Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.9
DOI: https://doi.org/10.24867/06BE45Djurdjevic

AUTOMATSKO ODREĐIVANJE TEMA KNJIGA POMOĆU TEHNIKA ZA PROCESIRANJE PRIRODNOG JEZIKA

AUTOMATIC BOOK TOPIC DETECTION USING NATURAL LANGUAGE PROCESSING

Vlada Đurđević, *Fakultet tehničkih nauka, Novi Sad*

Oblast - RAČUNARSTVO I AUTOMATIKA

Kratak sadržaj - *Ovaj rad bavi se analizom performansi LDA modela kreiranog sa ciljem određivanja tema koje se pojavljuju u nekom korpusu knjiga. Opisan je skup podataka sa kojim se radi kao i svi problemi koji se javljaju prilikom implementacije ovakvog modela. Detaljno su analizirana četiri glavna koraka kreiranja modela, pretpocesiranje podataka, NER metoda, određivanje optimalnog broja tema i izbor konkretnog algoritma za implementaciju. Za svaki od koraka su demonstrirani različiti pristupi rešavanju problema koji se javljaju. Izvršena je evaluacija rezultata za svaki od ovih pristupa nakon čega je odabran optimalan pristup sa ciljem da čini sastavni deo krajnjeg modela.*

Ključne reči: *Latent Dirichlet Allocation, Named Entity Recognition*

Abstract - *This paper presents a performance analysis of an LDA model created for determining topics from a book corpus. A detailed analysis of four crucial steps regarding the implementation of the model is presented, data preprocessing, NER method, determining the optimal number of topics and choosing the best implementation algorithm. For each of the steps, a number of different methods for overcoming the problems that arise are demonstrated. The obtained results for each of the different methods are presented and discussed in detail. Finally, the optimal method is chosen to be a part of the resulting model.*

Keywords: *Latent Dirichlet Allocation, Named Entity Recognition*

1. UVOD

Procesiranje prirodnog jezika (*Natural language processing* - NLP) predstavlja granu veštačke inteligencije koja se bavi interakcijom između ljudi i računara uz pomoć prirodnog jezika. Krajnji cilj NLP-a je mogućnost čitanja, razumevanja i upotrebe različitih ljudskih jezika.

Jedna od metoda koja se koristi prilikom procesiranja ljudskog jezika je LDA (*Latent Dirichlet Allocation*), sa ciljem modelovanja kolekcija dokumenata i određivanja tema koje su u njima javljaju. Svaka pojedinačna tema predstavlja distribuciju verovatnoće nekih reči, tj. pruža nam verovatnoću da će se u njoj pojaviti neki određen skup reči.

Ideja na kojoj se temelji LDA jeste da se svaki dokument sastoji od skupa različitih tema a da se svaka od tih tema sastoji od skupa različitih reči. Ovaj master rad opisuje upotrebu i analizira rezultate dobijene uz pomoć LDA metode u okviru sistema za predikciju popularnosti knjiga [1]. Cilj ovakvog sistema je da pronađe i analizira sve atribute od interesa koji mogu uticati na popularnost neke knjige.

Pod popularnošću se smatra rejting knjige, tj. ocena koju je knjiga dobila na sajtu *Goodreads* [2]. LDA metod je korišćen prilikom analize naslova i sažetka knjige radi određivanja tema koje se pojavljuju u samoj knjizi. Takođe je upotrebljena NER (*Named Entity Recognition*) metoda radi uklanjanja ličnih imena iz tekstova koji čine sažetke knjiga. Isprobano je nekoliko različitih implementacija LDA modela pa je nakon toga odabran model koji daje najbolje rezultate. Ovaj model je potom primenjen na analizu sažetaka kao i na analizu naslova same knjige pošto se iz rezultata moglo zaključiti da nema potrebe za kreiranjem dva različita modela.

2. SRODNI RADOVI

Postoji velik broj radova koji se bave određivanjem skupa tema koje se pojavljuju u nekoj kolekciji dokumenata. U ovoj sekciji su predstavljeni radovi koji su najviše uticali na pristupe rešavanju problema koji se javljaju pri svakom od koraka kreiranja modela.

