Modeliranje tema u društvenim znanostima: studija slučaja na bazi Web of Science

Maja Buhin Pandur, Jasminka Dobša
Sveučilište u Zagrebu
Fakultet organizacije i informatike
Pavlinska 2, Varaždin, Croatia
{mbuhin, jasminka.dobsa}@foi.unizg.hr

Luka Kronegger
Sveučilište u Ljubljani
Fakultet društvenih znanosti
Kardeljeva ploščad 5, Ljubljana, Slovenia
luka.kronegger@fdv.uni-lj.si

Sažetak. *Modeliranje tema jedan je od najpopularnijih zadataka koji se istražuju u području obrade prirodnog jezika, a latentna Dirichletova alokacija (engl. Latent Dirichlet Allocation, LDA) je jedna od najčešće korištenih tehnika za modeliranje tema. To je nenadzirana tehnika strojnog učenja koja kreira teme koristeći zbirku dokumenata na temelju riječi ili n-grama sa sličnim značenjem.*
U ovom smo radu primijenili metodu strukturalnog modeliranja s metodom LDA za izlučivanje tema iz znanstvenih radova s područja društvenih znanosti. Provedeno je strukturalno modeliranje na 3663 članaka iz Web of Science Core Collection baze iz razdoblja od 1999. do 2019. godine. Dobiveni rezultati ukazuju na to da se optimalan broj tema podudara s postojećim brojem definiranih područja istraživanja u društvenim znanostima ili s njegovim cjelobrojnim višekratnikom. Time se otvara područje za istraživanje usporedbe postojeće taksonomije i taksonomije predložene modelom LDA te za buduću identifikaciju interdisciplinarnosti.

Ključne riječi. modeliranje tema, latentna Dirichletova alokacija, strukturalno modeliranje tema, društvene znanosti

1 Uvod

U posljednjih nekoliko godina sve se više istražuje interdisciplinarnost između znanstvenih disciplina. Razvijaju se mnogi interdisciplinarni programi za rješavanje ključnih problema koje pojedinačna disciplina ne može riješiti. Interdisciplinarnost nije važna samo u akademskom svijetu već i u drugim područjima gdje donosi inovacije.

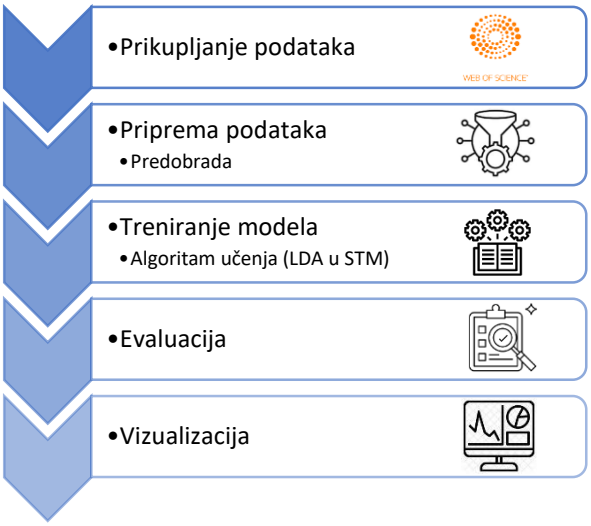
U literaturi visokog obrazovanja interdisciplinarna istraživanja definirana su kao „proces odgovora na pitanje, rješavanje problema ili rješavanje teme koja je preširoka ili presložena da bi se adekvatno mogla riješiti jednom disciplinom i u rješavanju se oslanja na više disciplina s ciljem integriranja njihovih uvida u izgradnju sveobuhvatnijeg razumijevanja“ (Repko, 2008). U scijentometriji, istraživanje interdisciplinarnosti se mjeri ispitivanjem mreže citata te, na primjer, mjerenjem postotka citata izvan glavne discipline citiranog rada.

Metode dubinske analize tekstih podataka već su korištene kako bi se automatski prepoznala interdisciplinarnost iz teksta (Ramage & Manning & Dumais, 2011), (Chuang i sur., 2012), (Nichols, 2014). Dietz i sur. (Dietz & Bickel & Scheffer, 2007) koristili su LDA za modeliranje tema kako bi mjerili utjecaj koji istraživački radovi imaju jedni na druge. Gerrish i Blei (Gerrish & Blei, 2010) pokazali su da se pomoću metode LDA može identificirati kvalitativno različit skup relevantnih članaka s obzirom na tradicionalnu mjeru brojanja citata. Na isti su način, Hall i sur. (Hall & Jurafsky & Manning, 2008) identificirali različite metodološke trendove u području računalne lingvistike u publikacijama tijekom 30 godina. Unatoč ovim studijama, koje su LDA uglavnom koristile kao metodu istraživanja korpusa, nije utvrđena korisnost LDA za prepoznavanje interdisciplinarnih radova. Nichols (Nichols, 2014) je predstavio novu metodu za mjerenje interdisciplinarnih istraživanja u portfelju nagrada Nacionalne zaklade za znanost. On je predložio korištenje tematskog modela Nacionalne zaklade za znanost te institucionalne strukture Nacionalne zaklade za znanost ispitivanjem prijedloga i nagrada za istraživačke stipendije, a ne publikacija. Nanni i sur. (Nanni & Dietz & Ponzetto, 2018.) istraživali su učinak LDA s rezultatima dobivenima korištenjem drugih metoda dubinske analize teksta, poput leksičkih značajki kod metode potpornih vektora (SVM) ili Rocchio klasifikatora za automatsku identifikaciju interdisciplinarnih djela iz korpusa sažetaka doktorskih disertacija. S obzirom na to, namjeravamo provjeriti korisnost modeliranja tema za prepoznavanje interdisciplinarnosti u člancima. U budućem radu uspoređivali bismo rezultate dobivene pomoću metode LDA s rezultatima dobivenim analizom društvenih mreža (SNA).

