Uporabna vrednost podatkov spletnih zajemov: arhiviranje spletnih mest in analiza spletnih vsebin

The practical value of web capture data: archiving Web sites and Web content analysis

Matjaž Kragelj, Mitja Kovačič

Oddano: 29. 3. 2017 – Sprejeto: 5. 6. 2017

1.04 Strokovni članek
1.04 Professional article
UDK 005.921.1-022.324:004.738

Izvleček

Zakon o obveznem izvodu publikacij (2006) Narodni in univerzitetni knjižnici (NUK) nalaga skrb za zajem, ohranjanje in nudenje dostopa uporabnikom do zajetih spletnih publikacij, spletnih mest in vsebin. Leta 2015 je NUK opravil prvi zajem slovenske domene .si, naslove spletnih domen je priskrbel Arnes (Akademska in raziskovalna mreža Slovenije). V prispevku se osredotočamo na pomen zajema spletnih vsebin zaradi vsakodnevnega propadanja spletnih domen. Poleg zajema in dejavnosti za zagotavljanje ohranjanja zajetih vsebin je v prispevku tematizirano tudi pridobivanje informacij iz nestrukturiranih vsebin (spletnih dokumentov). Omenjeni so primeri in delovanje aplikacij za zajemanje specifičnih informacij iz različnih spletnih dokumentov, npr. zajem cene določenega artikla v določeni trgovini z namenom obveščanja končnega uporabnika o najugodnejši ponudbi na trgu. Večji del prispevka je namenjen analizi zajetih spletnih vsebin in možnosti luščenja ter uteževanja besedišča, pridobljenega iz spletnih dokumentov. Z algoritmi in statistikami za označevanje in razvrščanje terminov v množici spletnih vsebin se spletni arhiv iz pasivne podatkovne zbirke spremeni v okolje, ki omogoča dodano vrednost povezovanja podatkov, iskanja sorodnosti znotraj podatkov spletnega arhiva in s podatki zunaj njega.

Ključne besede: *spletni arhivi, frekvenca pojavljanja, tf-idf, luščenje podatkov, spletni zajemi, domena .si*

Abstract

The Legal Deposit Act imposes to the National and University Library the concern and rights for capturing, preserving and providing access to online publications, web sites, and other content to library users. In 2015, the Library started the first capture of Slovenian .si internet domain. The domain addresses were provided by ARNES (the Academic and Research Network of Slovenia). The article focuses on the importance of covering the web content due to the deterioration of daily web domains. In addition to covering and activities to ensure the conservation of web contents, the paper also covers the subject of how to obtain information from unstructured content (documents on the web). The article shows some examples and applications to capture specific information from a variety of online documents (scraping), like the price of a selected item in a particular web store in order to inform the end user about the best offer on the market. The major part of the article is devoted to the analysis of captured web content and the possibility of scaling and ranking the vocabulary derived from web documents. The algorithms and statistics for marking and document ranking in a mass of web content can help transform the web archive from a passive database to the environment that creates the added value of data integration, finding similarities within a web archive data and the data from the outside of a web archive.

Keywords: *web archives, term frequency – inverse document frequency, data scraping, web harvesting, .si domain*

1 Spletne vsebine in zajemanje

Po mnenju Dramowicza (2016) so spletna mesta najpomembnejši repozitoriji informacij. Informacije, ki so nam tam na voljo, so temeljne za številne aktivnosti, kot so: medsebojna komunikacija, poslovanje, raziskovanje, zabava itd. Kopica brezplačnih informacij je zbranih na nepregledni množici spletnih mest, v različnih formatih in po navadi v slabo definiranih strukturah (Dramowicz, 2016). Vsak dan nastane in ugasne nepredstavljivo veliko spletnih strani in mest. Po eni izmed spletnih raziskav (February 2016 web server survey, 2017), opravljeni 29. 2. 2016, je živih več kot 996 milijonov spletnih mest. V času pisanja tega članka (marec 2017) pa je zaslediti podatek o tem, da je trenutno aktivnih več kot milijarda in sto šestdeset milijonov spletnih mest. Pri tem je treba upoštevati dejstvo, da je od tega zgolj četrtina spletnih mest aktivnih, ostala so zgolj rezervirana – registrirane domene in podobno (How many active sites, B. l.).