Rad [3] najbolje ilustruje osnovni cilj jednog LDA modela i predstavlja osnovu na kojoj se temelji ovaj master rad. U njemu je implementirano klasterovanje i vizualizacija dokumenata uz pomoć LDA metode i samoorganizujućih mapa (*Self-Organizing Maps* - SOM). Autori ovog rada su implementrirali LDA-SOM model i primenili ga na dva različita skupa podataka da bi evaluirali njegovu efektivnost. Korišćeni skupovi podataka su *20 Newsgroups*, koji se sastoji od 11,269 dokumenata i *Neural Information Processing Systems* (NIPS), koji se sastoji od 1500 dokumenata. Implementirani su samo osnovni koraci pretprocesiranja podataka a evaluacija modela je izvršena upotrebom metrika za evaluaciju SOM.

Naredni rad [4] se bavi isključivo veličinom skupa podataka nad kojim će se obučavati LDA model, tačnije brojem potrebnih dokumenata za kreiranje dobrog LDA modela. Intuitivno, deluje da će veći obučavajući skup uvek rezultovati boljim krajnjim modelom ali to ne mora uvek da bude slučaj. U okviru ovog rada autori su došli do zaključka da je obučavajući skup koji se sastoji od oko 40.000 dokumenata i 100.000 različitih termina dovoljan za kreiranje kvalitetnog LDA modela.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Aleksandar Kovačević, vanr.prof.

U okviru rada [5] fokus je bio na pretprocesiranju podataka. Cilj rada je bio pronalaženje tema koje se javljaju u grupi tvitova. Autori ovog rada su implementirali 7 različitih koraka pretprocesiranja, neophodnih za funkcionisanje krajnjeg LDA modela, nakon čega se izgubila skoro trećina obučavajućeg skupa.

Prilikom kreiranja LDA modela broj tema koje model treba da odredi u okviru datog korpusa mora biti unapred zadat. U navedenim radovima broj tema je određen ručno, na osnovu samog označenog skupa podataka ili nasumično. Nasuprot tome, rad [6] daje pregled metoda za automatsko određivanje optimalnog broja tema koje se pojavljuju u bilo kakvom skupu podataka. Jedna od najpopularnijih metoda koje se koriste sa ovim ciljem je računanje *perplexity*[1] vrednosti za niz modela sa različitim brojem tema a zatim odabir modela sa najnižom *perplexity* vrednošću.

Poslednji od radova koji će biti navedeni u ovom poglavlju se odnosi na vizualizaciju rezultata nekog LDA modela. Bez obzira na kvalitet samog LDA modela krajnji rezultati su samo skup brojeva koji predstavljaju verovatnoće tema. Takve rezultate je prilično teško analizirati a još teže prikazati korisniku na jednostavan način. U radu [7] autori predstavljaju potencijalno rešenje ovog problema, opisuju jedan od alata za vizualizaciju i poređenje rezultata modela.

3. METODOLOGIJA I ALATI

U ovoj sekciji se nalazi pregled svih metoda, algoritama, alata, okruženja i biblioteka korišćenih u okviru ovog rada.

3.1 *Latent Dirichlet Allocation*

Latent Dirichlet Allocation (LDA) predstavlja generativni probabilistički model neke kolekcije diskretnih podataka, poput tekstualnog korpusa. Sačinjava ga troslojni Bajesovski hijerarhijski model kod koga se svaki element neke kolekcije modeluje kao konačna kombinacija postojećih tema. Opšti cilj LDA modela je pronalaženje kratkog opisa za svaki element neke kolekcije, radi efikasnog procesiranja ogromnih kolekcija podataka, dok se u isto vreme očuvava esencijala statistička zavisnost među njima. LDA model podrazumeva da se svaki dokument W unutar korpusa D može dobiti prateći generativni proces definisan na sledeći način.

1. Odabere se N~ na osnovu Poasonove distribucije
2. Odabere se θ~ na osnovu Dirihleove distribucije
3. Za svaku od N rečiw_n:

3.1 Odabere se tema z_n ~ na osnovu Multinominalne distribucije

3.2 Odabere se reč w_n iz multinominalne verovatnoće $p(w_n|z_n,\beta)$

Kao posledica ovako definisanog generativnog procesa, dobija se sledeća verovatnoća za ceo korpus:

$$p(D|\alpha,\beta) =$$

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

Parametri α i β su jedinstveni na nivou korpusa i pretpostavlja se da se sempluju samo jednom prilikom generisanja korpusa. Varijable θ_d su jedinstvene na nivou pojedinačnog dokumenta i sempluju se jednom za svaki dokument, a varijable z_{dn} i w_{dn} su jedinstvene na nivou pojedinačnih reči i sempluju se jednom za svaku reč u svakom dokumentu.