Krajnji je cilj predstavljene analize istražiti mogu li metode dubinske analize teksta, poput modeliranja tema korištenjem metode LDA, predstavljati valjanu alternativu za identifikaciju interdisciplinarnih područja istraživanja izravno iz tekstnih sadržaja naslova radova, sažetaka ili ključnih riječi. Cilj analize je pokazati kako je određeno znanstveno područje okarakterizirano pojmovima te kreiranje novih tema

korištenjem metode LDA za modeliranje. Kako bismo pronašli prikladan broj tema, istrenirali smo nekoliko modela koristeći različite izbore za broj tema te ih evaluirali mjerama semantičke koherencije (engl. semantic coherence), vjerojatnosti zadržavanja za skupove podataka (engl. likelihood for held-out datasets), rezidualima i donjom granicom marginalne vjerojatnosti.

Rad je organiziran na sljedeći način. U idućem poglavlju opisane su metode koje se koriste za modeliranje tema. Zatim je opisan skup podataka sa tehnikom predobrade, analiza podataka, nakon čega slijedi opis rezultata. Rad završava poglavljem koji se odnosi na budući rad. Dizajn istraživanja prikazan je na slici 1.



Slika 1. Grafički prikaz dizajna istraživanja.

2 Metode

2.1 Dubinska analiza teksta

Cilj dubinske analize teksta je otkrivanje relevantnih znanja koja su možda nepoznata ili prikrivena ispod očiglednih. Postoji nekoliko tipičnih nenadziranih i nadziranih tehnika dubinske analize teksta, kao što su kategorizacija teksta, grupiranje teksta, sažimanje dokumenata te ekstrakcija ključnih riječi. Modeliranje tema je tehnika dubinske analize teksta koja koristi nadzirane i nenadzirane tehnike strojnog učenja.

2.2 Modeliranje tema metodom latentne Dirichletove alokacije

Modeliranje tema je statistička metoda kojoj je cilj otkrivanje apstraktnih "tema" u skupu dokumenata. To je tehnika nenadziranog strojnog učenja s obzirom da ne zahtijeva trenirani skup podataka ili unaprijed definirani popis tema. Teme se kreiraju iz različitih dokumenata na temelju riječi ili izraza sa sličnim značenjem. Jedna od najpopularnijih metoda modeliranja tema je latentna Dirichletova alokacija (LDA). LDA pokušava sve dokumente organizirati prema temama na način da latentne teme prvenstveno definiraju riječi.

Metoda latentne Dirichletove alokacije za modeliranje tema tretira svaki dokument kao mješavinu tema, a svaku temu kao mješavinu riječi. Ova metoda omogućava da se dokumenti međusobno "preklapaju" u smislu sadržaja, umjesto da se razdvajaju u zasebne skupine, što na neki način odražava tipičnu upotrebu prirodnog jezika ("Text Mining with R", 2020). LDA se široko koristi u brojnim primjenama strojnog učenja, obrade prirodnog jezika (engl. natural language processing, NLP) i aplikacijama za pronalaženje informacija. Griffiths i Steyvers (Griffiths & Steyvers, 2004.) koristili su metode LDA za izlučivanje znanstvenih tema u zbirci dokumenata.

LDA je generativni, probabilistički hijerarhijski Bayesov model koji inducira teme iz zbirke dokumenata u tri koraka (Blei & Ng & Jordan, 2003) (slika 2):

1. Svaki dokument u zbirci raspoređen je po temama koje su uzorkovane za taj dokument na temelju Dirichletove distribucije.
2. Svaka riječ u dokumentu povezana je s jednom temom na temelju odabrane Dirichletove distribucije.
3. Svaka je tema označena kao multinomijalna distribucija nad riječima koje su dodijeljene uzorkovanoj temi.