Ključen izziv institucij, katerih poslanstvo predstavlja arhiviranje spletnih mest, je bil najti format zapisa, ki bi v isti datoteki čim preprosteje združeval nabor različnih informacij (metapodatki, podatki). Datoteka naj bi vsebovala tako tekstovne kot arhivske datoteke, binarne, komprimirane itd. Leta 2005 se je pričel razvoj datotečnega formata WARC (Web ARChive), ki je leta 2009 pridobil certifikat ISO kot standard za shranjevanje spletnih vsebin (ISO 28500:2009).

Poskus zajemanja in upravljanja velikanskih količin podatkov, ki so na spletu na voljo, je (še vedno) resen izziv za organizacije, ki se s tem ukvarjajo. Zajemanje spletnih mest je metoda, s katero, podobno kot to počne uporabnik z uporabo spletnega brskalnika, program zajema spletno vsebino. Razlika je v tem, da strojno zajemanje poteka veliko hitreje in se zanj navadno uporabljajo programi oziroma metode, ki poskrbijo za ponovno reprezentacijo podatkov v izvorni obliki. Med najbolj znane aplikacije oziroma orodja za zajem spletnih vsebin sodijo Wget (GNU Wget, 2017), Heritrix (Jack, 2014), HTTrack (2017) in druga. Med spletnimi aplikacijami, ki jih lahko uporabimo neposredno, v živo, pa omenimo WebCite (B. l.), Wayback Machine (Internet archive wayback machine, 2014) in Archive-it (2014).

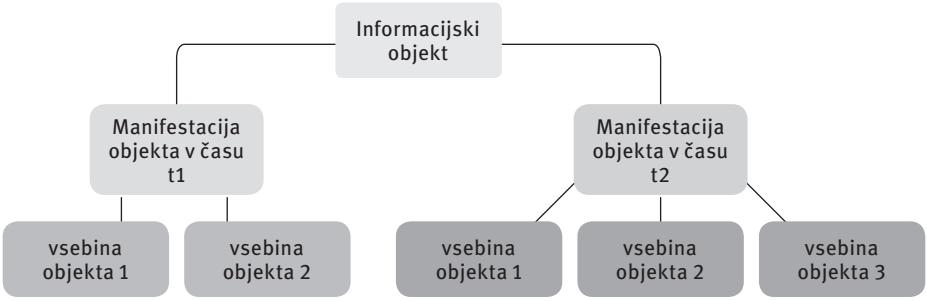
V prispevku se bomo osredotočili na pomen in dodano vrednost arhiviranja publikacij, objavljenih na spletu. Princip delovanja lahko po analogiji prenesemo na pridobivanje drugih virov podatkov. Poudariti želimo namen ohranjanja podatkov ter (pol)avtomatskega pridobivanja informacij iz besedil, ki jih imamo na voljo.

2 Manifestacije spletnih objektov in strategije ohranjanja

Pri arhiviranju vsebin na spletu gre za sistematično zajemanje in arhiviranje določenih delov svetovnega spleta z namenom ohranjanja čim natančnejše vsebine. Vsebine na spletu se neprestano spreminjajo, zato je treba zajeme posameznega spletnega mesta izvajati kontinuirano, s čimer lahko zagotovimo zajem podatkov, ki vsak dan nastajajo in tudi izginjajo iz svetovnega spleta. Frekvenca zajemov je odvisna od pogostosti spreminjanja spletnega mesta, njegove velikosti in od kapacitete strojne opreme, ki jo imamo na voljo. Z večkratnim zajemanjem neke spletne vsebine zmanjšamo možnost izgube podatkov oziroma njihove pristnosti.

