NOVI PRISTOPI
Odkrivanje semantičnih struktur v etnoloških vsebinah

Izvirni znanstveni članek | 1.01

Izvleček: Članek obravnava strojne (računalniške) metode, ki omogočajo avtomatsko semantično analizo in ekstrakcijo smiselnih vzorcev iz vsebin. Te metode so v zadnjem času vse aktualnejše in pogosto nujne, zlasti pri analizi obsežnejših digitalnih zbirk, kjer ročna klasifikacija ali tipološka ureditev gradiva nista mogoči. S semantično analizo raziskovalcu omogočimo pogled na širšo konceptualno zasnovo gradiva, pregled vsebin po izbranih parametrih ter možnost odkrivanja semantičnih struktur (npr. tematskih vzorcev) in konteksta.
V članku gre za folkloristične vsebine, konkretno za analizo besedil ljudskih pesmi.

Ključne besede: strojne metode, semantična analiza, obdelava naravnega jezika, folkloristika, ljudske pesmi, LSA (latentna semantična analiza), LDA (latentna Dirichletova razporeditev)

Abstract: The article addresses computational approaches to semantic analysis and extraction of meaningful structures from the contents. These methods are becoming more relevant and often necessary, especially in the analysis of large digital collections where manual classification of materials is not possible. Moreover, computational semantic analysis offers insight into the broader conceptual structure of the contents; selection of desired parameters; and discovery of a general semantic structure (e.g. thematic patterns) and context. The materials analyzed for this purpose contain a corpus of Slovene folk songs.

Key Words: computational methods, semantic analysis, natural language processing, folklore, folk songs, LSA (Latent Semantic Analysis), LDA (Latent Dirichlet Allocation)

Uvod

Razvoj informacijsko-komunikacijskih tehnologij in težnje po ustvarjanju digitalnih vsebin so močno vplivali na pomen digitalnih virov v humanistiki. Medtem ko je bila v preteklosti pozornost namenjena predvsem digitalni prezervaciji in razvoju sistemov, ki bi v kar se da veliki meri ta proces avtomatizirali (npr. s standardizacijo formatov, metapodatkovnih shem in ontologij, izdelavo orodij za produkcijo in manipuliranje digitalnih vsebin, avtomatizacijo postopkov arhiviranja itn.), je danes vse aktualnejša uporaba digitalnih virov pri raziskovalnem delu. S premostitvijo fizičnih omejitev so se namreč odprle možnosti novih, inovativnih raziskav, ki temeljijo na razvoju in uporabi strojnih (računalniških) metod in tehnologij pri iskanju, analizi in interpretaciji digitalnih vsebin. Tu se bomo osredinili na metode strojne analize naravnega jezika.

