabilnost. Drugim riječima, što je broj označivača veći to je i broj ključnih izraza veći, pa se uspješnost sustava prema F mjeri automatski povećava relaksacijom uvjeta ključnosti izraza. Stoga smo uzeli u obzir još jedan tip mjere. Kendall Tau, za razliku od F mjere, zahtijeva da je izlaz sustava rangirana lista ključnih izraza. Zlatni je skup rangiran na način da se rang ključnog izraz za određeni dokument računa kao broj označivača koji ga je označio u danom dokumentu. S druge strane, rang ključnih izraza izlaza sustava je u korelaciji s rangom ključne riječi u svojoj temi temi.

Jednom kada su sortirane liste formirane, Kendall Tau je mjera definirana kao broj zamjena koje je potrebno napraviti kako bi se poredak jedne liste izjednačio s poretkom druge liste.

6.1. Predobrada podataka

U svim ispitnim okruženjima podaci su obrađeni na isti način. Svaki je dokument najprije obrađen sustavom za označavanje vrste riječi. Kao označivač vrste riječi korišten je označivač iz sustava za strojno prevođenje Apertium (Forcada et al., 2011). Nakon što je riječima dodijeljena oznaka vrste, iz svakog su dokumenta izbačene riječi koje nemaju oznaku imenice ili pridjeva. Time smo prije svega smanjili dimenzionalnost i broj parametara za procjenu, te smo ujedno drastično reducirali broj kandidata koji bi mogli predstavljati ključnu riječ dokumenta. Očekivali smo da će ključne riječi uglavnom biti imenice, a pridjeve smo zadržali zbog povećanja konteksta.

6.2. Rezultati

U ovom su dijelu prikazani dobiveni rezultati te je njihova analiza.

6.2.1. Okruženje *60x2*

Kao što je i prije napomenuto, u okruženju *60x2* za učenje i za evaluaciju korišten je samo skup od 60 dokumenata. Unatoč veličine korištenog skupa, od eksperimenta ovakvog tipa očekivali smo da će pronaći gornju granicu uspješnosti razvijenog sustava.

Kao hiperparametri prve metode korišteni ukupan broj tema te parametri T_1 i T_2 opisani u 5.2.1, dok je kao maksimalni rang riječi odabran broj 20. Hiperparametari druge metode su broj tema te rang riječi.
Dobiveni rezultati prikazani su u tablicama 6.1 i 6.2

Tablica 6.1: Rezultati prve metode za okruženje *60x2*

broj tema	n	T_1	T_2	P	R	F1
10	20	0.0	0.6	**44.5**	19.7	27.3
10	20	0.0	1.4	44.5	**25.3**	25.3
10	20	0.0	0.0	42.6	23.9	**30.7**
20	20	0.0	1.8	**48.4**	23.7	31.8
20	20	0.0	0.0	38.9	**35.6**	37.1
20	20	0.0	0.0	38.9	35.6	**37.1**
30	20	0.0	1.1	**51.6**	36.6	42.8
30	20	0.0	0.0	43.1	**44.6**	44.7
30	20	0.0	0.0	43.1	46.4	**44.7**
50	20	0.0	1.3	**49.5**	36.6	42.8
50	20	0.0	0.0	41.0	**44.5**	42.7
50	20	0.0	0.0	41.0	44.5	**42.7**
100	20	0.0	0.7	**46.4**	26.3	33.6
100	20	0.0	0.0	38.5	**57.5**	46.0
100	20	0.0	0.0	38.4	57.5	**46.0**

Hiperparametar T_2 korišten kod prve metode kontrolira osjetljivost sustava prilikom određivanja riječi koje se smatraju ključnim za određeni dokument. Što je vrijednost hiperparametra veća, to će broj izvučenih ključnih riječi biti manji. Također smanjivanje vrijednosti T_2 će ujedno povećati broj označenih ključnih riječi. Ova veza nadalje utječe i na mjere preciznosti i odziva. Što je više riječi označeno ključnim, to će se odziv povećavati, jer će i vjerojatnost da se ključna riječ ujedno nalazi u izlazu