Pri večkratnem zajemanju spletnih vsebin se moramo zavedati, da se z večkratnim zajemanjem količina podatkov veča, spreminja pa se (oziroma navadno dodaja) vsebina zajetega gradiva. Podatki, ki so na voljo na spletu v času $t = 0$, so že v naslednjem trenutku $t = 1$ lahko precej drugačni. Že najmanjša akcija upravljavca spletnega mesta lahko povzroči precejšnjo spremembo podobe spletnega mesta. Vsak posamezni zajem je označen s časovnim žigom; ta označuje določeno količino in vsebino gradiva, ki ga na nekem spletnem mestu v nekem trenutku lahko pridobimo oziroma zajamemo. Na Sliki 1 vidimo, da se tako vsebina kot količina podatkov na nekem spletnem mestu skozi čas lahko precej spreminjata. Informacijski objekt, ki ga zajemamo z namenom ohranjanja vsebine, bo imel lahko v drugem, lahko še tako kratkem časovnem intervalu, ko smo ga poskusili

ponovno zajeti enako, spremenjeno, ali pa – povsem drugačno vsebino. Še tako redno in natančno zajemanje vsebine nekega spletnega mesta ni zagotovilo za popolno ohranitev vsebine, poleg tega pa lahko pride do izgube podatkov zaradi drugih vzrokov.



Slika 1: *Manifestacija informacijskega objekta v različnih časovnih obdobjih*

Kljub večkratnim zajemanjem neke spletne vsebine lahko pride do izgube (dela) podatkov. Po Brownu (2013) manipulacija podatkov lahko nastopi zaradi:
- nesreče, zlonamerne aktivnosti;
- poškodbe ali propada medija, ki hrani arhiv;
- napake pri zapisovanju (npr. vpliv kozmičnih žarkov);
- uničenja medija (npr. naravne nesreče);
- poškodbe medija;
- napak na strežniški ali mrežni opremi;
- napake na programski opremi;
- napake pri replikaciji;
- zastaranja strojne in/ali programske opreme;
- slabega vodenja revizije dostopov;
- izgube podatkov, šifrantov ali slovarjev za interpretacijo podatkov ali zaradi kulturne spremembe (nezmožnost interpretacije podatkov).

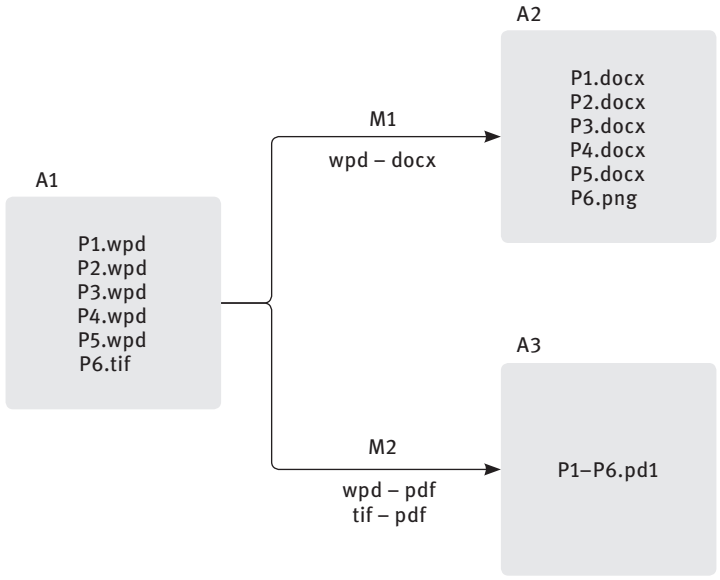
Ob izvajanju procesa zajemanja in arhiviranja spletnih vsebin je treba upoštevati tudi, da se skozi čas spreminjata format zapisa podatkov in uporabnikova zmožnost njihove interpretacije. Tako naletimo na težavo, ko uporabnik nima (več) nameščene tehnologije, aplikacije za vpogled v vsebino, ustvarjeno in objavljeno na spletu pred desetletji.