Predstavitev problema

Strojne oz. računalniške metode so v humanistiki, predvsem pa na tu obravnavanem področju etnologije, konkretno folkloristike, še danes redke. Eden izmed razlogov je sorazmerno velika nepovezanost področij kakor tudi razpoložljivost digitalnih virov. Rokopisne zbirke, terenski zapisi, dnevniki in na splošno večina starejšega gradiva so pogosto v obliki digitalnih slik in le izjemoma transkribirani, kar je predpogoj za strojno analizo besedil. Nadalje ima vsako področje svoje posebnosti, ki jih je treba upoštevati pri uporabi strojnih metod. Etnološke vsebine so za strojne metode dodaten izziv, saj je te treba prilagoditi tudi jezikovnim značilnostim. Tako je npr. ljudska pesem drugačna od formalizma umetne, saj so besedila ljudskih pesmi močno narečno in kontekstualno obarvana, njihovi motivi pa so pogosto lahko razumljivi le v specifičnem kulturnem, geografskem, zgodovinskem ali sociološkem kontekstu. Glavni raziskovalni problem strojnih metod za obdelavo naravnega jezika predstavlja izgradnja algoritmov za ekstrakcijo, analizo in interpretacijo ter konceptualizacijo kvalitativnih dimenzij, na podlagi katerih lahko gradimo semantične relacije. Strojne metode so učinkovite in primerne predvsem za analizo večje količine podatkov. Potrebno jih je razumeti kot komplementarne ročni analizi vsebin, saj z njimi lahko razširimo obseg analize in hkrati dobimo širši pogled v splošnejše značilnosti. Medtem ko ročna analiza večinoma poteka »vertikalno« in pogosto v njej odseva subjektivni vidik raziskovalca, sicer pa je njena prednost v natančnosti in raziskovanju/odkrivanju specifičnih značilnosti v omejenem izboru vsebin, nam po drugi strani strojne metode omogočijo »horizontalen« pregled obsežnejšega zbira vsebin ter ekstrakcijo in analizo splošnih značilnosti in pomenskih vzorcev, ki jih ni mogoče ročno pridobiti in analizirati (Abello idr. 2012). Na ta način raziskovalcu omogočimo pogled v širšo konceptualno zasnovo gradiva in odkrivanje novih semantičnih struktur in povezav, ki zaradi obsežnosti zbirk pogosto ostajajo skrite.
V angleškem jeziku se za uporabo strojnih metod na področju folkloristike večinoma uporablja izraz *Computational Folkloristics* (glej npr. prav tam), kar smo prevedli v *računalniško folkloristiko*. S folkloističnega vidika gre v prvi vrsti za odkrivanje pomena in semantičnih struktur v večjih korpusih besedil ljudskih pesmi, npr. odkrivanje medbesedilnosti in razmerij med variantnimi tipi, motivičnih in tematskih vzorcev, žanrske pripadnosti itn. Osnovni izziv je zajeti temeljne elemente ljudske pesmi in ustvariti formalno strukturo za sistematično računalniško analizo, ki jo lahko potem razširimo na širši vzorec primerov ali pa analizo sorodnih vsebin.
V nadaljevanju sta na kratko predstavljena različna pristopa strojne obdelave naravnega jezika (angl. *natural language processing* oz. NLP), statistični in verjetnostni pristop.

* Dr. Gregor Strle, univ. dipl. filozof, asistent z doktoratom, Glasbenonarodopisni inštitut ZRC SAZU. 1000 Ljubljana, Novi trg 5, gregor.strle@zrc-sazu.si; doc. dr. Matija Marolt, univ. dipl. inž. rač. in inf., docent, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani. 1000 Ljubljana, Tržaška cesta 25, matija.marolt@fri.uni-lj.si.

Metode

Latentna semantična analiza

Med metodami strojne obdelave naravnega jezika je najbolj razširjena *latentna semantična analiza* (angl. *latent semantic analysis* oz. LSA; Landauer in Dumais 1997; Landauer idr. 2006). LSA je statistična metoda, ki korpus besedil obravnava kot vrečo besed in *pomene* povzame na podlagi asociacij med besedami tako, da vzorce uporabe/pojavnosti besed analizira v več dokumentih. Relativne jakosti teh asociacij predstavi kot vektorje v visoko-dimenzionalnem prostoru.[1] Frekventnost pojavljanja besed je definirana v matrici besed in dokumentov. Pri ustvarjanju matrice je pomembna funkcija uteževanja, ki temelji na pogostnosti pojavljanja besed v odstavkih in je v obratnem sorazmerju s pojavljanjem besed v vseh dokumentih – s tem se izniči pomembnost visoko frekventnih izrazov, ki ne prispevajo bistveno k razlagi pomena (Martin in Berry 2006). Da bi ohranili le bistvene značilnosti, se izvede dekompozicija SVD (oz. *singular value decomposition*), ki dimenzionalnost originalne matrice zmanjša na vnaprej določeno število dimenzij.[2] Dobimo semantičen prostor oz. asociacijsko mrežo besed in globalno oceno podobnosti[3] med besedami, ki je generirana na podlagi pojavljivosti besede v korpusu. V tem semantičnem prostoru so pomensko podobne besede predstavljene kot vektorji tesno skupaj, nasprotno velja za pomensko nepovezane besede. Ker LSA ne uporablja le informacije o tem, kako pogosto se *beseda1* in *beseda2* pojavita skupaj, marveč tudi kako pogosto se pojavita z vsemi drugimi besedami v zbirki, lahko do neke mere zazna tudi sinonime (Landauer idr. 2006).[4]

