PAM Pokrajinski arhiv Maribor

Moderna arhivistika

Časopis arhivske teorije in prakse
Journal of Archival Theory and Practice

Letnik 4 (2021), št. 2 / Year 4 (2021), No. 2

Maribor, 2021

Pokrajinski arhiv Maribor

Moderna arhivistika

Časopis arhivske teorije in prakse
Journal of Archival Theory and Practice

Letnik 4 (2021), št. 2 / Year 4 (2021), No. 2

Maribor, 2021

VSEBINA

Prejeto / Received: 18. 08. 2021

1.02 Pregledni znanstveni članek *1.02 Review Article*

OBDELAVA PODATKOV V ARHIVSKIH PODATKOVNIH ZBIRKAH Z UPORABO NEKATERIH METOD ANALIZE VSEBINE

Dr. Tanja Martelanc
Pokrajinski arhiv v Novi Gorici, Slovenija
tanja.martelanc@pa-ng.si

Izvleček:

V članku so predstavljene različne metode analize vsebine, ki temeljijo na poznavanju umetne inteligence, procesiranju naravnega jezika in tekstualnega rudarjenja in bi lahko znatno pripomogle k hitrejšemu in bolj natančnemu popisovanju arhivskega gradiva, iskanju informacij in določevanju tematike arhivskih dokumentov ter tako posledično omogočile uporabnikom prijaznejši in efektivnejši način uporabe arhivskega gradiva. Članek temelji na nedavnih izsledkih tujih in slovenskih avtorjev, ki so predstavljene metode uporabili bodisi v arhivski stroki ali drugih humanističnih strokah, kot so npr. lingvistika in bibliotekarstvo. Tujejezične in domače literature na temo metod analize vsebine je izjemno veliko, dnevno se število člankov in prispevkov eksponentno povečuje, zato so v prispevku predstavljene le nekatere od metod, ki bi se po mnenju avtorice lahko uporabljale pri analizi vsebine arhivskega gradiva.

Ključne besede:

metoda analize vsebine, umetna inteligenca, modeliranje tem, arhivistika, tehnike poizvedovanja

Abstract:

Data Processing in Archival Databases Using Certain Methods of Content Analysis

The article presents different content analysis methods, which are based on the knowledge of artificial intelligence, processing of human language and text mining, and could substantially facilitate a faster and more detailed description of archival records, search for information and defining the theme of archival documents. Consequently, this would enable a kinder and more effective manner of archival records use for the users. The article is based on recent findings by home and foreign authors, who tested presented methods in either archival or other humanist fields, e.g. linguistics and library science. Home and foreign literature, dealing with content analysis methods is abundant, the number of articles grows exponently. Therefore, the author presents only those methods which could be used for the analysis of archival records content.

Key words:

content analysis method, artificial intelligence, topic modelling, archival science, search techniques

1. Uvod

Količina arhivskega gradiva se iz leta v leto eksponentno povečuje. Ker arhivisti zaradi kadrovske podhranjenosti ne bodo zmogli zadovoljivo popisati prevzetega gradiva, tako da bi uporabniku omogočali kar največ različnih načinov poizvedb oz. se bodo le-ti v množici podatkov, ki niso natančneje opredeljeni, "izgubili", bi bilo nujno

potrebno razmisliti o uvedbi in uporabi novih metod, ki bi pripomogle k hitrejšemu priklicu želenih informacij in avtomatskemu popisovanju ter kategorizaciji arhivskega gradiva.

Nove metode s področja umetne inteligence, ki se iz leta v leto izboljšujejo in nadgrajujejo, se množično uporabljajo predvsem v marketinški stroki, gospodarstvu, zdravstvu, varnostnih sistemih, sistemih nadzora ipd. (Allahyari et al., 2017). Na področju humanistike so metode najbolj poznane v lingvistiki in bibliotekarstvu; arhivska stroka se v zadnjem času zaveda prednosti uporabe umetne inteligence pri obdelovanju podatkov v arhivskih podatkovnih zbirkah in jo poskuša implementirati na svojem področju delovanja (Semlič Rajh, Šabotić in Šauperl, 2013, str. 125–144; Semlič Rajh in Šauperl 2013, str. 145–157). Pri tem v veliki meri uporablja nova spoznanja s področja analize vsebine.

2. Metoda analize vsebine

Metoda analize vsebine (angl. *Content Analysis*) je empirična raziskovalna metoda, s katero se obdeluje oz. išče in ugotavlja vzorce v besedilih, podobah, avdiovizualnih posnetkih ipd. Ti so bili ustvarjeni z namenom, da jih vidimo, preberemo, interpretiramo, se nanje odzovemo. Tako pridobljene rezultate je lažje obvladovati, saj je metoda sposobna preoblikovati velike količine podatkov v manjše vsebinske kategorije. Metoda analize vsebine je lahko kvantitativna (osredotoča se na štetje in merjenje) ali pa kvalitativna (osredotoča se na interpretacijo in razumevanje). V obeh primerih je potrebno kategorizirati oz. kodirati besede, teme, koncepte, ki se pojavljajo v dokumentih, in nato analizirati rezultate. Analiza vsebine se prednostno ukvarja z raziskovanjem pomena posredovane vsebine, saj odgovarja na vprašanja, kot so: kaj je pošiljatelj želel povedati z nekim besedilom/umetnino/filmom ipd., kaj to besedilo/umetnina/film dejansko pove sprejemniku, kako besedilo/umetnino/film sprejemnik razume itd. (Churchill, 2013, str. 256–257, 268; Know Your Audience, 2012).