Po Brownu (2013) je na voljo več rešitev, in sicer:
- računalniški muzej,
- emulacija,
- transformacija.

Cilj **računalniškega muzeja** je uporaba avtentične strojne in programske opreme za dostop do informacij v primerni obliki. Uporaba je nepraktična, ne dopušča »uporabe od kjerkoli, kadarkoli in kolikokrat«, predvsem pa je predraga za široko oziroma pogosto uporabo.

Emulacija je princip delovanja programske opreme, kjer z vzvodi prilagoditve poskusimo vzpostaviti pogoje, ki bi uporabniku omogočili dostop do informacij in njihovo uporabo tudi prek oddaljenih sistemov, in to časovno in prostorsko poljubno ter z možnostjo večkratne uporabe. Slaba stran emulacije je dolgotrajnost razvoja potrebnih aplikacij na trenutni strojni opremi, poleg tega pa upravljanje s tako aplikacijo (npr. Word Perfect) od uporabnika zahteva znanje.

Transformacija oziroma migracija je princip, pri katerem na upravljavski strani poskrbimo za prenos vsebine v zapis, ki bo uporabniku na voljo vedno in povsod, ne da bi mu bilo treba za dostopanje do vsebin pridobiti nova znanja ali veščine.



Slika 2*: Migracija v novo manifestacijo*

Na Sliki 2 prikazujemo primer transformacije oziroma migracije. Ta poteka tako, da iz ene verzije objekta (A1) ustvarimo dva druga (A2, A3). Kot vhodni vir podatkov uporabimo že izdelano množico več Word Perfect poglavij (vsako v svoji datoteki, na koncu kot element dodamo še logotip dokumenta; A1) in jih, zaradi zagotavljanja možnosti uporabe v današnjem času, pretvorimo v dveh migracijah (M1, M2) v dve novi manifestaciji (A2 in A3). Eno manifestacijo obdržimo za

arhiviranje in morebitne kasnejše transformacije v nov zapis (A2), rezultat druge pa je objekt, namenjen uporabniku – v našem primeru je to pdf. datoteka (A3).

Primer nakazuje, da uporabniku ni treba vedeti ničesar o tehniki in postopkih transformacije in migracije ter formatu zapisa vhodnih ali izhodnih podatkov. Upravljavec mora poskrbeti, da s postopki, ki so na voljo, in z upoštevanjem pravil in arhivskih formatov vhodne podatke transformira v izhodne. Te pripravi na način, da jih je mogoče v prihodnosti na novo transformirati, pri čemer mora biti transformacija neizgubna (v primeru sprememb pravil, dobrih praks, standardov zapisa), hkrati pa mora uporabniku ponuditi verzijo zapisa na način, da bo informacija ostala popolna in nespremenjena, vendar transformacija ne bo zahtevala novega znanja za uporabo. Praktičen primer je nudenje zapisa uporabniku v pdf obliki, medtem ko upravljavec hrani podatek v obliki, ki trenutno predstavlja standard za določeno vrsto zapisa.