Latenta Dirichletova razporeditev

Latenta Dirichletova razporeditev (angl. *latent Dirichlet allocation* oz. LDA; Blei idr. 2003) je generativni verjetnostni tematski model. Temelji na osnovni predpostavki tematskih modelov: *dokumenti* vsebujejo številne *teme*. Osrednje vprašanje LDA in drugih tematskih modelov je, kako izluščiti tematsko strukturo, skrito v teh dokumentih. LDA skuša prepletanje latentnih tem modelirati kot verjetnostno porazdelitev prek dokumentov in besed. V LDA je treba število tem opredeliti predhodno, nato LDA po dokumentih v korpusu izračuna skrito tematsko strukturo. Podatki so del generativnega procesa, ki definira *skupno*

A	Nəč predowga, nəč prekratka, sej ne bom plesala u nji.
B	Nič predolga, nič prekratka, sej ne bom plesala v nji.
C	nič predolg nič prekratek saj ne biti plesati v on

Slika 1: Učinek lematizacije. A prikazuje originalno besedilo, B besedilo po odstranitvi naglasnih glasov in uporabi naglasnega slovarja in C lematizirano besedilo.
Vir: lastni prikaz.

verjetnostno porazdelitev nad opazovanimi in skritimi naključnimi spremenljivkami, pri čemer so znane spremenljivke besede in vnaprej določeno število tem, skrita pa je tematska struktura. LDA uporabi *skupno porazdelitev* za izračun *posteriorne porazdelitve* skritih spremenljivk v danih dokumentih. Tako ima vsak dokument drugačno porazdelitev tem in vsaka beseda drugačno verjetnost, da pripada določeni temi.

> To je značilna lastnost latentne Dirichletove razporeditve – vsem dokumentom v zbirki je skupen isti nabor tem, hkrati pa vsak dokument kaže pripadnost tem temam v drugačnem razmerju (Blei 2012: 4).

Analiza

Z analizo smo želeli preveriti primernost uporabe statističnega in verjetnostnega pristopa pri raziskavah folklorističnih vsebin. Osredinili smo se na splošne značilnosti korpusa na ravni pesemskih tipov in tem, ki se v njih prepletajo. S tem smo razkrili konceptualno strukturo ljubezenskih in družinskih pripovednih pesmi in hkrati odkrili pomanjkljivosti preskušenih metod. Poudarek je bil na praktični uporabi osnovnih modelov LSA in LDA, da v primeru večjih razlik izberemo primernejši pristop za nadaljnje raziskave.