Metoda analize vsebine je zanesljiva metoda, ki je ponovljiva in daje verodostojne rezultate. V preteklosti se je metoda analize vsebine izvajala ročno, njeni zametki so znani že iz 19. stoletja, danes se razvijajo matematični algoritmi, ki pomagajo pri avtomatskem analiziranju korpusa podatkov. Predvsem se s pomočjo analize vsebine lahko:

- identificira in opiše razvoj, vzorce in razlike, vidne v dokumentih,
- klasificira, kategorizira in vrednoti dokumente,
- povzema pomen dokumentov,
- raziskuje odnose med objekti v dokumentih in med dokumenti samimi ter odnose v kontekstu, v katerem so bili uporabljeni (Churchill, 2013, str. 255–256).

2.1 Metoda analize besedila

Ker se v arhivski stroki srečujemo predvsem z besedili, so v nadaljevanju predstavljene metode analize besedila, ki se v povezavi z razvijanjem umetne inteligence pridružujejo področju tehnologije naravnega jezika (angl. *Human Langugage Technology*). Ta zajema procesiranje naravnega jezika, prepoznavanje govora, strojno prevajanje, sintezo besedila in tekstovno rudarjenje (Brezovnik, 2009, str. 5).

2.1.1 Procesiranje naravnega jezika

Procesiranje naravnega jezika (angl. *Natural Language Processing* oz. NLP) ali računalniško jezikoslovje (angl. *Computational Linguistics*) se ukvarja z obdelavo nestrukturiranih besedil, zapisanih v naravnem jeziku. Glavni namen procesiranja naravnega jezika je človeku razumljive podatke pretvoriti v jezik, ki ga bo razumel tudi računalnik, tj. besedilo v naravnem jeziku pretvoriti v nekaj, kar bo strojno berljivo (Brezovnik, 2009, str. 5; Horvat, 2013, str. 1). Procesiranje naravnega jezika se množično uporablja pri predpripravi besedila za nadaljnjo obdelavo v okviru tekstovnega rudarjenja in zajema več opravil (povzeto po: Brezovnik, 2009, str. 5–11; Pavlinek, 2016, str. 11–18):

- razčlenjevanje (besedilo razdelimo na manjše enote, npr. posamezne besede),
- korenjenje oz. krnjenje besed (besedam odstranimo končnice),
- lematizacijo (besede nadomestimo z njihovimi lemami, tj. osnovnimi oblikami besed, ki so npr. zapisane v slovarju jezika),
- normalizacijo sinonimov (sinonim se zamenja z normalizirano obliko),
- označevanje besednih vrst (angl. *Part-of-Speech* oz. PoS; v stavku označimo besedno vrsto, npr. samostalnik, glagol, pridevnik itd.),
- določanje pomena besed (pomen besede ugotavljamo v nekem kontekstu),
- razreševanje sklicev (razrešimo besede, ki predstavljajo sklice na druge besede ali besedne zveze, npr. zaimke).

Poznamo nekaj orodij za procesiranje naravnega jezika v slovenščini, med njimi orodje Obeliks, ki omogoča tokenizacijo, oblikoskladenjski označevalnik, lematizacijo ali geslenje idr. (Horvat, 2013, str. 28).

2.1.2 Tekstovno rudarjenje

Tekstovno rudarjenje (angl. *Text Mining*) je raziskovalno področje, ki išče vzorce v večji količini nestrukturiranih tekstovnih podatkov v naravnem jeziku tako, da iz njih izlušči želene informacije. Tekstovno rudarjenje spada v t. i. podatkovno rudarjenje (angl. *Data Mining*). Tekstovno rudarjenje besedila ne poskuša razumeti, ampak v njem išče zgolj vzorce. Zajema pa (povzeto po: Brezovnik, 2009, str. 11–16; Allahyari et al., 2017; Likhitha, Harish in Keerthi Kumar, 2019, str. 1):

- iskanje informacij (angl. *Information Retrieval*), npr. s spletnimi in namiznimi iskalniki, pri katerih se ugotavlja stopnjo ustreznosti zapisov glede na iskalni niz, pri čemer so rezultati prikazani glede na stopnjo ustreznosti iskalnega niza od najbolj do najmanj ustreznega bodisi z upoštevanjem odnosov med posameznimi besedami ali brez,
- kategorizacijo oz. klasifikacijo besedil (angl. *Text Classification*), ko dokumente uvrščamo v eno ali več vnaprej znanih kategorij glede na njihovo vsebino, npr. naivni Bayes-ov klasifikator, frekvenca besed – inverzna frekvenca dokumentov ali FB-IFD (angl. *Term Frequency-Inverse Document Frequency* ali TF-IDF), latentna semantična analiza (angl. *Latent Semantic Analysis* ali LSA), latentno semantično indeksiranje (angl. *Latent Semantic Indexing* ali LSI), algoritem K-ti najbližji sosed, odločitvena drevesa idr. Razločujemo med nadziranimi metodami, ki za pravilno delovanje potrebujejo zunanjo informacijo, in nenadziranimi, ki delujejo popolnoma samostojno,

- razvrščanje v gruče (angl. *Text Clustering*), ko dokumente uvrščamo v eno ali več vnaprej neznanih kategorij glede na njihovo vsebino; besedila se v gruče razdeli glede na podobnosti,
- ekstrakcijo entitet ali konceptov (angl. *Named Entity Extraction*), s prvo odkrivamo entitete v besedilu, kot so npr. osebe, organizacije, lokacije ipd., z drugo pa koncepte, ki nastopajo v dokumentih,
- izdelavo povzetkov besedil (angl. *Text Summarization*).

Tekstovno rudarjenje se bolj ali manj uspešno uporablja tudi pri analizi t. i. *Big Data,* npr. zbirke govorov, zapisnikov raznih sestankov, člankov, blogov, elektronskih sporočil, spletnih strani, Twitter-ja, Facebook-a in drugih socialnih platform (več o tem v: Hassani et al., 2020, str. 1–34).