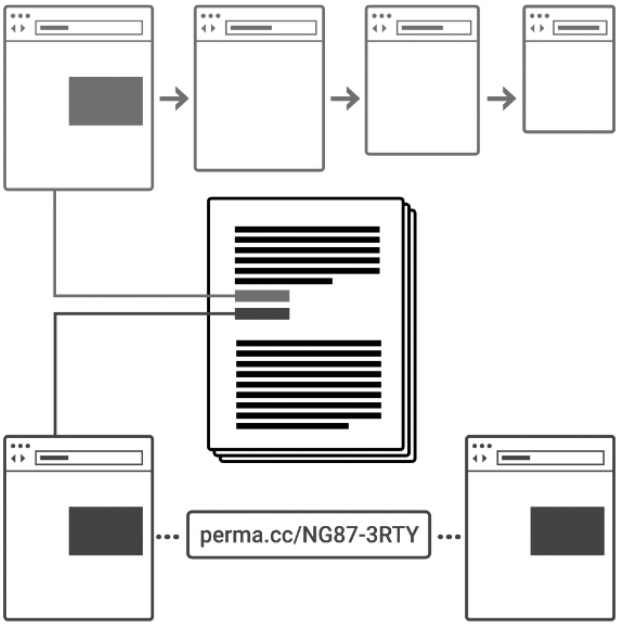
3 Ohranjanje spletnih objektov za potrebe citiranja in navajanja virov

Zakaj je nujno arhiviranje spletnih vsebin? Potrebo po arhiviranju pokaže dejstvo, da več kot 50 % navedenih povezav v mnenjih Vrhovnega sodišča (v ZDA) ne kaže več na citirano spletno stran ali mesto. Približno 70 % navedenih povezav v akademskih pravnih revijah in 20 % vseh navedb v člankih znanosti, tehnologije in medicine je podvrženih istim težavam (Perma, 2013).

Ena od rešitev za ohranjanje citirane spletne strani ali vsebine, če nimamo časa ali nujnih sistemskih virov za arhiviranje na način, opisan v poglavju 2, je prosto dostopna spletna aplikacija Perma.[1] Ta namesto nas shrani in arhivira spletno stran ter nam ponudi trajno povezavo do arhiviranega vira. Namesto citiranja spletnega mesta, ki smo ga uporabili v dokumentu, tako navajamo arhivirano stran, za katero nam upravljavci jamčijo, da bo vedno na voljo.

Primer na Sliki 3 prikazuje, kako sčasoma s spletne strani izgine podatek (kvadrat), medtem ko se v spodnjem primeru podatek ohrani, se pa spremeni povezava do njega.

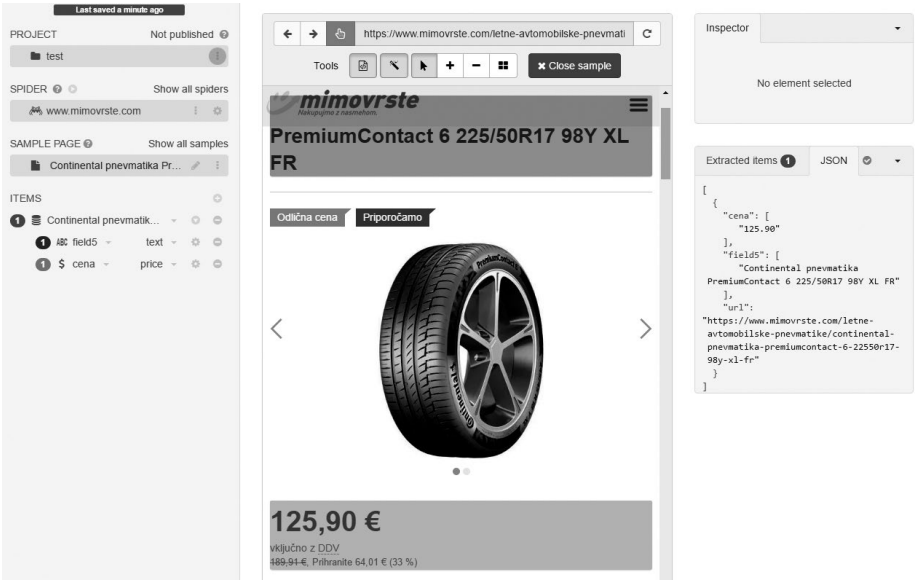
[1] Perma: https://perma.cc.