Korpus

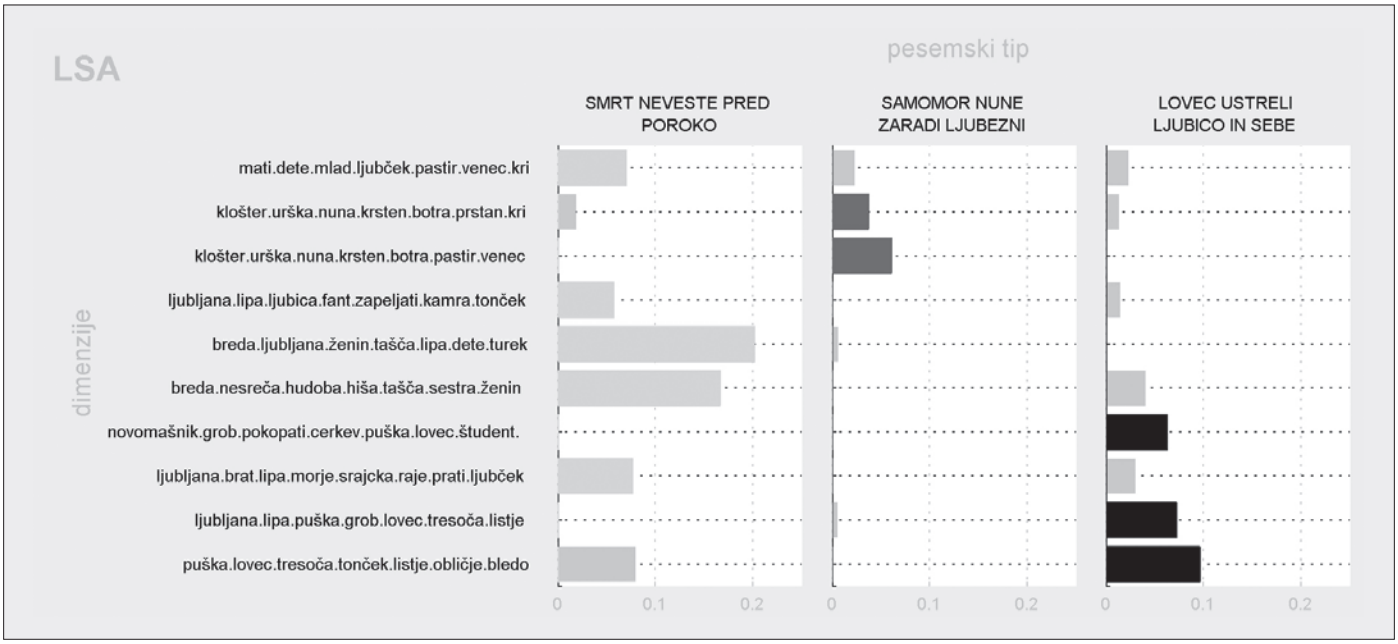
Metodi smo preskusili na zbirki 1965 variant ljudskih pripovednih pesmi iz knjig *Slovenske ljudske pesmi IV* in *V* (v nadaljevanju: SLP 4 in 5; Golež Kaučič idr. 1998, 2007). Izbor vsebuje 36 pesemskih tipov pripovednih pesmi o ljubezenskih usodah in konfliktih s 1073 variantami ter 54 tipov pripovednih pesmi o družinskih usodah in konfliktih z 857 variantami, ostalo so hrvaške in srbske pesmi na slovenskem ozemlju. Gre za tematsko zelo sorodno gradivo, kar je razvidno že iz naslovov tipov variant (glej SLP 4 in 5): *smrt*, *uboj*, *umor*, *samomor*, *nezvestoba*, *kazen* ipd. Nadalje je tudi tu opazna močna medbesedilnost, tj. potovanje verzov, kitic in tematskih vzorcev oz. motivov iz pesmi v pesem, kar je značilno za ljudsko pesem (glej npr. Kumer 1996;

1 Izrčrpna matematična predstavitev LSA v: Martin in Berry 2006.

2 Po mnenju raziskovalcev (prav tam) dobimo najboljše rezultate pri zmanjšanju dimenzionalnosti na približno 300 dimenzij, odvisno tudi od velikosti korpusa.

3 Mera podobnosti med besedami je opredeljena z vrednostmi: vrednosti proti 1 kažejo na visoko sorodnost, medtem ko nizke oz. negativne vrednosti kažejo na nepovezanost.

4 Poudariti je primerno, da so statistične metode učinkovite predvsem pri analizi zelo obsežnih korpusov. Zgled: Landauer in Dumais (1997) sta LSA preskusila pri preskusu razpoznavanja sinonimov, ki ga ameriške univerze uporabljajo za preskus znanja angleščine pri sprejemu tujih študentov (TOEFL – Test of English as a Foreign Language). Za podlago sta uporabila korpus TASA (Touchstone Applied Science Associates Inc.), ki ga sestavlja 92.409 besed iz 37.651 besedil, knjig, člankov in ostalega splošnega gradiva, ki ga ameriški študent pozna do 1. letnika univerze. LSA je test opravil z 64,4 % uspešnostjo, kar je primerljivo z rezultatom velikega vzorca študentov in zadovoljivo za sprejem na večino ameriških univerz.



Slika 2: Razporeditev LSA najpomembnejših besed in dokumentov prek prvih 10 dimenzij.
Vir: lastni prikaz.

Golež Kaučič 2003). Kakor bomo videli v nadaljevanju, se to močno izraža pri rezultatih strojne analize, kjer omenjene teme in motivi prevladujejo v večini pesemskih tipov.