2.2 Modeliranje tem

Najbolj poznana metoda analize besedila je modeliranje tem (angl. *Topic Modeling*), ki pokriva tako področje obdelave naravnega jezika (NLP) kot tudi tekstovno rudarjenje. Modeliranje tem je skupek algoritmov, ki omogočajo analizo večje količine dokumentov (angl. *Bags of Words*) z namenom opredeliti njihovo tematiko (kategorizacija besedil); podlago za kategoriziranje išče v predpostavki, da se podobne besede pojavljajo v podobnih kontekstih oz. vzorcih, ki so sorodni določeni tematiki (Bail, *Topic Modeling*; Debenjak, 2019, str. 9).

Modeliranje tem je probabilistična statistična metoda, ki omogoča razumevanje skritih semantičnih struktur v nestrukturiranem besedilu oz. dokumentih (Dieng, Ruiz in Blei, 2019). Probabilistične metode so tiste metode, kjer ne obstaja nedvoumen odnos med vsebino zapisa in verjetnostjo, da bo zapis poiskan na dano iskalno zahtevo. Vzemimo npr. arhivsko podatkovno zbirko ScopeArchiv; v njej poizvedujemo po determinističnem postopku, se pravi, da poizvedujemo po relacijski zbirki podatkov, kjer so iskani podatki že vnaprej znani, konkretni. V tekstovnih zbirkah pa poizvedujemo po vsebini dokumentov; vsebine pa se ne da izraziti z enostavnimi vrednostmi, zato z metodami, kot so modeliranje tem, le predvidevamo stopnjo verjetnosti, da posamezen dokument ustreza iskani tematiki (Bail, *Topic Modeling*; van Hooland in Coeckelbergs, 2018, str. 79).

Modeliranje tem avtomatično razporeja, razume, išče in povzema podatke v ogromni količini besedil brez predhodnega učenja. Pri tem ne prešteva zgolj pogostosti pojavljanja neke besede v besedilu, ampak poskuša razumeti pomen in kontekst besedila in v njem uporabljenih besed. Končni produkt modeliranja tem je prikaz verjetnosti, kateri dokument ustreza iskani tematiki, npr. izbran dokument ustreza 30 % tematiki šolstva in 70 % tematiki urejanja javnega prometa, govori pa o ureditvi prometnega režima v okolici šol (Erčulj, 2019, str. 69).

Modeliranje tem je le ena od metod, ki se uporabljajo pri analizi vsebine besedil; poznani so namreč tudi drugi načini, kot npr. razvrščanje dokumentov v gruče oz. grozde, vendar pa slednji delujejo na nekoliko drugačen način. Kljub dobrim obetom uporabe metode modeliranja tem pri razvrščanju nestrukturiranega digitalnega gradiva pa je slaba lastnost te metode predvsem subjektivna interpretacija tem na podlagi nekaj ključnih besed, pri čemer se lahko skriti pomen "izgubi med vrsticami". Vrh vsega se besedilo ne dotika izključno le ene ali druge tematike, ampak se lahko teme v besedilu tudi prepletajo, kar vodi do dvoumnih rezultatov (van Hooland in Coeckelbergs, 2018, str. 79).

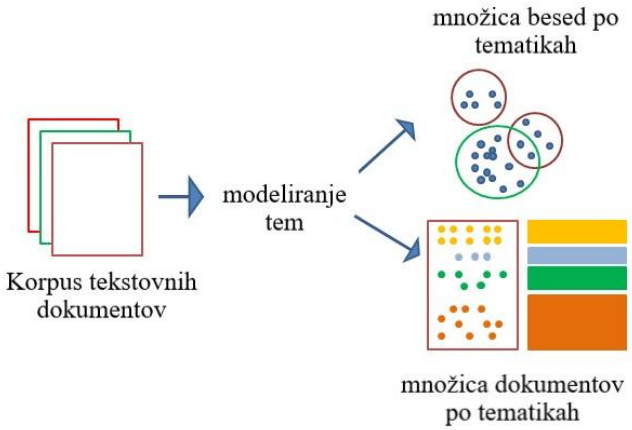
2.2.1 Latentna Dirichletova alokacija

Med različne pristope modeliranja tem uvrščamo verjetnostno latentno semantično analizo (angl. *Probabilistic Latent Semantic Analysis* ali PLSA), ki je statistično orodje za analizo sopojavitvenih podatkov, alokacijo Pachinko (angl. *Pachinko allocation*), hierarhično latentno drevesno analizo (angl. *Hierarchical latent tree analysis* ali HLTA) in najbolj priljubljeno in daleč največkrat uporabljeno latentno Dirichletovo alokacijo (angl. *Latent Dirichlet Allocation* ali LDA) (Hengchen et al., 2016, str. 3245–3249).

LDA je hierarhična probabilistična metoda, ki predstavlja vsako temo kot skupek ključnih besed in vsak dokument kot mešanico različnih tem (Dieng, Ruiz in Blei, 2019). Pri tej metodi moramo najprej opredeliti število tem, ki jim bodo dokumenti dodeljeni, kar je ena težjih nalog, ki ima hkrati tudi ključne posledice za rezultate analize. Vsaka beseda v besedilu je dodeljena eni od vnaprej predvidenih tem, hkrati je vsak dokument dodeljen eni ali več predvidenim temam (Dieng, Ruiz in Blei, 2019; Likhitha, Harish in Keerthi Kumar, 2019, str. 2). Vendar dokumenti temam niso dodeljeni naključno, ampak glede na prevladujoče besede v neki tematiki in glede na prevladujočo tematiko v nekem besedilu. Se pravi, da se z LDA poskuša "*oceniti verjetnost teme ob opazovanih besedah glede na dano porazdelitev besed po temah in glede na dano porazdelitev dokumentov po temah*" (Vidmar, 2010, str. 21). Teme so z uporabo LDA predstavljene z vektorji, vektorji pa so sestavljeni iz besed in njihovih uteži (Debenjak, 2019, str. 10.; Allahyari et al., 2017).