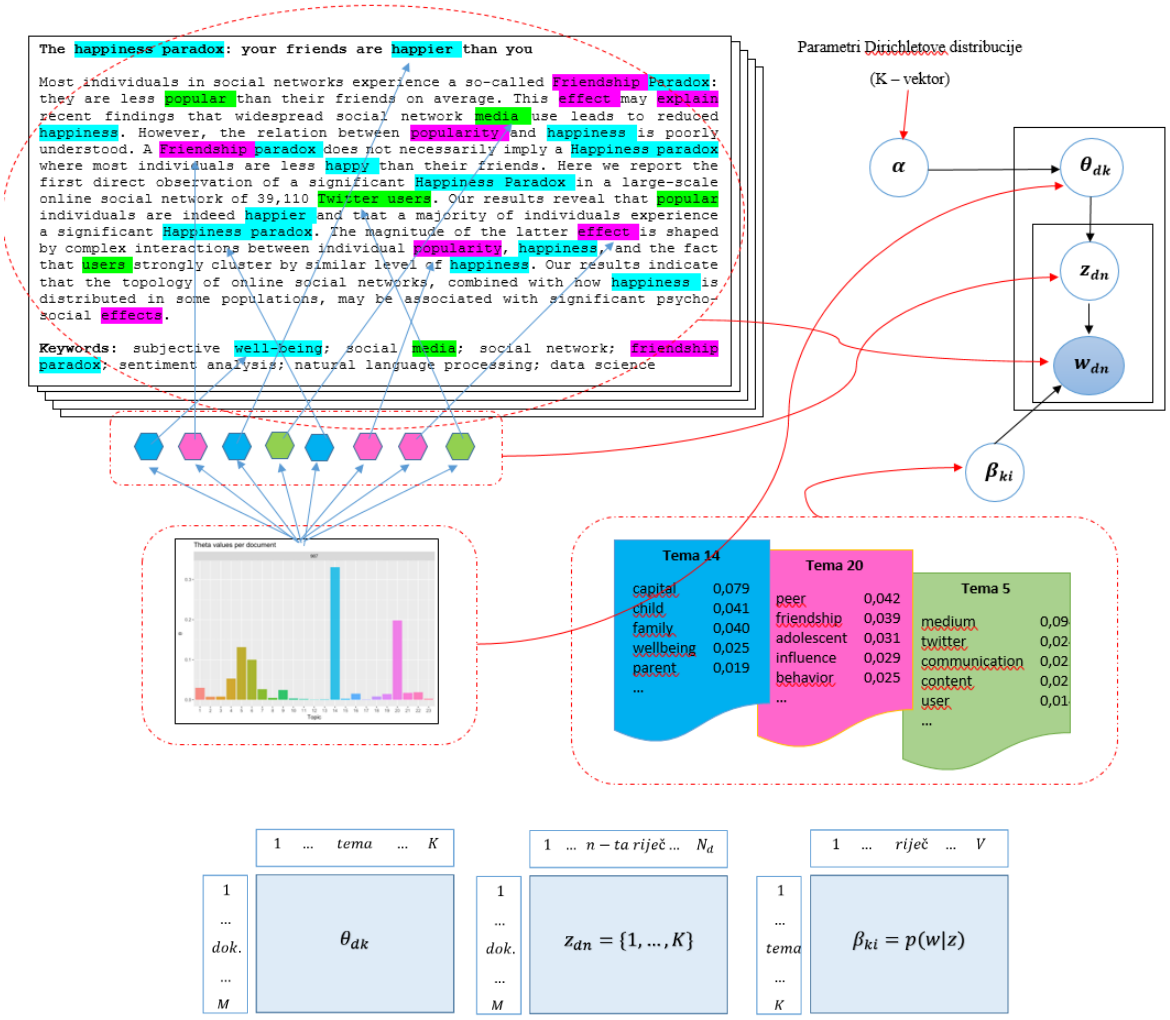
U radu su korištene sljedeće oznake:

- M – broj dokumenata,
- N – broj riječi u svakom dokumentu,
- $\boldsymbol{w}$ – reprezentacija dokumenta prikazanog kao jedinični osnovni vektor $\boldsymbol{w} = (w_1, w_2, \dots, w_N)$, gdje je w_n $n-ta$ riječ u nizu; vektor $\boldsymbol{w}$ ima jednu komponentu jednaku 1, a sve ostale su 0,
- V – veličina rječnika gdje je $v-ta$ riječ u rječniku predstavljena s V – dimenzionalnim vektorom takvim da je $w^v = 1$ i $w^u = 0$ za $u \neq v$,
- D – korpus od M dokumenata, prikazan kao $D = (\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_M)$
- k – broj tema kojima dokument pripada,
- z – tema iz skupa od k tema.

Vjerojatnost promatranog skupa podataka izračunava se i dobiva iz korpusa D na sljedeći način:

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

gdje su varijable θ_d varijable na razini dokumenta, uzorkovane jednom po dokumentu, a varijable z_{dn} i w_{dn} su varijable na razini riječi te su uzorkovane jednom za svaku riječ u svakom dokumentu.



Slika 2. Grafički prikaz LDA modeliranja tema prikazanog na jednom od dokumenata iz korpusa

Parametar α ima distribuciju tema θ za svaki dokument (slika 2). Svaki od M dokumenata ima neku θ distribuciju. θ je nasumična matrica veličine $(M \times k)$ gdje $\theta(i,j)$ predstavlja vjerojatnost $i-tog$ dokumenta koji sadrži riječi koje pripadaju $j-toj$ temi. θ ima Dirichletovu distribuciju $\mathrm{Dir}(\alpha)$.

Pretpostavimo da postoji jedan dokument s N riječi te je svaka riječ iz tog dokumenta generirana temom. Generirali smo N tema koje sadrže riječi. Na temelju skalarnog parametra η za svaku temu, β također ima Dirichletovu distribuciju. β generira k pojedinačnih riječi za svaku temu prema Dirichletovoj distribuciji. Slično tome, β je matrica veličine $(k \times V)$ gdje $\beta(i,j)$ predstavlja vjerojatnost da $i-ta$ tema pripada $j-toj$ riječi.