Predpriprava gradiva

Ker je slovenščina sintetični jezik z veliko morfologije, pesmi pa so tudi močno narečno obarvane, je bilo pred analizo nujno izvesti lematizacijo besedila (spremembo v besedne leme), saj sicer oba modela ne bi znala dobro povezati besed v različnih oblikah. Lematizacijo smo izvedli v dveh korakih. Najprej smo posebne znake, ki se uporabljajo za označevanje narečnih glasov (npr. polglasniki, rezijanski zasopli vokali, primorski zveneči 'h' ipd.), zamenjali s slovničnimi ekvivalenti in nato s pomočjo narečnega slovarja pogosto uporabljene narečne izraze prevedli v slovnično obliko. V drugem koraku smo z uporabo oblikoslovnega označevalnika Obeliks (Grčar idr. 2012) besedila lematizirali. Obeliks s segmentacijskem in tokenizacijskim modulom najprej besedilo razdeli na stavke in besede, nato z oblikoslovnim označevalnikom besedam pripiše besedno vrsto in lastnosti ter jim na koncu z lematizatorjem pripiše njihovo osnovno obliko (npr. delam -> delati, mizama -> miza). Primer učinka lematizacije je ponazorjen na sliki 1.

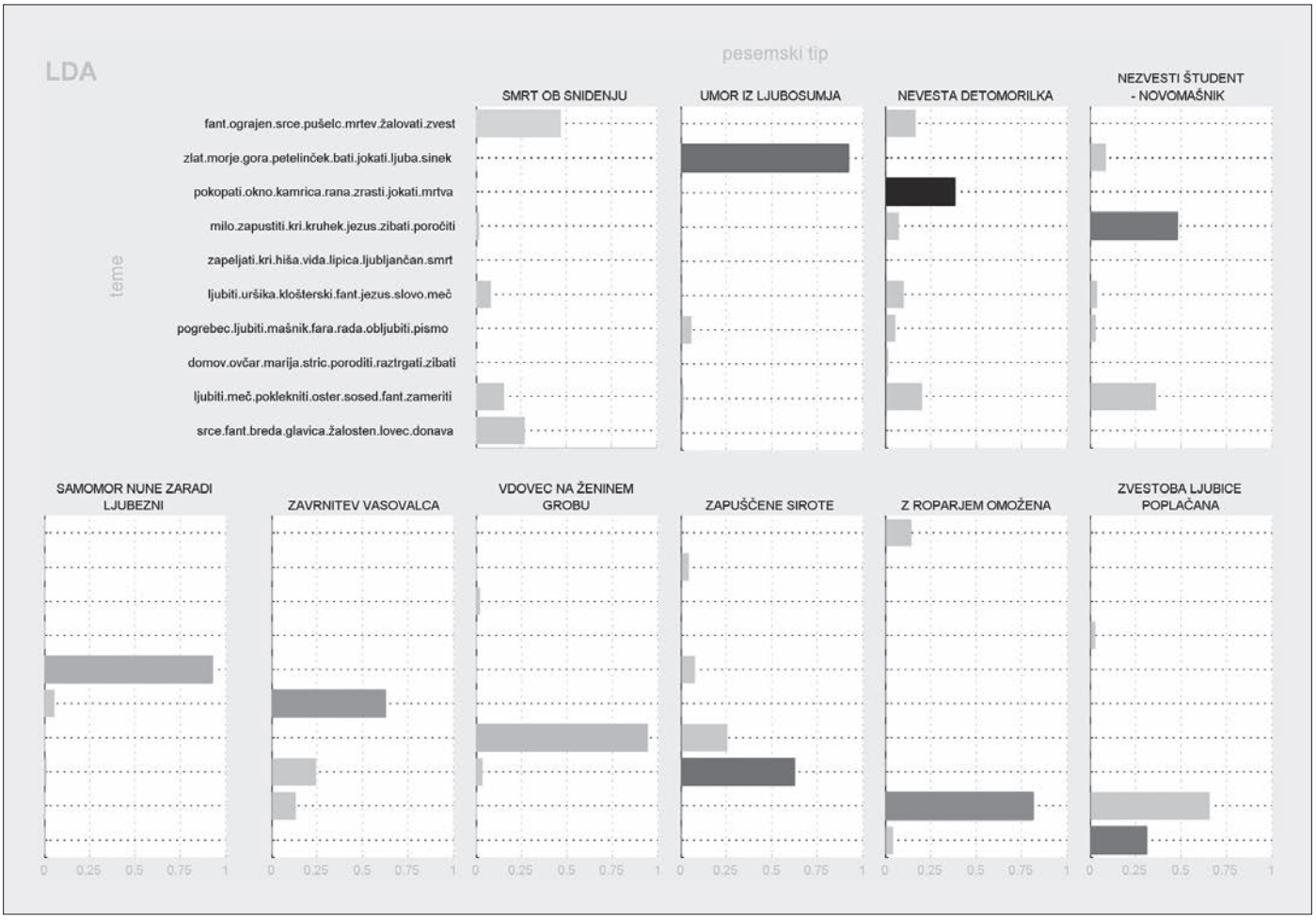
Rezultati

Analizo in vizualizacijo rezultatov smo izvedli s knjižnicami za LSA in LDA v okolju Matlab. Vektorje pojavitev besed v pesmih, ki so vhod v obe metodi, smo izračunali z mero TF-IDF (*term-frequecy, inverse document frequency*), ki logaritmično uteži pogostnosti pojavitev besed, hkrati pa manj poudari besede, ki se pojavljajo v veliko pesmih. Po analizi smo rezultate vizualizirali s projekcijo na mrežo SOM[5] (Kohonen 1995; glej sliki 4 in 5). Na slikah 2–5 so grafični prikazi rezultatov analize obeh metod. Projekcijo smo omejili na najpomembnejše besede in dokumente v prvih 10 dimenzijah oz. temah, generiranih z LSA in LDA.[6] Rezultati relativne podobnosti in pomembnosti besed v posamičnih pesemskih tipih so primerljivi pri obeh metodah, npr.: za pesemski tip *Samomor nune zaradi ljubezni* so značilne besede 'Urška', 'klošter', 'nunca', za tip *Nezvesti študent - novomašnik* besede 'novomašnik', 'študent', ‘zvestoba’, za tip *Lovec ustreli ljubico in sebe* besede 'lovec', 'puška', 'gozd', za tip Mačeha in sirota pa 'dete', ‘'mačeha', 'česati' itn.