3.2 *Named Entity Recognition*

Named Entity Recognition (NER) se koristi kao naziv za podkategoriju zadataka koje obavlja *Information Extraction* (IE), a koja se fokusira na prepoznavanje informacionih jedinica unutar nestruktuiranog teksta, poput imena ljudi, organizacija, lokacija, kao i datuma, vremena, i novčanih iznosa. Ključan problem koji NER sistemi treba da reše je identifikovanje unapred nepoznatih entiteta. Postoji nekoliko pristupa rešavanju ovog problema. U okviru ovog rada korišćena je alat koji implementira *Linear Chain Conditional Random Fields* model, koji spada u metode nadgledanog učenja.

3.3 NLTK

NLTK (*Natural Language ToolKit*) čini platformu za izgradnju *Python* programa koji se bave obradom običnog teksta. Pruža interfejse ka velikom broju različitih korpusa i leksičkih resursa, poput *WordNet*-a, kao i velik broj biblioteka za procesiranje teksta. Prilikom implementacije sistema, NLTK platforma je najviše korišćena za pretprocesiranja teksta, uz pomoć metoda tokenizacije, stemovanja, lematizacije i POS tagovanja.

3.4 MALLET

MALLET (*MAchine Learning for LanguagE Toolkit*) je *Java*-baziran paket za procesiranje prirodnog jezika, klasifikaciju dokumenata, određivanje tema koje se pojavljuju u dokumentima, ekstrakciju informacija a sadrži i velik broj drugih metoda mašinskog učenja za obradu teksta.

3.5 Stanford NER

Stanford NER predstavlja *Java*-baziranu implementaciju *Named Entity Recognition* sistema, koja u svom centru sadrži *Linear-chain Conditional Random Fields* model. U sklopu ove implementacije postoje tri različita klasifikaciona modela: 7-class, 4-class i 3-class model. U okviru ovog rada korišćen je 3-class model jer obuhvata entitete neophodne za rešavanje problema predstavljenog u radu a u isto vreme pruža značajno bolje rezultate od druga dva modela.

3.6 *Multidimensional scaling*

Multidimensional scaling (MDS) predstavlja tehniku za nelinearnu redukciju dimenzionalnosti. Podatke koji se nalaze u visoko dimenzionom prostoru, predstavlja preko matrice udaljenosti. Zatim pokušava da očuva taj kriterijum nakon projekcije podataka u prostor niže dimenzije. Svodi se na optimizacioni problem minimizacije funkcije koja predstavlja gubitak informacija.

3.7 Jensen-Shannon divergencija

Jensen-Shannon divergencija (JSD) predstavlja metod za merenje sličnosti između dve distribucije verovatnoće. Ova divergencija je simetrična pa se u opštem slučaju može koristiti kao mera udaljenosti, a takođe je i konačna što je čini pogodnom za računanje sličnosti između distribucija verovatnoće.

3.6 pyLDAvis

PyLDAvis predstavlja *Python* biblioteku kreiranu sa ciljem da pomogne korisnicima prilikom interpretacije

[1] *Perplexity* predstavlja meru koja određuje koliko dobro neki statistički model opisuje određen skup podataka, gde manja vrednost ukazuje na bolji statistički model.

tema koje se pojavljuju kao rezultati nekog modela za određivanje tema. Uzima potrebne informacije iz obučenog LDA modela i predstavlja ih u vidu interaktivne web aplikacije koja omogućava globalan pogled na sve teme, kao i detaljnu analizu svake pojedinačne teme i svih termina koji se u njoj pojavljuju. Teme se predstavljaju u 2D prostoru upotrebom MDS tehnike i Jensen-Shannon divergencije.

4. EKSPERIMENTALNA POSTAVKA

U ovoj sekciji će biti predstavljeni problemi rešavani u okviru ovog rada. Fokus rada je prvenstveno bio na kreiranju optimalnog LDA modela, i posebna pažnja je posvećena problemima koji treba da se reše prilikom kreiranja modela.

4.1 Problemi

Kreiranje modela je predstavljeno preko četiri glavna koraka: pretprocesiranje podataka, NER metoda, određivanje optimalnog broja tema i izbor konkretnog algoritma za implementaciju.