Za treniranje LDA modela trebamo procijeniti parametre α, θ, η i β. α je parametar povezan s distribucijom koja određuje kako izgleda distribucija tema za sve dokumente u korpusu, θ je slučajna matrica za koju je $\theta(i,j)$ vjerojatnost da $i-ti$ dokument sadrži $j-tu$ temu, η je parametar povezan s distribucijom koja određuje način na koji se riječi raspoređuju u svakoj temi, a β je slučajna matrica za koju je $\beta(i,j)$ vjerojatnost da $i-ta$ tema sadrži $j-tu$ riječ. LDA je probabilistički model, stoga trebamo izračunati zajedničku distribuciju za mješavinu tema θ, $P(\theta, z, \beta | D; \alpha, \eta)$. Za skup od M dokumenata, gdje svaki dokument sadrži N riječi i svaka riječ je generirana jednom od k tema, trebamo tražiti posteriornu zajedničku vjerojatnost od θ, z i β, danu s D te pomoću parametara α i η. Rješenje ovog problema dano je u radu Bleia i sur. (Blei & Ng & Jordan, 2003).

Za primjenu LDA kod modeliranja tema, koristili smo strukturalno modeliranje tema (STM) koji je implementiran u **stm** R paketu ("stm: R Package for Structural Topic Models", 2020).

2.3 Mjere za evaluaciju modela

Mjere koje se koriste za pronalaženje prikladnog broja tema opisane su u nastavku.

Semantička koherencija je mjera koju su uveli Mimno i sur. (Mimno i sur., 2011). Ta se mjera za temu k izračunava korištenjem liste od N najvjerojatnijih riječi u temi k:

$$C_k = \sum_{n=2}^{N} \sum_{m=1}^{n-1} \log\left(\frac{D(v_n, v_m) + 1}{D(v_m)}\right) \quad (2)$$

gdje je $\boldsymbol{v}$ vektor od prvih N riječi u temi poredanih silaznim redoslijedom, $D(v)$ je broj dokumenata od najmanje jedne riječi v, a $D(v, v')$ je broj zajedničkog pojavljivanja riječi v i v' u dokumentu. Intuitivno, ovo je zbroj po svim parovima riječi u top tematskim riječima, pri čemu se djeluje logaritamskom funkcijom na frekvenciju zajedničkog pojavljivanja riječi podijeljenu s osnovnom frekvencijom riječi. Brojniku se dodaje jedan kako bi se spriječilo da vrijednost logaritma bude nula u slučaju da se par riječi nikad ne pojavi. Semantička koherencija postiže maksimum kada se riječi s najvećom vjerojatnošću u određenoj temi često javljaju zajedno. Također, to je mjera koja dobro korelira sa ljudskom prosudbom kvalitete teme. Ako postoji nekoliko tema koje dominiraju s prevladavajućim riječima, tada je potrebno sagledati, uz semantičku koherentnost, i ekskluzivnost.

Ekskluzivnošću se mjeri razlika između tema uspoređujući sličnosti distribucije riječi β u različitim temama. Tema je ekskluzivna ako se top riječi te teme ne pojavljuju među ostalim temama. Ekskluzivnost za riječ v u temi k je definirana kao:

$$EX_{k,v} = \frac{\beta_{k,v}}{\sum_{j=1}^{K} \beta_{j,v}} \quad (3)$$

Frekvencija i ekskluzivnost su važni čimbenici u određivanju semantičkog sadržaja riječi. Prema tome, univarijantna mjera važnosti teme može biti korisna aproksimacija za različite zadatke poput smanjenja dimenzionalnosti, odabira značajki i otkrivanja sadržaja. Stoga je modificirana harmonijska sredina za pomicanje „prosječnog“ ranga na niži rezultat (Bischof & Airoldi, 2012). Metrika označavanja frekvencijske ekskluzivnosti (FREX) je vagana harmonijska sredina ranga riječi dana kao:

$$FREX_{k,v} = \left(\frac{w}{ECDF(EX_{k,v})} + \frac{1 - w}{ECDF(\beta_{k,v})}\right)^{-1} \quad (4)$$

gdje ECDF predstavlja empirijsku kumulativnu funkciju distribucije za riječ v u njezinoj distribuciji teme β_k, a w je težina za ekskluzivnost postavljena na 0.7 u našim pokusima.

Još jedna mjera za usporedbu modela kojom se provjerava koliko dobro svaki model predviđa riječi u dokumentu je procjena vjerojatnosti zadržavanja (engl. held-out likelihood estimation). U **stm** paketu, procjena vjerojatnosti zadržavanja koristi dvije funkcije (“stm: R Package for Structural Topic Models”, 2020). Prva funkcija koristi metodu vjerojatnost zadržavanja potpunosti dokumenta (engl. document-completion held-out likelihood), koja je procjena vjerojatnosti pojavljivanja riječi u dokumentu kada su te riječi uklonjene iz dokumenta u koraku procjene. Druga funkcija evaluira vjerojatnost zadržavanja (engl. held-out likelihood) za riječi koje nedostaju na temelju modela izvedenog na zadržanim dokumentima. Mjera procjene vjerojatnosti zadržavanja slična je unakrsnoj validaciji te pomaže u procjeni izvedbe predviđanja modela.