Drugače je z razporeditvijo pesemskih tipov. Kakor je razvidno iz slik 2 in 3, LSA iz celotnega korpusa v prvih 10 dimenzijah generira le tri pesemske tipe, medtem ko LDA najde za vsako dimenzijo svoj pesemski tip. LSA kot statističen model upošteva zgolj pojavljanje besed v dokumentih in ne zazna latentnih tem, kar je očitna pomanjkljivost. Naslednja vzroka za tako slab rezultat LSA sta omejena tematska raznovrstnost korpusa in že omenjena močna medbesedilnost, kar se pokaže v razmeroma majhnem število motivov in tem, ki posledično prevladujejo v večini pesemskih tipov. Tako ima npr. temo 'smrt' (smrt, uboj, umor, samomor, detomor) že v naslovu kar 35 % pesemskih tipov, še višji pa je odstotek v pesmih. Na analizo dodatno vplivajo pogostejši tipi z več variantami, npr. zgodba o študentu novomašniku (tragična ljubezen) ima 150 različic (SLP 4), o nevesti detomorilki kar 188 (SLP 5), medtem ko pesem o galjotu (kaznjencu, ki odkloni vrnitev k družini; SLP 5) le dve. To posledično povzroči nesorazmerno porazdelitev in prekrivanje

5 Mreža SOM nam omogoči vizualizacijo visoko-dimenzijskih podatkov, da bi predstavili osnovno strukturo in razmerja med podatki. Gre za nekakšen zemljevid (nizkodimenzionalni pogled na visokodimenzionalen prostor), saj ohranja topološke lastnosti vhodnega prostora in s tem sosedskih razmerij med vozlišči.

6 S spreminjanjem števila dimenzij se spreminjajo tudi število, izbor in razporeditev tem, besed in pesemskih tipov.



Slika 3: Razporeditev LDA prvih 10 tem prek najpomembnejših besed in dokumentov.
Vir: lastni prikaz.

dimenzij pri LSA: SOM LSA nazorno kaže prekrivanje dimenzij 2/3, 4/5 in 6/9 (primerjaj SOM projekciji rezultatov LSA in LDA na slikah 4 in 5).