LDA ima dva rezultata:

1. identificira besede, ki se najpogosteje povezujejo s tematiko, ki jo je uporabnik predvidel,
2. dokument glede na verjetnost dodeli vsaki od tem, ki jih je uporabnik predvidel.



Slika 1: Primer vizualne predstavitve LDA (povzeto po: Naskar).

Na podlagi tega se lahko dokument dodeli eni ali več temam, ki jih je uporabnik predvidel. LDA za vsako temo predvidi seznam ključnih besed. Te ključne besede so reprezentativni simboli posamezne teme; povežemo jih lahko tudi z večjezičnim evropskih tezavrom EuroVoc. Hkrati ima LDA tudi to prednost, da razumljivo poda rezultate za neznani dokument, se pravi za dokument izven učne množice (Hengchen et al., 2016, str. 3245–3247). Kljub temu pa ima LDA tudi nekatere slabosti, saj v postopku obdelave besedil odstrani najbolj in najmanj pogoste besede in s tem pripomore k ne preveč natančnemu rezultatu, saj lahko "oklesti" pomembne dele besedila (Dieng, Ruiz in Blei, 2019).

V zadnjem času je postala priljubljena tudi metoda strukturiranega modeliranja tem (angl. *Structural Topic Model* ali STM), ki je precej podobna LDA, vendar v ozir jemlje še metapodatke o dokumentu, npr. avtorja, čas nastanka itd., z namenom izboljšanja rezultatov poizvedbe (Bail, *Topic Modeling*).

2.2.2 Modeliranje tem z uporabo besednih vložitev

Modeliranje tem z uporabo besednih vložitev (angl. *Embedded topic model* ali ETM) je še eden od načinov, ki je bil v tuji literaturi uporabljen pri klasificiranju arhivskih dokumentov. ETM je generativna probabilistična metoda, kar pomeni, da je vsak dokument mešanica tem, vsaka beseda je dodeljena dotični temi, moč relacije med besedo in temo ter dokumentom in temo pa je številčno izražena in vektorsko predstavljena (Dieng, Ruiz in Blei, 2019). Za razliko od enostavnega modeliranja tem je pri ETM beseda razumljena v kontekstu, v katerem je bila uporabljena. Ker mora, kot je bilo že zgoraj poudarjeno, pri modeliranju tem tematiko posameznega dokumenta uporabnik ugotoviti sam, kar je pravzaprav problematično, saj je teme na podlagi rezultatov včasih tudi težko interpretirati, so metodi modeliranja tem raziskovalci pritegnili še vektorsko besedno vložitev in jo na ta način izboljšali (van Hooland in Coeckelbergs, 2018, str. 79).

Vektorska besedna vložitev je model, ki z uporabo vektorjev predstavlja relacije med besedami in pomeni besed: besede s sorodnim pomenom so si npr. bližje. V nasprotju z modeliranjem tem, ki omogoča razumevanje odnosov med dokumenti na podlagi identificiranih tem, se vektorska besedna vložitev uporablja z namenom razumevanja semantične povezave med besedami pa tudi semantične povezave med različnimi temami v nekem dokumentu. Kot je razvidno iz imena, gre za vektorsko prezentacijo, ki pomaga razumeti razdaljo oz. bližino med besedami, semantično sorodstveno razmerje besed najdenih v isti temi, s tem pa omogoča lažje določanje tematike dokumenta. Z uporabo besedne vložitve so raziskovalci poenostavili metodo modeliranja tem, saj je besedna vložitev pomagala pri avtomatičnem zaznavanju različnih konceptov, ki se skrivajo v eni temi. Vrh vsega so jo preizkusili tudi na besedilih, iz katerih niso bile predhodno odstranjene manj pomembne besede (npr. predlogi, vezniki in členki), in ugotovili, da je dala uporabne rezultate (van Hooland in Coeckelbergs, 2018, str. 80–86; Dieng, Ruiz in Blei, 2019).[1]

[1] *Modeliranje tem z uporabo besednih vložitev uporablja programsko orodje Word2Vec, ki je eno najbolj priljubljenih programskih orodij na trgu za naravno procesiranje jezika (van Hooland in Coeckelbergs, 2018, str. 80; Škvorc, Robnik Šikonja, 2019, str. 110–114). Več o tej temi glej tudi: Esposito, Corazza in Cutugno, 2016.*

2.3 Razpoznavanje pojavnih oblik entitet

Razpoznavanje pojavnih oblik entitet (angl. *Named Entity Recognition* ali NER) je način tekstualnega rudarjenja za razpoznavo določenih informacij v besedilu, ki predstavljajo imenske identitete, npr. lastno ime osebe, organizacije, lokacijo, časovno opredelitev ipd. Pri tem se uporabljajo skriti markovski modeli (angl. *Hidden Markov Models*), pogojna naključna polja (angl. *Conditional Random Fields*) ali model Stanford NER, ki jih v praksi implementirajo z nadzorovanim učenjem na besedilu, kjer so identitete že vnaprej označene. Slabost te metode je, da včasih prihaja do nejasnosti glede razločevanja med nekaterimi entitetami (Nemec je npr. lahko priimek, hkrati pa tudi označuje prebivalca države Nemčije). Glavni namen razpoznavanja pojavnih oblik entitet je, da se iz nestrukturiranih ali polstrukturiranih oblik podatkov pridobi strukturirane podatke. Na podlagi pridobljenih strukturiranih podatkov pa so možne aplikacije drugih metod poizvedovanja po informacijah (Štajner, Erjavec in Krek 2013, str. 58–81; Allahyari et al., 2017). To metodo bi v arhivistiki lahko s pridom uporabili tudi pri procesu anonimizacije tajnih podatkov, davčnih skrivnosti in drugih občutljivih podatkov, ki jih varuje slovenska zakonodaja.