Prvi problem koji treba da se prevaziđe prilikom kreiranja LDA modela predstavlja pretprocesiranje podataka. U osnovi podaci predstavaljaju nestruktuirani tekst i treba da prođu dosta koraka pretprocesiranja pre nego što model može da ih upotrebi na željeni način. U okviru rada su upoređeni rezultati dobijeni različitim vidovima pretprocesiranja. Prvo je kreiran model bez pretprocesiranja podataka, zatim model koji implementira metod stemovanja a za kraj model koji implementira metod lematizacije i POS tagovanja. Takođe postoje neki koraci koji su zajednički za sve metode, poput tokenizacije reči, prebacivanje velikih slova u mala, izbacivanja znakova interpukcije i izbacivanja reči koje se izrazito često pojavljuju.

Naredni korak predstavlja NER metoda, koja takođe u nekoj meri predstavlja vid pretprocesiranja podataka.U okviru ovog rada se koristila *Stanford NER 3-class* imeplementacija ove metode, koja pronalazi i identifikuje imena ljudi, organizacija i lokacija u nestruktuiranom tekstu. Razlog korišćenja NER metode proističe iz činjenice da postoji velik broj knjiga koje imaju više nastavaka a glavni likovi ostaju isti. Zbog ovoga, imena tih likova mogu da predstavljaju reči od velikog značaja za temu koja je dominantna u tim knjigama. Posledica ove činjenice je da ukoliko se pojavi neka knjiga sa istim imenima likova postoji veća verovatnoća da će dobiti istu temu, iako ime lika ne bi trebalo da utiče na temu knjige.

Najkompleksniji korak prilikom kreiranja bilo kog LDA modela predstavlja određivanje optimalnog broja tema koji taj model treba da odredi. U okviru rada je isproban broj tema određen na osnovu sličnih radova, broj tema određen na osnovu žanrova knjige i broj tema određen na osnovu koherentnosti[2] krajnjeg modela.

Poslednji korak kreiranja modela predstavlja izbor algoritma za implementaciju. Isprobane su dve različite implementacije, Gensim i MALLET implementacija i pokazalo se da MALLET implementacija pruža bolje rezultate..

[2] Koherentnost predstavlja meru za evaluaciju kvaliteta tema koje odredi neki model, a ustvari određuje koliko je neka tema pogodna za interpretaciju od strane čoveka. U okviru ovog rada korišćena je Gensim implementacija mere koherentnosti koja predstavlja implementaciju četvorodelnog procesa definisanog u [8].

4.2 Rezultati

Za problem pretprocesiranja podataka najbolji rezultati su postignuti upotrebom metode lematizacije i POS tagovanja, ali treba imati u vidu da je ovaj pristup značajno vremenski zahtevniji od metode stemovanja. NER metoda je takođe rezultovala poboljšanim modelom ali se javlja isti problem vremenske zahtevnosti. Isto tako treba primetiti da je problem koji se rešava ovom metodom specifičan za domen problema definisanog u ovom radu, dok se u velikom broju drugih slučajeva on uopšte ne javlja. Najkompleksniji korak kreiranja modela je predstavljalo određivanje optimalnog broja tema. U ovom radu je cilj bio da se ostvari jedna od osnovnih ideja LDA modela, predstavljanje velike količine podataka uz pomoć relativno kratkog opisa. Radi ostvarivanja ovog cilja je pronađen najmanji mogući skup tema koje mogu da opišu dokumente iz korpusa na osnovu vrednosti za koherentnost modela, ali u opštem slučaju čak i dosta veći broj tema se ne bi mogao klasifikovati kao pogrešan. Konačno, za samu implementaciju modela je odabrana MALLET biblioteka na osnovu rezultata koherentnosti modela, iako je Gensim implementacija značajno brža. Rezultujuća koherentnost za MALLET implementaciju bila je 0.426 dok je za Gensim implementaciju bila samo 0.311. Teme određene krajnjim, optimalnim modelom su prikazane na slici 1.

Slika 1: *Teme optimalnog modela*

5. ZAKLJUČAK

Problem rešavan u okviru ovog rada je implementacija i analiza performansi LDA modela kreiranog sa ciljem određivanja tema koje se pojavljuju u nekom korpusu knjiga. Model je implementiran uz pomoć programskog jezika Python uz korišćenje NLTK, Gensim i MALLET programskih biblioteka kao i Stanford NER i pyLDAvis alata.