Sklep

V članku smo predstavili statistično (LSA) in verjetnostno (LDA) metodo semantične analize naravnega jezika. Namen je bil predvsem praktičen: izbrati učinkovit pristop za semantično analizo folklorističnih vsebin, ki bi ga lahko v prihodnosti nadgradili za zahtevnejše raziskave. S predstavljenima metodama smo skušali izluščiti splošne značilnosti, prevladujoče teme, motive in pesemske tipe v korpusih družinskih in ljubezenskih pripovednih pesmi ter na ta način ugotoviti prednosti in pomanjkljivosti obeh metod.

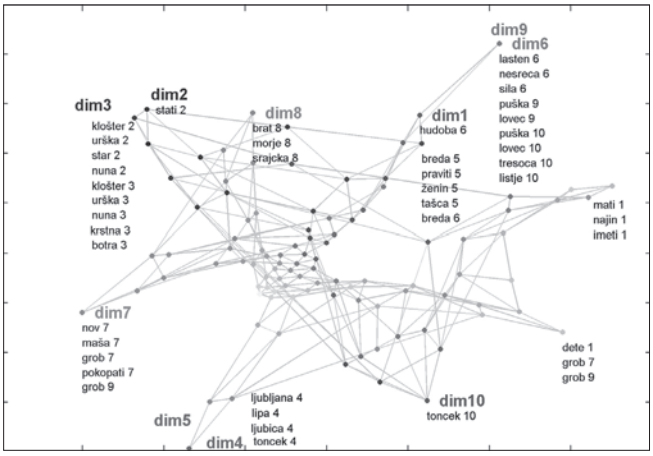
Rezultati so razkrili očitno pomanjkljivost statističnih modelov naravnega jezika (v našem primeru LSA), tj. nezmožnost zaznavanja, generaliziranja in porazdelitve prek latentnih tem. S statističnim pristopom izgubimo pomemben del informacij, kar pri tako koherentnem korpusu, kakršen je testni, vodi v osiromašeno semantično analizo. Po drugi strani pa nam LDA z verjetnostno porazdelitvijo omogoča ekstrakcijo latentnih tem in posledično zadostno semantično strukturo za zajem nekaterih kvalitativnih vidikov naravnega jezika, npr. sinonimijo, polisemijo in kontekst (glej npr. Griffiths idr. 2007). To, in pa zmožnost učenja oz. posploševanja na nove dokumente zunaj učne množice, je ena izmed bistvenih prednosti modela LDA pred statističnimi pristopi. V konkretnem primeru večje tipološke raznovrstnosti nismo zaznali, kar je bilo pričakovano glede na tematsko omejenost in medbesedilno prepletenost korpusa. Kot smo omenili v uvodu, nam strojne metode večinoma omogočajo splošen pregled in zato zahtevajo pogosto dopolnjevanje z ročno analizo. Prednost metod računalniške folklorisitke je predvsem v avtomatski analizi večje količine podatkov, ekstrakciji splošnega pomena in smiselnih vzorcev. Za podrobnejšo analizo, npr. raziskavo specifičnih (oz. manj zastopanih) tem in motivov, bi bilo treba upoštevati vzorce relevantnih delov besedila, jih predhodno ročno označiti in na njih uporabiti kako naprednejšo različico LDA. Izziv za prihodnost.

Literatura

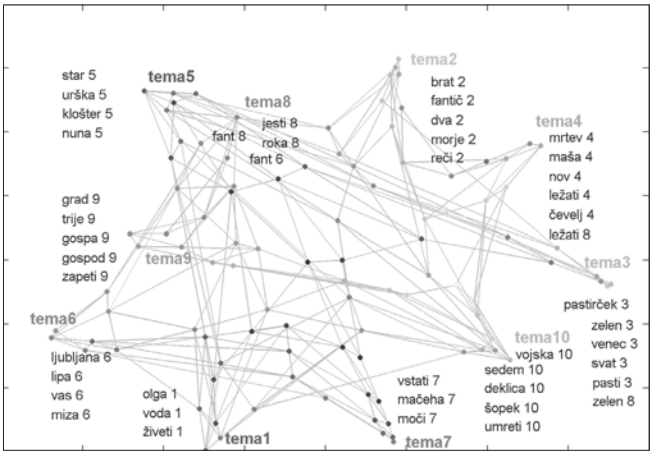
ABELLO, J. idr. Computational folkloristics. *Communications of the ACM* 55/7, 2012, 60–70.