2.4 Vizualizacija rezultatov

Na podlagi predstavljenih metod analize vsebine se lahko pridobljene rezultate tudi vizualno prikaže.[2] Vizualizacija informacij poskuša preseči tipično grafično predstavitev podatkov s tem, da odkriva zakonitosti v podatkih. Posledično abstraktne podatke uporabniku sistema predstavi v njemu razumljivem jeziku. Možnosti vizualizacije so različne: 1D-, 2D- in 3D-tehnika, multidimenzionalna in časovna vizualizacija, drevesa, mreže in delovni prostor … (Merčun in Žumer, 2008, str. 97–106).[3]

V arhivistiki že uporabljamo drevesni diagram za prikaz strukture fondov in zbirk, lahko pa bi ga uporabljali tudi za prikaz družinskih dreves in sorodstvenih vezi med različnimi družinami/rodbinami, sistema organizacije javne uprave in sodstva v dotičnem časovnem obdobju in njihovih medsebojnih relacij, prikaz posameznih tematik dokumentov, ki bi bile hierarhično predstavljene, itd. Poleg te metode bi lahko s pridom uporabili še vsem znane tehnike vizualizacije podatkov, kot so stolpasti, paličasti, tortni kolači ali grafikoni, s pomočjo katerih bi lahko predstavili pojavnost različnih tematik v enem fondu/seriji/združenih dokumentih/dokumentih. Med zanimivejše oblike vizualizacije podatkov pa brez dvoma spadajo oblak besed, mrežna vizualizacija ali t. i. vizualni tezaver, Eulerjev diagram, Vennov diagram idr.

2.4.1 Oblak besed

Oblak besed vizualno prikaže besede tekstovnega korpusa tako, da so najbolj pogoste besede bodisi odebeljene in/ali večjih velikosti. Na ta način bi v arhivistiki lahko nazorno prikazali, katere besede se v dotičnem tekstovnem korpusu največkrat pojavljajo, npr. v nekem fondu ali seriji (WordItOut).

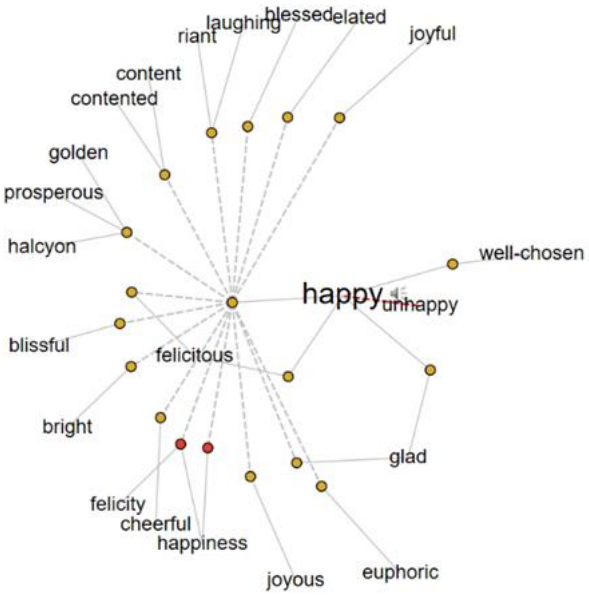
[2] *Več o tem v: Mladenić in Grobelnik, 2013, str. 27–33.*

[3] *Več o tezavrih glej: Developing a Functions Thesaurus, 2003.*

Slika 2: Oblak besed, ustvarjen z aplikacijo [WordItOut dne 10. 5. 2021.]

2.4.2 Vizualni tezaver

Vizualni tezaver je tezaver, pri katerem so relacije med pojmi oz. deskriptorji vizualno predstavljene, ali drugače povedano: vizualni tezaver je nazorno vidna zbirka deskriptorjev in odnosov med njimi. Pri jezikovnih vizualnih tezavrih je iskani pojem v središču zaslona, z njim povezani pojmi pa na periferiji. Odnosi z deskriptorji so različno opredeljeni, tako barvno kot tudi z uporabo različnih črt (sklenjena črta, črtkana črta itd.). Predstavljeni odnosi so lahko zelo kompleksni, abstraktni in večdimenzionalni, med seboj zapleteni, prepleteni itd. (Visual Thesaurus). Vizualni tezaver je lahko predstavljen v 2D- ali 3D-tehniki s pomočjo t. i. mrežne vizualizacije. (Merčun in Žumer, 2008, str. 108).