Model je kreiran sa ciljem da pronađe što manji skup tema koji je dovoljan da na ispravan način predstavi sve knjige iz obučavajućeg korpusa. Teme određene uz pomoć ovog modela su predstavljene kao vektor vrednosti verovatnoća i zatim korišćene u okviru sistema za predikciju popularnosti knjiga, gde se analizom rezultata algoritma koji određuje važnost atributa moglo primetiti da ovako određene teme u velikoj meri utiču na razultujuću popularnost neke knjige.

Detaljno su analizirana četiri glavna koraka kreiranja modela, pretpocesiranje podataka, NER metoda, određivanje optimalnog broja tema i izbor konkretnog algoritma za implementaciju. Preptocesiranje je obavljeno upotrebom Gensim i NLTK biblioteka i predstavlja osnovni korak prilikom implementacije modela ali takođe

i vremenski najzahtevniji. NER metoda je implementirana uz pomoć Stanford NER alata i fokus je bio isključivo na pronalaženju i uklanjanju ličnih imena iz korpusa dokumenata. Demonstrirani su različiti pristupi određivanju optimalnog broja tema, a za potrebe ovog rada je odabran pristup koji uključuje računanje vrednosti koherentnosti modela. Kao poslednji korak kreiranja modela su predstavljena dva različita načina implementacije modela, upotrebom Gensim i MALLET biblioteka, kao i razlika među njima.

Dalji pravci istraživanja uključuju analizu uticaja još različitih metoda pretprocesiranja, poput predstavljanja karakterističnih fraza u obliku bigrama ili trigrama, modifikaciju NER metoda sa ciljem uklanjanja imena lokacija i organizacija, izmenu samih algoritama za implementaciju sa ciljem optimizacije za konkretan problem definisan u ovom radu.

6. LITERATURA

[1] O. Hrnjaković, V. Đurđević, D. Bujiša, *Predikcija popularnosti knjiga*, Fakultet tehničkih nauka, Novi Sad, 2019

[2] Goodreads. (2018). [online] Dostupno na: https://www.goodreads.com/

[3] J. Millar, G. Peterson, M. Mendenhall, *Document Clustering and Visualization with Latent Dirichlet Allocation and Self-Organizing Maps*, Air Force Institute of Technology, 2009

[4] S. Crossley, M. Dascalau, D. McNamara, *How Important Is Size? An Investigation of Corpus Size and Meaning in both Latent Semantic Analysis and Latent Dirichlet Allocation*

[5] D. Alvarez-Melis, M. Saveski, *Topic Modeling in Twitter: Aggregating Tweets by Conversations*, Massachusetts Institute of Technology, Cambridge, MA, USA, 2016

[6] W. Zhao, J. Chen, R. Perkins, Z. Liu, W. Ge, Y. Ding, W. Zou, *A heuristic approach to determine an appropriate number of topics in topic modeling*, 2015

[7] J. Murdock, C. Allen, *Visualization Techniques for Topic Model Checking*, Program in Cognitive Science, Indiana University, USA

[8] M. Roder, A. Both, A. Hinneburg, *Exploring the Space of Topic Coherence Measures*, Leipzig University, R&D, Unister GmbH, Martin-Luther University, Germany

Kratka biografija:

Vlada Đurđević je rođen 24.04.1995. godine u Novom Sadu, Republika Srbija. Osnovnu školu „23. Oktobar" u Sremskim Karlovcima završio je 2010. godine. Nakon toga upisuje opšti smer u gimnaziji „Svetozar Marković" u Novom Sadu. Srednju školu završava 2014 godine. Iste godine se upisuje na Fakultet tehničkih nauka u Novom Sadu, odsek Elektrotehnika i računarstvo, smer Računarstvo i automatika. Školske 2016/17 se opredeljuje za usmerenje Primenjene računarske nauke i informatika, a potom školske 2017/18 za usmerenje Inteligentni sistemi. Zvanje diplomirani inženjer elektrotehnike i računarstva stekao je 2018. godine, sa prosečnom ocenom 8.72. Iste godine upisao je master akademske studije na smeru Računarstvo i automatika na Fakulteta tehničkih nauka u Novom Sadu. Uža specijalnost na master studijama bila je inteligentni sistemi. Položio je sve ispite predviđene planom i programom.