BLEI, David M.: Probabilistic topic models. *Communications of the ACM* 55/4, 2012, 77–84.

BLEI, David M. idr.: Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3/4,5, 2003, 993–1022.



Slika 4: SOM projekcija rezultatov LSA.
Vir: lastni prikaz.



Slika 5: SOM projekcija rezultatov LDA.
Vir: lastni prikaz.

GOLEŽ KAUČIČ, Marjetka idr. (ur.): *Slovenske ljudske pesmi: Knj. 5: Pripovedne pesmi*. Ljubljana: Založba ZRC, ZRC SAZU, 2007.

GOLEŽ KAUČIČ, Marjetka: *Ljudsko in umetno: Dva obraza ustvarjalnosti*. Ljubljana: Založba ZRC, ZRC SAZU, 2003 (Zbirka Folkloristika).

GOLEŽ KAUČIČ, Marjetka idr. (ur.): *Slovenske ljudske pesmi: Knj. 4: Pripovedne pesmi*. Ljubljana: Slovenska matica, 1998.

GRČAR, Miha idr.: Obeliks: Statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V: Tomaž Erjavec in Jerneja Žganec Gros (ur.), *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan, 2012, 89–94.

GRIFFITHS, Thomas L. idr.: Topics in semantic representation. *Psychological Review* 114/2, 2007, 211–244.

KOHONEN, Timmo: *Self-Organizing Maps*. Series in Information Sciences 30, Berlin; New York: Springer, 1995, 362.

KUMER, Zmaga: *Vloga, zgradba, slog slovenske ljudske pesmi*. Ljubljana: Založba ZRC, 1996.

LANDAUER, Thomas K. in Susan T. Dumais: A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104, 1997, 211–240.

LANDAUER,Thomas K. idr. (ur.). *Handbook of latent semantic analysis*. Mahwah, N.J.: Erlbaum, 2006.

MARTIN, Dian I. in Michael W. Berry: Mathematical foundations behind latent semantic analysis. V: Thomas K. Landauer idr. (ur.): *Handbook of latent semantic analysis*. Mahwah, N.J. Erlbaum 2006, 35–55.

New Approaches: Uncovering Semantic Structures in Ethnological Materials

The paper explores computational approaches to semantic analysis of the natural language. The principal research problem of these methods is the construction of algorithms for the extraction, analysis, and interpretation as well as conceptualization of qualitative dimensions that will represent the basis for the construction of semantic relations. These methods are particularly effective in the analysis of extensive digital contents and, since they can extend the scope of research and give a better insight into more general characteristics of the materials, should be perceived as a complementary addition to manual analysis. The researcher is thus able to gain a wider view of the conceptual structure of the contents and discover semantic structures that often remain concealed because of the sheer extensiveness of the contents.
As an example of these methods, this text discusses an analysis of a corpus of Slovene folk songs. A comparison was made between two basic approaches of computational analysis: the statistical approach based on LSA (Latent Semantic Analysis) and the probabilistic approach on the basis of LDA (Latent Dirichlet Allocation). The principal objective of the study was to examine the suitability of both methods through an analysis of a body of songs that, in addition to their other linguistic peculiarities such as the dialect, for example, are highly intertextually colored, which presents an additional challenge for computational methods. The principal question of this analysis was to what extent these two methods can bring insight into the conceptual structure of the corpus and capture its general characteristics on the level of individual song types, and themes that are intertwined in them. Analysis results indicate an obvious disadvantage of the statistical method. While the two methods are comparable in the calculation of relative similarity between words there is a marked difference when it comes to the question of themes. As a statistical model, the LSA method takes into account only how many times a given word appears in a text but does not detect latent themes; this is apparent in a relatively small number of song types detected by the LSA although the entire corpus contains numerous song types. Research results indicate that the probabilistic method is more suitable for the analysis of a corpus with a strong intertextuality.