Tablica 6.2: Rezultati druge metode za okruženje *60x2*

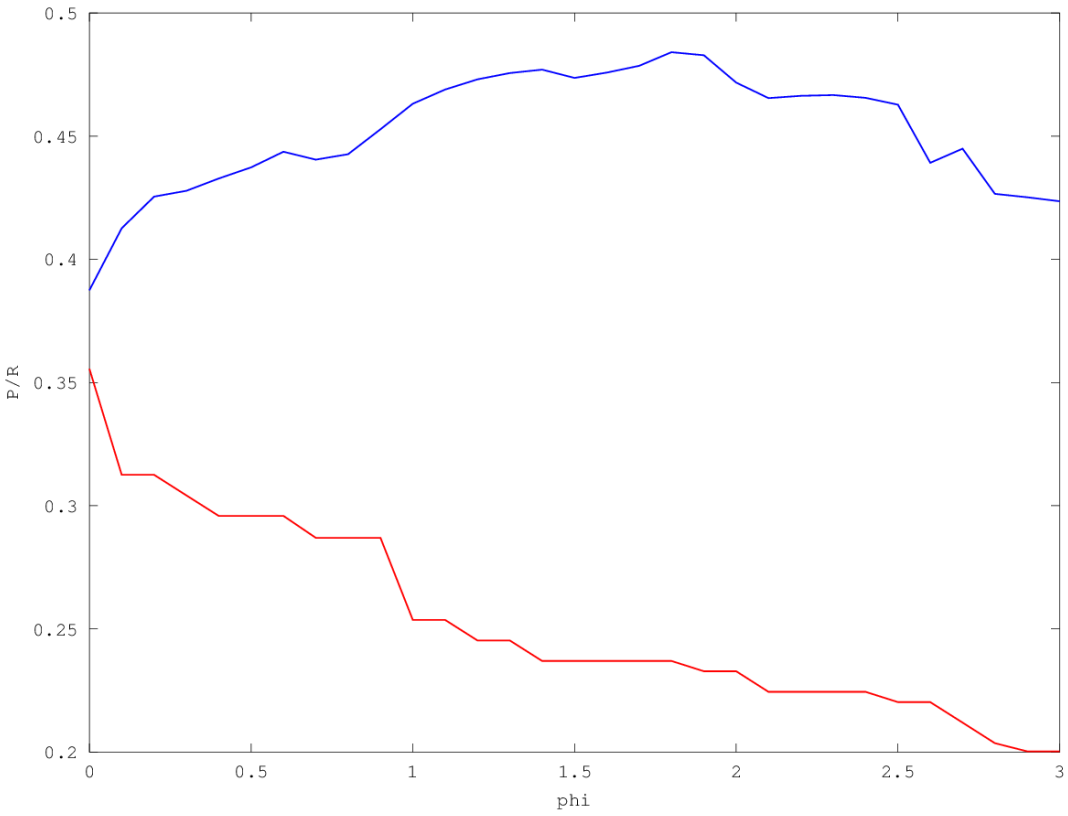
broj tema	n	P	R	F1
10	25	**40.8**	24.2	30.4
10	25	40.8	**24.2**	30.4
10	25	40.8	24.2	**30.4**
20	23	**46.8**	18.7	26.7
20	23	42.6	**33.3**	37.4
20	23	42.6	33.3	**37.4**
30	18	**48.2**	31.4	38.1
30	24	45.6	**35.0**	31.5
30	23	46.4	35.0	**40.0**
50	2	**47.6**	15.4	23.1
50	47	37.9	**53.1**	43.7
50	47	37.9	53.1	**43.7**
100	2	**44.3**	15.4	22.9
100	36	33.4	**42.1**	37.3
100	36	33.4	42.1	**37.3**

sustava biti veća. Jedini slučaj gdje se odziv sustava može smanjiti sa smanjivanjem vrijednosti T_1 jest ukoliko se dvije ili više ključne riječi spoje u jedan ključni izraz. S druge strane, povećanje broja izvučenih ključnih riječi može smanjiti, ali i povećati preciznost sustava, jer preciznost, za razliku od odziva, izravno ovisi o broju riječi u izlazu sustava. Pa tako će dodavanje nove riječi u izlaz sustava povećati preciznost ukoliko je riječ ispravno označena ključnom, ili smanjit će preciznost sustava ukoliko je dodana riječ neispravno označena ključnom. Utjecaj parametra T_2 za model s 30 tema prikazana je na slici 6.1. Crvena linija označava vrijednost odziva za različite vrijednosti T_2, dok plava linija označava promjenu preciznosti.