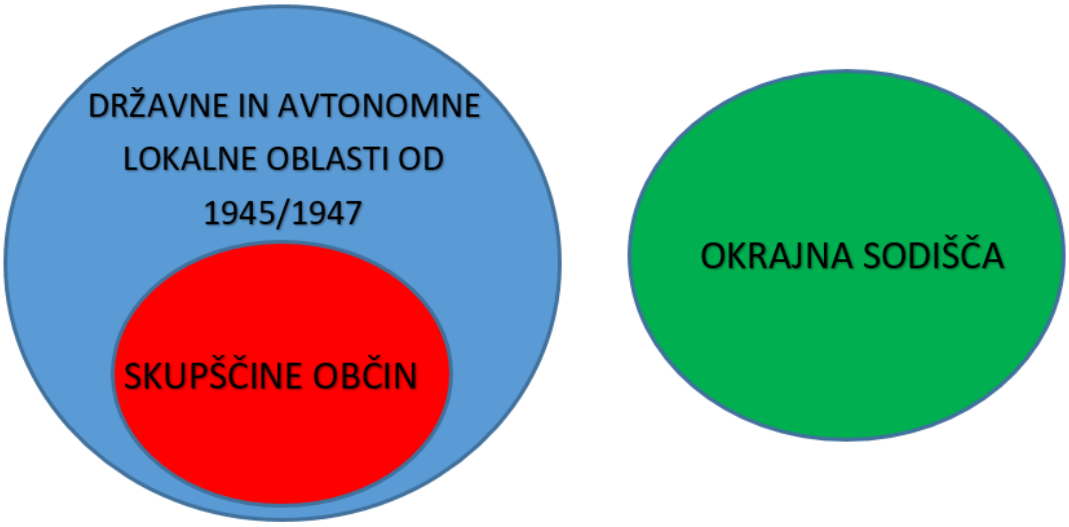


Slika 3: Primer uporabe vizualnega tezavra za besedo vesel (angl. happy) z aplikacijo VisualThesaurus, dne 10. 5. 2021.

Vizualizacija informacij ter vsebinske povezave med njimi postajajo vedno bolj pomembne tudi v sistemih poizvedovanja, ki se uporabljajo v arhivistiki. Trenutno je najbolj v uporabi linearen sistem poizvedovanja oz. poizvedovanje po drevesni strukturi arhiva, medtem ko so bile do nedavnega povezave in odnosi med dokumenti zapostavljeni. Preboj na tem področju predstavlja konceptualni model arhivskega popisovanja RIC (Records in Contexts), ki opredeljuje 14 entitet, ki so med seboj lahko povezane na skoraj 800 načinov. Gre za izredno kompleksno strukturo, s pomočjo katere se lahko prikaže odnose med samimi dokumenti, sestavnimi deli posameznih dokumentov, ustvarjalci itd.[4]

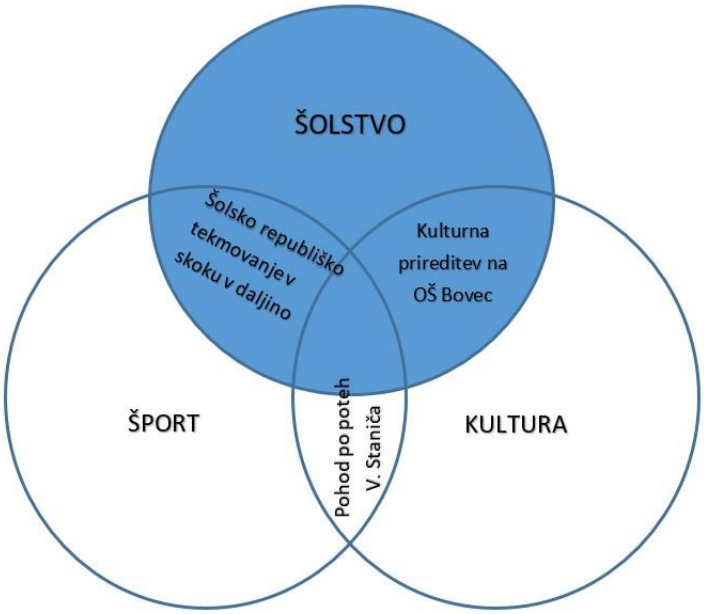
2.4.3 Eulerjev in Vennov diagram

Eulerjev diagram, ki naj bi ga razvil švicarski matematik, fizik in astronom Leonhard Euler v 18. stoletju, je zgrajen iz enostavnih zaprtih krivulj (običajno so to krogi), ki predstavljajo množice. Krogi se med seboj lahko prekrivajo v celoti (enaki elementi ali podmnožica), delno ali pa se sploh ne prekrivajo, ker nimajo skupnih elementov (Eulerjev diagram).



Slika 4: Primer Eulerjevega diagrama.

[4] *Več o tem v dokumentu: Records in Contexts.*



Slika 5: Primer Vennovega diagrama

Na podoben način deluje tudi Vennov diagram, ki se uporablja v teoriji množic, verjetnosti, logiki in računalništvu. Gre za grafični prikaz odnosa med množicami, ki ga je leta 1880 izumil John Venn. Razlika med Vennovimi in Eulerjevimi diagrami je ta, da morajo Vennovi diagrami vsebovati vsa možna področja, ki se prekrivajo (Vennov diagram).

S pomočjo Eulerjevega in Vennovega diagrama bi lahko v arhivistiki prikazovali rezultate poizvedovanja, ki so med seboj tesno povezani ali pa sploh niso povezani.

3. Zaključek

Uporaba novih metod s področja umetne inteligence procesiranja naravnega jezika in tekstovnega rudarjenja bi pripomogla k hitrejšemu in učinkovitejšemu poizvedovanju po arhivskih dokumentih in arhivskih podatkovnih zbirkah ter poenostavila in pohitrila izdelavo arhivskih pomagal. V ta namen bi nove metode lahko služile za:

- nazoren vsebinski opis arhivskega gradiva in relacij med njimi,
- enako velja za nazoren opis kontekstov arhivskega gradiva med njimi in
- nazoren prikaz relacij med opisi arhivskega gradiva in njihovimi konteksti.

Metode analize vsebine bi:

- omogočale hitrejši ponoven priklic zajetih vsebin iz arhivskih informacijskih sistemov,
- lahko bi se uporabljale kot iskalnik za ciljano raziskovalno zbirko podatkov,
- omogočale bi natančnejšo poizvedbo po arhivskih podatkovnih zbirkah, saj bi ta temeljila na entitetah, ki so med seboj vsebinsko, logično ali kako drugače povezane,
- rezultati poizvedovanja bi bili vsebinsko, tematsko itd. ustrezno predstavljeni,
- tak pristop bi omogočal poizvedovanje oz. pregledovanje zadetkov po različnih nivojih, zornih kotih, časovnih obdobjih itd.,
- s pomočjo interaktivnih funkcionalnosti bi se lahko uspešno reševalo mnoge arhivske strokovne naloge,
- omogočen bi bil interaktiven prikaz različnih vsebin tako besedila kot tudi fotografije, zvoka, video posnetkov itd.