Pretpostavke modela mogu se testirati i pomoću reziduala. Funkcija `residuals` u **stm** paketu mjeri postoji li prekomjerna disperzija varijance multinomijalne varijance unutar LDA metode generiranja podataka. Kao što je spomenuto u (Taddy, 2012), ako su reziduali previše raspršeni, možda će biti potrebno više tema kako bi se apsorbirale neke dodatne varijance. Iako ne postoji određena metoda za odabir broja tema, i provjera reziduala i procjena vjerojatnosti zadržavanja su korisne mjere za određivanje broja tema.

Donja granica marginalne vjerojatnosti je mjera konvergencije modela. Jednom kada granica ima dovoljno malu promjenu između iteracija, model se smatra konvergiranim.

3 Rezultati eksperimenta

3.1 Skup podataka i predobrada podataka

Analizirani skup podataka dobiven je iz baze podataka Web od Science (WoS) Core Collection pretraživanjem članaka koji sadrže frazu *social netwok** u WoS-ovom području društvenih znanosti u razdoblju od 1999. do 2019. godine. Fraza *social network** koristi se kako bi se suzio promatrani skup podataka. Pretraga je izvršena u ožujku 2020. godine, a ukupno je preuzeto 3664 članaka. Svaki od ovih članaka opisan je nizom metapodataka kao što su autor(i), naslov, sažetak, ključne riječi, područje istraživanja i godina izdanja. Glavna ideja je istražiti istraživačke teme na području društvenih znanosti iz brojnih riječi iz naslova, sažetaka i ključnih riječi.

Prema klasifikaciji iz WoS-a, postoji 25 kategorija u području istraživanja društvenih znanosti. Svi su članci razvrstani u najmanje jednu kategoriju. Samo dvije kategorije, *arheologija* i *razvojne studije*, nisu sadržavale članke iz našeg skupa podataka. Većina članaka kategorizirana je u *biomedicinskim društvenim znanostima, poslovanju i ekonomiji, matematičkim metodama društvenih znanosti, psihologiji, društvenim znanostima - ostale teme* te *sociologiji*. Naslovi, sažeci i ključne riječi autora izvučeni su iz svakog članka te spojeni kako bi se dobio skup podataka jedne varijable s 3664 instanci. Korpus je indeksiran ukupno s 29199 indeksnih pojmova.

Prije analize, skup podataka uređen je pomoću **tm** paketa iz **R** za uklanjanje stop riječi, interpunkcija, brojeva, nepotrebnih znakova i razmaka. Riječi *social*, *network*, *study*, *analysis*, *model* i *datum* također su uklonjene iz vektora riječi kako bi se smanjio negativan utjecaj u analizi te je provedena lematizacija.

Kreirana je matrica dokumenata i pojmova u kojoj redovi predstavljaju dokumente, a stupci pojmove iz dokumenata. Nakon opisanih koraka predobrade, matrica dokumenata i pojmova, stvorena na takav način, sadržavala je 3663 dokumenta i 20718 pojmova. Da bismo smanjili rijetkost matrice, izbacili smo pojmove koji se pojavljuju u samo jednom dokumentu. Nakon toga, matrica dokumenata i pojmova je sadržavala 3663 dokumenta i 9096 pojmova. Konačno, svaka vrijednost matrice imala je vrijednosti iz frekvencije pojmova – inverzne frekvencije dokumenata (TF-IDF). Frekvencija pojmova (TF) mjera je važnosti pojma u dokumentu. Na slici 3 prikazan je stupčasti grafikon pojmova koji se najčešće pojavljuju u korpusu od 3663 dokumenta. Inverzna frekvencija dokumenta (IDF) je mjera koja penalizira najčešće pojmove. Množenjem IDF-a s TF, dobije se TF-IDF mjera važnosti pojma u dokumentu korpusa. TF-IDF ocjena za pojam t u dokumentu d iz skupa dokumenta D , a izračunata je na sljedeći način:

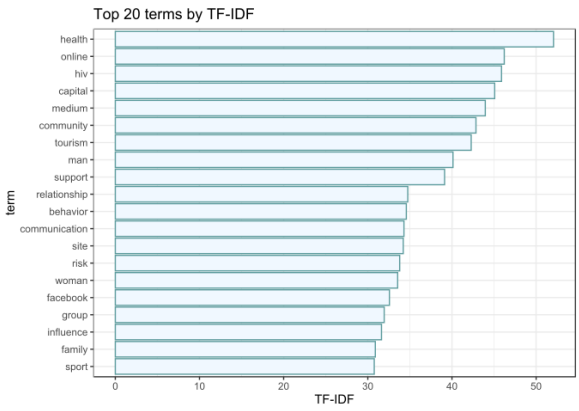
$$TF - IDF = TF \cdot IDF \tag{5}$$

gdje je

$$TF = \log(1 + freq(t, d)) \tag{6}$$

$$IDF = \log\left(\frac{M}{count(d \in D : t \in d)}\right) \tag{7}$$

M je ukupni broj dokumenata, a $freq(t, d)$ frekvencija pojma t u dokumentu d .



Slika 3. Stupčasti dijagram pojmova s najvećom TF-IDF težinom u skupu podataka.

3.2 Analiza podataka

U sljedećem smo koraku iz korpusa izgradili matricu dokumenata i pojmova temeljem samo frekvencije pojavljivanja pojma u dokumentima i primijenili metodu LDA. Prije procjene LDA, trebalo se definirati broj tema. Istrenirana je grupa modela s različitim brojem tema te su ti modeli evaluirani kako bi se procijenilo koliko je tema prikladno za dani korpus. Nakon postavljanja različitih vrijednosti za broj tema (k) od 2 do 100, istraženo je koliko je tema prikladno. Te su vrijednosti intuitivno uzete kako bi odgovarale modelu jer u WoS-u postoji 25 kategorija iz područja društvenih znanosti.