Na prikazanom grafu se vidi da se odziv sustava monotono smanjuje s povećanjem vrijednosti $\phi_{k,ij}$. Stoga, ukoliko želimo maksimalni odziv, postavili bi vrijednost hiperparametra na nulu. S druge strane, vidljivo je i da se preciznost mijenja nemonotono s povećanjem vrijednosti T_2, pa stoga vrijednost koja daje maksimalnu preciznost nije unaprijed poznata.

Na slici 6.2 prikazano je ponašanje preciznosti i odziva sustava za različite vrijednosti hiperparametra n korištenog kod druge metode. I kod ovog hiperparametra se

analitički može doći do sličnih zaključaka kao i kod hiperparametra $\phi_{k,ij}$. Povećanje vrijednosti minimalnog ranga n izravno dovodi do povećanja odziva sustava. Slično kao i kod prve metode, povećajnje vrijednosti n može dovesti do smanjivanja odziva sustava samo ukoliko povećanje ranga dovede do spajanja ključnih riječi u jedan izraz. Vrijednost koja daje maksimalnu preciznost opet ne može biti poznata unaprijed te se mora na neki način optimirati.

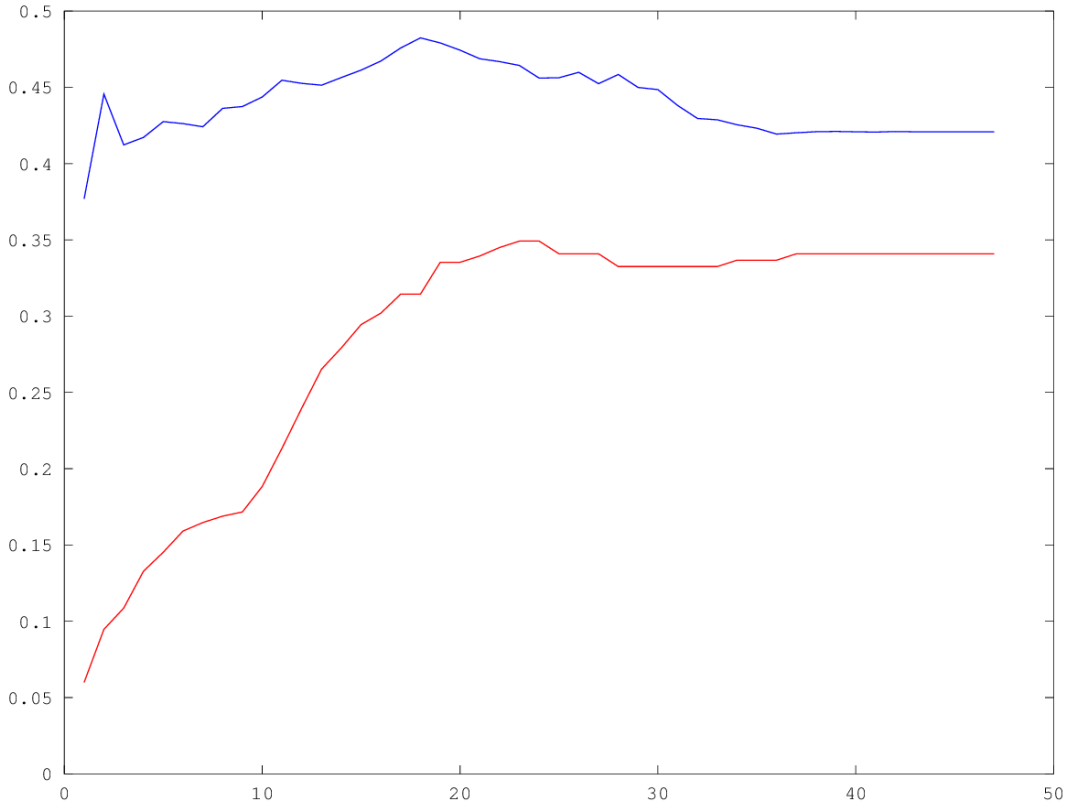


Slika 6.1: Preciznost i odziv prve metode za različite vrijednosti hiperparametra T_2

6.2.2. Okruženje *960-60*

U okruženju *960-60* za procjenu parametara korišten je skup od 960 neoznačenih dokumenata, dok je za testiranje korišten skup od 60 dokumenata. Obzirom da procjena parametara LDA modela traje znatno duže od same primjene modela, ovaj način izvlačenja ključnih riječi je puno praktičniji od načina opisanog u prethodnom dijelu.