Iskalniki, ki delujejo na podlagi vsebinskih kriterijev in odnosov med posameznimi entitetami, so bolj precizni kot navadni iskalniki. Uporabnik iskane pojme lahko vsebinsko in tematsko razporedi ter tako lažje opredeli semantične odnose med njimi. Poleg tega lahko tudi hitreje določi, kateri dokumenti so zanimivejši in kateri ne. Poizvedovanje po arhivski podatkovni zbirki bi z uporabo predstavljenih metod postalo tudi interaktivno in vizualno atraktivnejše. Jasno opredeljevanje vsebin in kontekstov ter povezav med njimi pa bo v prihodnosti predstavljajo osnovo obvladovanja velikih količin ohranjenega arhivskega gradiva tako v fizični kot tudi v elektronski obliki, zato je uporaba predstavljenih metod v arhivistiki dobrodošla.

4. Viri in literatura

Allahyari, M. et al (2017). *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques*. Pridobljeno 10. 5. 2021 s spletne strani: https://www.academia.edu/43750935/A_Brief_Survey_of_Text_Mining_Classification_Clustering_and_Extraction_Techniques.

Bail, C. *Topic Modeling*. Pridobljeno 10. 5. 2021 s spletne strani: https://cbail.github.io/SICSS_Topic_Modeling.html.

Brezovnik, J. (2009). *Programsko orodje za procesiranje besedil v naravnem jeziku*. Magistrsko delo. Maribor: Fakulteta za elektrotehniko, računalništvo in informatiko.

Churchill, B. (2013). Content Analysis. V.: Walter M. (ur.). V *Social Research Methods*, Melburne. Pridobljeno 10. 5. 2021 s spletne strani: https://www.academia.edu/5647773/Content_Analysis.

Debenjak, M. (2019). *Sledenje razvoju raziskovalnih tematik*. Diplomsko delo. Ljubljana: Fakulteta za računalništvo in informatiko.

Developing a Functions Thesaurus. Guidelines for Commonwealth Agencies (2003). Canberra: National Archives of Australia. Pridobljeno 10. 5. 2021 s spletne strani: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.93.4611&rep=rep1&type=pdf.

Dieng, A. B., Ruiz, F. J. R. in Blei, D. M. (2019). *Topic Modeling in Embedding Spaces*. Pridobljeno 10. 5. 2021 s spletne strani: https://arxiv.org/pdf/1907.04907.pdf.

Erčulj, V. I. (2019). *Analiza diskusij spletnih podpornih skupin z metodami strojnega učenja za namen pridobivanja informacij o psiholoških vidikih zdravljenja.* Doktorska disertacija. Ljubljana: Fakulteta za družbene vede.

Esposito F., Corazza, A. in Cutugno, F. (2016). Topic Modelling with Word Embeddings. *Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016: 5-6.* Napoli [online]. Torino: Accademia University Press. Pridobljeno 10. 5. 2021 s spletne strani: https://pdfs.semanticscholar.org/01b5/ea4cf2bbb20946b23841c2aa112816b0aa8d.pdf?_ga=2.15499894.1522787987.1601540317-1857418969.1601540317.

Eulerjev diagram. Pridobljeno 10. 5. 2021 s spletne strani: https://sl.wikipedia.org/wiki/Eulerjev_diagram.

Hassani, H. et al. (2020). Text Mining in Big Data Analytics. *Big Data and Cognitive Computing, 4 (1).* Pridobljeno 10. 5. 2021 s spletne strani: https://www.mdpi.com/2504-2289/4/1/1/htm.

Hengchen, S. et al. (2016). Exploring archives with probabilistic models: Topic Modelling for the valorisation of digitised archives of the European Commission. *Proceedings of the IEEE International Conference on Big Data*, str. 3245–3249. Pridobljeno 10. 5. 2021 s spletne strani: https://biblio.ugent.be/publication/8520997/file/8521049.

Horvat, M. (2013). *Orodja za tekstovno rudarjenje v slovenščini*. Diplomsko delo. Ljubljana: Fakulteta za računalništvo in informatiko.

Know Your Audience: chapter 16: Content analysis. Pridobljeno 10. 5. 2021 s spletne strani: http://www.audiencedialogue.net/kya16a.html :

Likhitha, S., Harish, B. S. in Keerthi Kumar, H. M. (2019). A Detailed Survey on Topic Modeling for Document and Short Text Data. *International Journal of Computer Applications (0975 – 8887), 178 (39).* Pridobljeno 10. 5. 2021 s spletne strani: https://www.researchgate.net/publication/335339697_A_Detailed_Survey_on_Topic_Modeling_for_Document_and_Short_Text_Data/link/5d5fb945a6fdccc32cc9ba1a/download.

Merčun, T. in Žumer, M. (2008). Vizualizacija informacij v sistemih za poizvedovanje. *Knjižnica : revija za področje bibliotekarstva in informacijske znanosti, 52 (2-3)*, str. 95-114. Ljubljana: Zveza bibliotekarskih društev Slovenije.

Mladenić, D. in Grobelnik, M. (2013). Automatic text analysis by artificial intelligence. *Informatica : an international journal of computing and informatics, 37 (1),* str. 27-33. Ljubljana: Informatika.

Naskar, A. *Latent Dirichlet Allocation for Beginners: A high level overview*. Pridobljeno 18. 8. 2021 s spletne strani: https://thinkinfi.com/latent-dirichlet-allocation-for-beginners-a-high-level-overview/.