Model LDA ima dva pristupa za istraživanje tema koje se procjenjuju. Prvi pristup je promatranje povezanosti pojmova s temama, a drugi je pristup ispitivanje dokumenata za koje se procjenjuje da su u značajno povezano s određenom temom.

Model je ocijenjen mjerama semantičke koherencije, vjerojatnosti zadržavanja skupova podataka, reziduala i donje granice te su izrađeni dijagnostički prikazi za razumijevanje funkcioniranja modela za različit broj tema temeljem čega je odabran određeni broj tema. Teme se zatim uspoređuju s centroidima kategorija iz WoS-a koristeći kosinus mjeru sličnosti. Centroidi za specifičnu WoS kategoriju k dani su sa:

$$Cent_k = \frac{1}{n_k}\sum_{j=1}^{n_k} \vec{d}_{k,j} \tag{8}$$

gdje je $\vec{d}_{k,j}$ vektorski prikaz dokumenta j u WoS kategoriji k, a n_k je broj dokumenata iz WoS kategorije k. Kosinus mjera sličnosti mjeri sličnost računajući kosinus kuta između dva vektora $\vec{a}$ i $\vec{b}$:

$$\cos(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{||\vec{a}|| \cdot ||\vec{b}||} \tag{9}$$

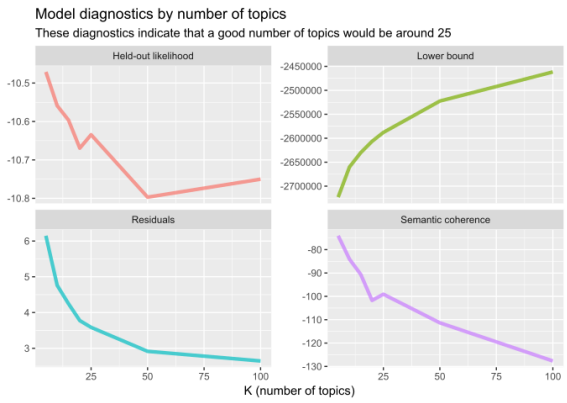
Sličnost između tema dobivenih metodom LDA i tema iz WoS kategorija mjeri se kao kosinus mjera sličnosti između vektora riječi distribucije vjerojatnosti tema i centroida za određenu kategoriju iz WoS-a. Vrijednosti kosinus kuta su između 0 i 1 s obzirom da oba vektora imaju pozitivne vrijednosti elemenata. Smatra se da su teme iz modela slične kategorijama iz WoS-a ako je vrijednost kosinusa veća od 0.5.

3.3 Rezultati

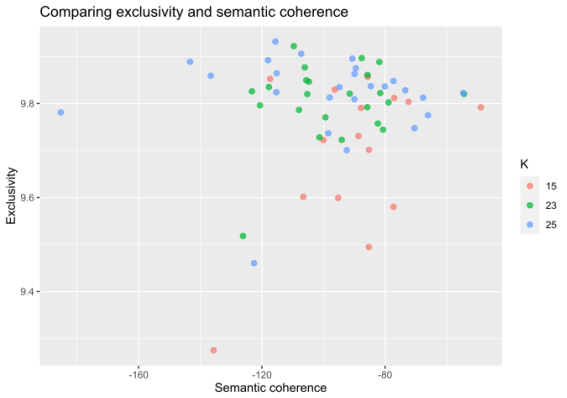
Nakon evaluacije modela sa semantičkom koherencijom, vjerojatnošću zadržavanja, rezidualima i donjom granicom napravljeno je nekoliko analitičkih grafikona koristeći ove iznose da bismo znali kako modeli funkcioniraju na nizu tema (slika 4). Iz dijagnostičkih grafikona možemo vidjeti da bi dobar broj tema mogao biti oko 25, jer oko te vrijednosti rast/pad pripadnih mjera evaluacije usporava. Nakon analize semantičke koherencije i ekskluzivnosti pojmova prema temama, može se pretpostaviti da bi dobar izbor broja tema bio 23 (slika 5).

Sljedeći rezultati opisani su na dva načina. Prvi je pristup pogledati skupove pojmova koji su zajednički s temama. Drugi pristup je sagledavanje stvarnih dokumenata za koje se procjenjuje da su značajno povezani s određenom temom. Oba ova pristupa prikazana su na slici 6. Među 23 najzastupljenije teme,

najviše dokumenata pripada temi 22. Također možemo vidjeti da je nekoliko tema usredotočeno na zdravlje (tema 17 s najvjerojatnijim pojmovima *HIV, man, sex, sexual, risk*; tema 16 s najvjerojatnijim pojmovima *health, support, age, old, adult*; tema 15 s najvjerojatnijim pojmovima *health, care, support, service, access*), komunikacija, internet i društvene mreže (tema 5 s najvjerojatnijim pojmovima *medium, Twitter, communication, content, user*, tema 4 s najvjerojatnijim pojmovima *online, site, Internet, Facebook, sns*), turizam, političke teme ili poslovanje i ekonomija.