Dobiveni rezultati prikazani su u tablicama 6.3 i 6.3. Rezultati pokazuju da prva metoda izvlačenja ključnih riječi u ovom okruženju daje usporedive rezultate s onima dobivenim u okruženju *60x2*. Nadalje, iako sama procjena parametara traje duže zbog



Slika 6.2: Preciznost i odziv druge metode za različite vrijednosti hiperparametra n

većeg broja dokumenata za učenje, proces izvlačenja ključnih riječi iz novig dokumenata jest znatno brži. Razlog zašto sustav radi podjednako dobro na neviđenom skupu jest činjenica da smo koristili veći skup za procjenu parametara.

Ukoliko rezultate za okruženje *60x2* gledamo kao uspješnost na skupu za učenje, možemo tvrditi da rezultati dobiveni u ovom okruženju ne indiciraju prenaučenost ili podnaučenost modela. Također, obzirom da smo na raspolaganju imali samo skup od 60 označenih dokumenata koji smo koristili isključivo za krajnu evaluaciju, nismo bili u mogućnosti provesti optimizaciju hiperparametara samog LDA modela kako bismo detaljnije istražili mogućnosti ovih metoda.

6.2.3. Okruženje *1020-60*

Na kraju, okruženje *1020-60* je slično okruženju *60x2*, s time da je skup za procjenu parametara znatno veći. Htjeli smo vidjeti hoće li povećanja skupa za učenje utjecati na uspješnost sustava ukoliko se ne koristi gotov model kao u prethodnom dijelu.

U tablicama 6.7, 6.8, 6.9 i 6.10 su prikazani dobiveni rezultati za obje metode na

Tablica 6.3: Rezultati prve metode za okruženje *960-60*

broj tema	n	T_1	T_2	P	R	F1
40	20	0.0	2.5	**42.7**	18.3	25.6
40	20	0.0	0.0	37.5	**45.9**	41.3
40	20	0.0	0.0	37.5	45.9	**41.3**
60	20	0.0	2.0	**45.9**	14.1	21.6
60	20	0.0	0.0	39.3	**34.5**	36.7
60	20	0.0	0.0	39.3	34.5	**36.7**
80	20	0.0	1.3	**43.8**	15.8	23.2
80	20	0.0	0.0	36.1	**27.0**	30.9
80	20	0.0	0.0	36.1	27.0	**30.9**
100	20	0.0	1.3	**44.6**	12.8	19.8
100	20	0.0	0.0	37.4	**26.5**	31.1
100	20	0.0	0.0	37.4	26.5	**31.1**

Tablica 6.4: Kendall tau prve metode za okruženje *960-60*

broj tema	n	T_1	T_2	Kendall tau
40	20	0.00	2.00	0.68
60	20	0.00	0.25	0.69
80	20	0.00	0.00	0.67
100	20	0.00	0.25	0.71

način da su za svaki broj tema prikazani parametri s kojima je postignut maksimum za jednu od korištenih mjera.