Pavlinek, M. (2016). *Razvoj modela za inteligentno podporo odločanju na osnovi analize nestrukturiranih vsebin*. Doktorska disertacija. Maribor: Fakulteta za elektrotehniko, računalništvo in informatiko.

Records in Contexts a Conceptual Model for Archival Description. Pridobljeno 17. 9. 2017 s spletne strani: https://www.ica.org/sites/default/files/ric-cm-0.2_preview.pdf.

Semlič Rajh, Z. in Šauperl, A. (2013). Analiza oblikovanja vsebine zajetih podatkov v podatkovni bazi SIRAnet. V: Gostenčnik, N. (ur.). V: *Tehnični in vsebinski problemi klasičnega in elektronskega arhiviranja [Elektronski vir] : arhivi in ustvarjalci gradiva : stanje in perspektive : zbornik mednarodne konference, Radenci, 10.–12. april 2013 (2013),* URL: http://www.pokarh-mb.si/uploaded/datoteke/Radenci/Radenci2013/12_Semlic_Sauperl_2013.pdf.

Semlič Rajh, Z., Šabotić, I. in Šauperl, A. (2013). Znanstvenoraziskovalno delo v arhivistiki: značilnosti uporabe dveh raziskovalnih metod. V: Gostenčnik, N. (ur.). V *Tehnični in vsebinski problemi klasičnega in elektronskega arhiviranja [Elektronski vir] : arhivi in ustvarjalci gradiva : stanje in perspektive : zbornik mednarodne konference, Radenci, 10.–12. april 2013 (2013),* URL: http://www.pokarh-mb.si/uploaded/datoteke/Radenci/Radenci2013/11_Semlic_Sabotic_Sauperl_2013.pdf.

Škvorc, T., Robnik Šikonja, M. (2019). Prepoznavanje idiomatskih besednih zvez z uporabo besednih vložitev. *Uporabna informatika, 27 (3),* str. 110–114. Ljubljana : Slovensko društvo informatika.

Štajner, T., Erjavec, T. in Krek, S. (2013). Razpoznavanje imenskih entitet v slovenskem besedilu. *Jezikovne tehnologije [Elektronski vir].* 1 (2), str. 58–81. Ljubljana : Trojina, zavod za uporabno slovenistiko.

Van Hooland, S. in Coeckelbergs, M. (2018). Unsupervised Machine Learning for Archival Collections: Possibilities and limits of topic modeling and word embedding. *Revista catalana d'arxivística*, št. 41, str. 73–90. Pridobljeno 10. 5. 2021 s spletne strani: https://arxivers.com/wp-content/uploads/2018/10/1.4_-Dossier_SVHooland_MCoeckelbergs.pdf :

Vennov diagram. Pridobljeno 10. 5. 2021 s spletne strani: https://sl.wikipedia.org/wiki/Vennov_diagram.

Vidmar, K. (2010). *Vizualizacija konceptualnega prostora besedilnih zbirk*. Diplomsko delo. Ljubljana: Fakulteta za računalništvo in informatiko.

VisualThesaurus. Pridobljeno 17. 9. 2017 s spletne strani: https://www.visualthesaurus.com/.

WortItOut. Pridobljeno 10. 5. 2021 s spletne strani:: https://worditout.com/.

SUMMARY

DATA PROCESSING IN ARCHIVAL DATABASES USING CERTAIN METHODS OF CONTENT ANALYSIS

Tanja Martelanc, Ph. D.
Regional Archives in Nova Gorica, Slovenia
tanja.martelanc@pa-ng.si

The amount of archival records grows every year. Due to staff shortage, archivist will not be able to describe those records in a way to provide users with sufficient data for various manners of searching. In the multitude of not sufficiently defined data, users will „get lost“. Therefore, it would be useful to think about implementing new methods, which would facilitate the use of archival records.

Methods from the field of artificial intelligence improve and update constantly. They are mostly used in marketing, economy, healthcare, security systems etc. In human science they are used in linguistics and library science. Lately, the international and Slovene archival profession recognized the advantages of using artificial intelligence when processing data in archival databases and began with the use of the empiric research method of content analysis, which processes and finds patterns in texts, images, audio-visual recordings etc. Large amounts of data can be formed into smaller content categories, which are easier to manage. At the same time, this method provides for a faster and more efficient search through archival records and databases and facilitates the creation of archival finding aids.

In archival profession, we mostly deal with text documents, therefore, the article presents content analysis methods, which, connected to the development of artificial intelligence, near the field of human language technology. The most known method of content analysis is topic modelling, which covers the field of human language processing and text mining. Topic modelling is a cluster of algorithms, which enable an analysis of large amounts of documents (Bags of Words) with the goal to define their theme (text categorisation). Topic modelling is a probabilistic statistical method which enables the understanding of hidden semantic structures in an unstructured text or document. With the use of such methods we foresee the level of probability that a certain document corresponds to the searched for theme. Topic modelling automatically arranges, understands, searches and summarizes data in a large amount of text without prior learning. It not only counts the occurrence of a certain word in the text, but also tries to understand the meaning and context of the text and words used in it.

The most popular and widely used among topic modelling approaches are the Latent Dirichlet Allocation or LDA, Embedded topic model or ETM, Vector Word Embedding, Named Entity Recognition or NER, etc.

Presented content analysis methods can be used also to visualize results, e.g. word clouds, visual thesaurus, Euler diagram, Venn diagram, etc.

The use of new methods from the field of artificial intelligence of processing human language and text mining would provide for a faster and more efficient search through archival records and databases and facilitates the creation of archival finding aids. Search engines, which operate on the basis of content criteria and relations between individual entities, are more precise than regular search engines. Using presented methods would make the search though the archival database interactive. A clear

defining of content and context and their correlations will present the basis of managing large amounts of archival records in physical and digital form in the future.