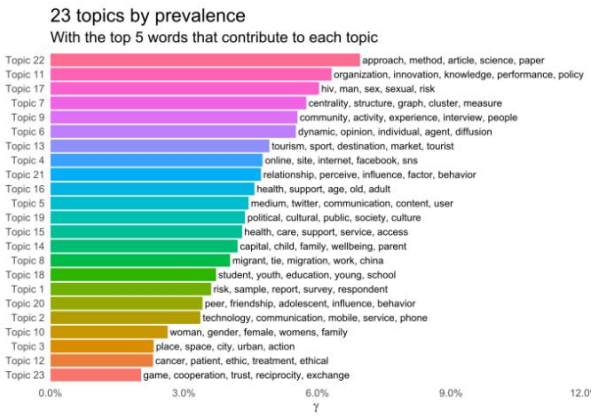
Slika 4. Dijagnostika modela prema broju tema pokazuje da je prikladan broje tema oko 25.



Slika 5. Usporedba semantičke koherencije i ekskluzivnosti.

Kosinus mjera sličnosti između vektora koja se temelji na distribuciji vjerojatnosti riječi iz tema i centroida za odabrane WoS kategorije (BE – *poslovanje i ekonomija*, BSS – *biomedicinske društvene znanosti*, COM – *komunikacija*, Edu – *obrazovanje i istraživanje obrazovanja*, FS – *obiteljske studije*, GL – *vlada i pravo*, MathM – *matematičke metode u društvenim znanostima*, Psy – *psihologija*, SI – *socijalna pitanja*, Soc – *sociologija*, SS – *društvene znanosti - ostale teme*) prikazana je u tablici 1.

Iz rezultata prikazanih u tablici 1 možemo vidjeti da neke teme imaju nešto zajedničko s kategorijama iz WoS-a. Na primjer, tema 7 povezana je s *matematičkim metodama u društvenim znanostima*, tema 17 s *biomedicinskim društvenim znanostima*, a tema 11 s *poslovanjem i ekonomijom.*



Slika 6. 23 tema poredanih prema prevalenciji unutar čitavog korpusa te prvih pet pojmova povezanih s temom.

Tablica 1. Kosinus mjera sličnosti između tema i kategorija iz WoS-a

	BE	BSS	COM	Edu	FS	GL	MathM	Psy	SI	Soc	SS
1	0.17	0.34	0.10	0.07	0.16	0.13	0.22	0.30	0.20	0.24	0.24
2	0.23	0.16	0.24	0.11	0.14	0.13	0.20	0.19	0.29	0.18	0.31
3	0.20	0.20	0.09	0.21	0.16	0.21	0.20	0.22	0.22	0.28	0.29
4	0.21	0.16	0.32	0.11	0.13	0.20	0.14	0.24	0.29	0.19	0.36
5	0.21	0.13	0.26	0.13	0.11	0.18	0.18	0.20	0.17	0.15	0.32
6	0.36	0.25	0.10	0.16	0.21	0.22	0.54	0.27	0.20	0.35	0.37
7	0.40	0.23	0.08	0.17	0.16	0.13	0.68	0.28	0.17	0.43	0.33
8	0.32	0.26	0.10	0.16	0.31	0.30	0.26	0.30	0.23	0.35	0.35
9	0.20	0.28	0.11	0.22	0.21	0.18	0.18	0.30	0.22	0.29	0.32
10	0.13	0.29	0.06	0.13	0.33	0.15	0.11	0.24	0.15	0.18	0.19
11	0.56	0.24	0.12	0.23	0.16	0.27	0.37	0.29	0.26	0.35	0.45
12	0.21	0.35	0.08	0.14	0.12	0.18	0.17	0.32	0.29	0.22	0.26
13	0.46	0.15	0.12	0.20	0.14	0.16	0.18	0.19	0.18	0.30	0.42
14	0.22	0.27	0.08	0.22	0.19	0.15	0.15	0.30	0.26	0.37	0.31
15	0.15	0.49	0.06	0.13	0.16	0.14	0.13	0.33	0.24	0.29	0.24
16	0.17	0.50	0.07	0.17	0.16	0.14	0.17	0.38	0.22	0.35	0.27
17	0.14	0.60	0.09	0.07	0.47	0.10	0.14	0.50	0.17	0.19	0.22
18	0.20	0.20	0.14	0.44	0.16	0.19	0.17	0.32	0.30	0.32	0.33
19	0.23	0.18	0.17	0.22	0.17	0.37	0.20	0.26	0.25	0.26	0.36
20	0.27	0.29	0.19	0.12	0.15	0.16	0.35	0.30	0.23	0.33	0.28
21	0.26	0.27	0.12	0.13	0.16	0.16	0.22	0.32	0.27	0.29	0.30
22	0.37	0.23	0.12	0.29	0.20	0.24	0.37	0.27	0.24	0.34	0.42
23	0.31	0.24	0.10	0.27	0.21	0.13	0.32	0.30	0.29	0.30	0.30