Tablica 6.5: Rezultati druge metode za okruženje *960-60*

broj tema	n	P	R	F1
40	47	5.5	0.8	1.4
40	9	3.3	0.7	1.3
40	47	5.5	0.8	1.4
60	47	3.7	0.8	1.3
60	9	3.3	0.7	1.3
60	47	3.7	0.8	1.3
80	47	13.3	1.2	2.2
80	9	11.1	1.7	2.2
80	47	13.3	1.2	2.2
100	9	3.3	0.8	1.3
100	9	3.3	0.8	1.3
100	9	3.3	0.8	1.3

Tablica 6.6: Kendall tau druge metode za okruženje *960-60*

broj tema	n	Kendall tau	
40	20	7	0.94
60	20	37	0.90
80	20	2	0.93
100	20	12	0.91

Tablica 6.7: Rezultati prve metode za okruženje *1020-60*

broj tema	n	T_1	T_2	P	R	F1
40	20	0.0	0.0	39.9	39.0	39.4
40	20	0.0	0.0	39.9	39.0	39.4
40	20	0.0	0.0	39.9	39.0	39.4
60	20	0.0	0.0	37.9	42.8	40.2
60	20	0.0	0.0	37.9	42.8	40.2
60	20	0.0	0.0	37.9	42.8	40.2
80	20	0.0	0.0	35.4	46.0	40.0
80	20	0.0	0.0	35.4	46.0	40.0
80	20	0.0	0.0	35.4	46.0	40.0
100	20	0.0	0.0	32.7	44.5	37.7
100	20	0.0	0.0	32.7	44.5	37.7
100	20	0.0	0.0	32.7	44.5	37.7

Tablica 6.8: Kendall tau prve metode za okruženje *1020-60*

broj tema	n	T_1	T_2	Kendall tau
40	20	0.00	0.00	0.71
60	20	0.00	0.00	0.72
80	20	0.00	0.00	0.72
100	20	0.00	0.25	0.64

Tablica 6.9: Rezultati druge metode za okruženje *1020-60*

broj tema	n	P	R	F1
30	10	**43.2**	17.9	25.3
30	31	41.6	**27.7**	33.3
30	31	41.6	27.7	**33.3**
60	19	**45.3**	24.3	31.7
60	47	43.3	**28.9**	34.7
60	42	43.3	28.9	**34.7**
80	11	**44.8**	17.1	24.8
80	47	38.5	**28.5**	32.8
80	37	38.7	28.5	**32.8**
100	7	**43.3**	14.0	21.1
100	34	33.1	**33.1**	33.1
100	30	33.4	33.1	**33.3**

Tablica 6.10: Kendall tau druge metode za okruženje *1020-60*

broj tema	n	Kendall tau
40	7	0.66
40	37	0.76
80	12	0.71
100	12	0.60

7. Zaključak

Sustavi za pretraživanje podataka vrlo često koriste ključne riječi kako bi utvrdili jesu li dva dokumenta na neki način povezana. Čitanje i sažimanje velikih dokumenata je još jedan zadatak koji je vremenski zahtjevan za ljude, pa se kao posljedica sve češće koriste računalne metode za rješavanje probleme ovog tipa. S popuplarizacijom interneta, dostupna količina podataka postala je veća nego ikada, a s povećanjem količine podataka povećava se i potreba za njezinim pretraživanjem i indeksiranjem. Ekstrakcija ključnih riječi jedan je način koji se može koristiti za kategorizaciju i brzo pretraživanje dokumenata.

U okviru ovog rada obrađen je model parcijalne pripadnosti za izvlačenje ključnih riječi iz dokumenata. Nakon što se ovaj pristup pokazao neuspješnim, nastavljeno je s primjenom modela Latentne Dirichletove alokacije, te su razvijene dvije metode oko ovog modela. Provedeno je vrednovanje razvijenih metoda te analiza dobivenih rezultata. Dobiveni su rezultati usporedivi s rezultatima ostalih metoda nenadziranog učenja.

Sljedeći korak bi uključivao detaljniju analizu otkrivenih tema kako bi se mogla razviti bolja mjera vrednovanja pripadnosti dokumenata pojedinim temama. To znači da bi se nakon primjene modela LDA teme dodatno obrađivale kako bi se broj neključnih riječi pokušao smanjiti na samom početku. Nakon ekspanzije ključnih riječi, dobiveni se izrazi također mogu vrednovati i rangirati kako bi neključni izrazi izbacili iz rezultata sustava. Konačno, potrebna je veća varijacija hiperparametara modela LDA kako bi se detaljnije utvrdile mogućnosti ovih metoda.

LITERATURA

Renee Ahel, B Dalbelo Bašic, i Jan Šnajder. Automatic keyphrase extraction from croatian newspaper articles. *The Future of Information Sciences, Digital Resources and Knowledge Sharing*, stranice 207–218, 2009.

David M Blei, Andrew Y Ng, i Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

George Casella i Edward I George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, i Francis M Tyers. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144, 2011.

Walter R Gilks, Sylvia Richardson, i David J Spiegelhalter. *Markov chain Monte Carlo in practice*, svezak 2. CRC press, 1996.

Kazi Saidul Hasan i Vincent Ng. Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art. U *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, stranice 365–373. Association for Computational Linguistics, 2010.

Katherine A Heller, Sinead Williamson, i Zoubin Ghahramani. Statistical models for partial membership. U *Proceedings of the 25th international conference on Machine learning*, stranice 392–399. ACM, 2008.

Finn V Jensen. *An introduction to Bayesian networks*, svezak 210. UCL press London, 1996.

David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

Geoffrey McLachlan i David Peel. *Finite mixture models*. Wiley. com, 2004.

Jure Mijic, B Dalbelo Bašic, i Jan Šnajder. Robust keyphrase extraction for a large-scale croatian news production system. *Proc. of FASSBL*, stranice 59–66, 2010.

Todd K Moon. The expectation-maximization algorithm. *Signal processing magazine, IEEE*, 13(6):47–60, 1996.

Gottfried E Noether. Why kendall tau. *Teaching Statistics*, 3(2):41–43, 1981.

Douglas A Reynolds. Gaussian mixture models., 2009.

Josip Saratlija, Jan Šnajder, i Bojana Dalbelo Bašić. Unsupervised topic-oriented keyphrase extraction and its application to croatian. U *Text, Speech and Dialogue*, stranice 340–347. Springer, 2011.

Marina Sokolova, Nathalie Japkowicz, i Stan Szpakowicz. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. U *AI 2006: Advances in Artificial Intelligence*, stranice 1015–1021. Springer, 2006.

Lonneke van der Plas, Vincenzo Pallotta, Martin Rajman, i Hatem Ghorbel. Automatic keyword extraction from spoken text. a comparison of two lexical resources: the edr and wordnet. *arXiv preprint cs/0410062*, 2004.

Martin J Wainwright i Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

Ian H Witten, Gordon W Paynter, Eibe Frank, Carl Gutwin, i Craig G Nevill-Manning. Kea: Practical automatic keyphrase extraction. U *Proceedings of the fourth ACM conference on Digital libraries*, stranice 254–255. ACM, 1999.

Primjena modela djelomične pripadnosti za ekstrakciju ključnih fraza iz dokumenata

Sažetak

Model parcijalne pripadnosti je poopćenje modela konačnog broja gustoća te je u mogućnosti modelirati djelomičnu pripadnost podataka pojedinim skupinama. Model pretpostavlja da je svaki podatak težinska suma uzoraka više izvora. S druge strane, latentna Dirichletova alokacija (LDA) modelira podatke kao diskretnu mješavinu jer je pretpostavka modela da je svaki atribut određenog podatka generiran neovisno o izvora ostalih atributa. U ovom je radu dana teorijska podloga modela parcijalne pripadnosti te modela latentne Dirichletove alokacije, opisana je njihova struktura, pretpostavke i način procjene njihovih parametara. Razvijene su dvije metode koje koriste parametre modela LDA te je vrednovana njihova uspješnost i ponašanje u različitim okruženjima.

Ključne riječi: obrada prirodnog jezika, PMM, LDA, ključne fraze, hrvatski jezik

Application of Partial Pembership Podels to Keyphrase Extraction from Croatian Documents

Abstract

The partial membership model (PMM) is a generalization of the standard finite mixture model since it can partial membership of each data point to different data sets. Every point from the data set is modeled as a weighted sum of samples from different sources. On the other hand, Latent Dirichlet Allocation (LDA) is a model which represents data points as discrete mixtures where each attribute of a given data point is generated independently of the sources of the other attributes. We have provided a theoretical background of the partial membership model and Latent Dirichlet Allocation and have developed two methods for keyword extraction using the parameters obtained by LDA. Finally, we have evaluated the performance of our methods, as well as their behaviour in different settings.

Keywords: natural language processing, PMM, LDA, keyphrase extraction, Croatian language