Također možemo primijetiti da se neke teme ne preklapaju sa samo jednom kategorijom iz WoS-a, već s nekoliko njih, pa možemo pretpostaviti da postoji interdisciplinarnost između ovih znanstvenih disciplina iz WoS-a, što i ne čudi s obzirom na to da je većina članaka kategorizirana u više disciplina. Tema 22 ima približno jednake vrijednosti kosinus mjere sličnosti za *poslovanje i ekonomiju, matematičke*

metode iz društvenih znanosti, *sociologije* i *društvenih znanosti - ostale teme*, dok tema 17 ima približno jednake vrijednosti kosinus mjere sličnosti za *biomedicinske društvene znanosti*, *obiteljske studije* i *psihologiju*. Možemo primijetiti da pojedina područja iz kategorija WoS-a imaju približno jednake vrijednosti kosinus mjere sličnosti za različite teme. Pretpostavljamo da je to zato što je algoritam LDA identificirao potkategorije unutar ove kategorije. Na primjer, *sociologija* ima približno jednake vrijednosti kosinus mjere sličnosti za teme 6, 7, 8, 11, 14, 16, 20 i 22.

4 Zaključak

U ovom je radu dani kratki uvod u modeliranje tema latentnom Dirichletovom alokacijom (LDA), koje je primijenjeno korištenjem strukturalnog modeliranja tema (STM) na skupu podataka iz Web of Science Core Collection.

Glavni cilj istraživanja bio je usporediti teme dobivene LDA modeliranjem tema s kategorijama iz baze Web of Science za područje društvenih znanosti. Istraživanje je provedeno na uzorku radova od 1999. do 2019. godine s ključnim upitom *social networks**, a rezultati su ograničeni na publikacije koje sadrže tu frazu. Usporedba teme koju je metoda LDA dobila i dane taksonomije u pogledu kosinus mjere sličnosti ukazuje na to da se društvene mreže uglavnom primjenjuju u disciplinama *poslovanje i ekonomija* (BE), *biomedicinske društvene znanosti* (BSS), *matematičke metode društvenih znanosti* (MathM) i *psihologiji* (Psy), koja se čini intuitivnim rezultatom. Nadalje, na temelju kosinus mjere sličnosti, uspjeli smo identificirati i interdisciplinarnost između disciplina BE i MathM, BSS i MathM, BSS i Psy, BSS, FS i Psy.

Na temelju intuitivnih rezultata dobivenih na ovom uzorku radova, planira se proširenje istraživanja na sve radove iz zbirke u području društvenih znanosti kako bi identificirali interdisciplinarna područja i proveli daljnja istraživanja analizom društvenih mreža i simboličkih podataka.

U budućim istraživanjima namjerava se istražiti interdisciplinarnost između skrivenih ili maskiranih znanstvenih disciplina te preispitati postojeća taksonomija istraživačkih područja u društvenim znanostima i njezine promjene kroz vrijeme.

Reference

Bischof, J., Airoldi, E. (2012). Summarizing Topical Content with Word Frequency and Exclusivity. *Proceedings of the 29th International Conference on Machine Learning, ICML '12* (pp. 201–208). New York: J. Langford, J. Pineau (eds.).

Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, pp. 993-1022.

Chuang, J., Ramage, D., Manning, C., Heer, J. (2012). Interpretation and trust: designing model-driven visualizations for text analysis. *SIGCHI Conference on Human Factors in Computing Systems*, (pp. 443-452). Austin, Texas, USA.

Dietz, L., Bickel, S., Scheffer, T. (2007). Unsupervised prediction of citation influences. *24th international conference on Machine learning* (pp. 233-240). Corvalis, Oregon, USA: Association for Computing Machinery, New York, United States.

Gerrish, S., Blei, D. (2010). A Language-based Approach to Measuring Scholarly Impact. *27th International Conference on Machine Learning*, (pp. 375-382). Haifa, Israel.

Griffiths, T., Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*, pp. 5228-5235.

Hall, D., Jurafsky, D., Manning, C. D. (2008). Studying the History of Ideas Using Topic Models. *Conference on Empirical Methods in Natural Language Processing*, (pp. 363–371).

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., McCallum, A. (2011). Optimizing semantic coherence in topic models. *Conference on Empirical Methods in Natural Language Processing (EMNLP '11)* (pp. 262-272). USA: Association for Computational Linguistics.

Nanni, F., Dietz, L., Ponzetto, S. P. (2018). Toward a computational history of universities: Evaluating text mining methods for interdisciplinarity detection from PhD dissertation abstracts. *Digital Scholarship in the Humanities, Volume 33, Issue 3*, pp. 612–620.

Nichols, L. G. (2014). A topic model approach to measuring interdisciplinarity at the National Science Foundation. *Scientometrics 100(3)*, pp. 741-754.

Ramage D., Manning C. D., Dumais S. (2011). Partially labelled topic models for interpretable text mining. *17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 457-465). San Diego California USA: Association for Computing Machinery New York NY United States.

Repko, A. F. (2008). *Interdisciplinary Research: Process and Theory.* California: Sage: Thousand Oaks.

Roberts, M. E., Stewart, B. M., Tingley, D. (August 2020). *stm: R Package for Structural Topic Models.* Preuzeto iz The Comprehensive R Archive Network: http://www.structuraltopicmodel.com/

Silge, J., Robinson, D. (August 2020). *Text Mining with R.* Preuzeto iz https://www.tidytextmining.com/topicmodeling.html

Taddy, M. A. (2012). On Estimation and Selection for Topic Models. *The 15th International Conference on Artificial Intelligence and Statistics.*, (pp. 1184-1193).