Slika 3: *Princip delovanja sistema »Perma«*

4 Luščenje podatkov s spletnih strani

V nasprotju s sistematičnim zajemanjem spletnih vsebin za potrebe ohranjanja celotne vsebine pri luščenju z (arhiviranega) spletnega mesta ciljno pridobivamo zgolj določene podatke oziroma specifične informacije na spletni strani. Na spletu je na voljo več brezplačnih rešitev, ki nam to omogočajo, npr. Import.io (2017), Dexi.io (2012), Scrapinghub (2010), Parsehub (2017) in druge. Metoda je uporabna in priljubljena predvsem pri tistih, ki z namenom trgovanja in prodaje zasledujejo npr. ceno izdelkov pri konkurenci. Omenjeni programi dovoljujejo nastavitev objektov na tarčah (npr. cena artikla »Continental pnevmatika PremiumContact 6 225/50R17 98Y XL FR« na eni izmed spletnih trgovin). Z avtomatizacijo postopkov imamo tako v vsakem trenutku vpogled v delovanje konkurence in po potrebi prilagajamo ceno izdelka v lastni spletni trgovini. Slika 4 prikazuje, kako smo z uporabo programa Scrapinghub izvedli zajemanje specifične informacije (cena in naziv izdelka) na neki spletni trgovini.

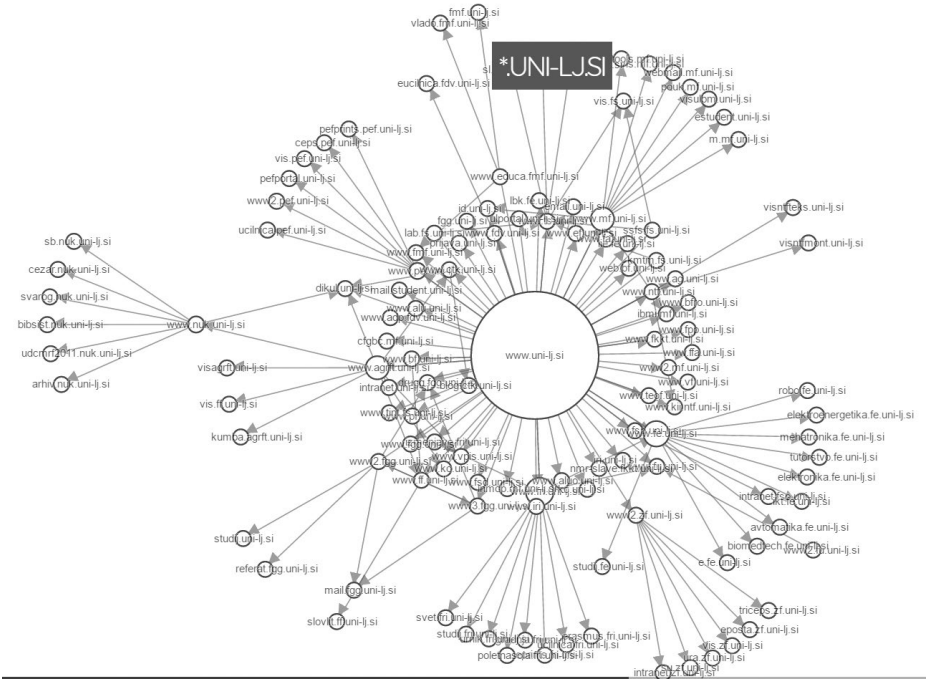


Slika 4: *Luščenje podatkov s pomočjo programa Scrapinghub*

5 Dostop do nestrukturiranih podatkov in njihova analiza

Spletni arhiv Narodne in univerzitetne knjižnice vsebuje spletna mesta, ki jih zaradi trajnega ohranjanja slovenske kulturne dediščine na svetovnem spletu periodično shranjujemo od leta 2008. K temu nas zavezuje Zakon o obveznem izvodu publikacij (2006), ki NUK-u nalaga skrb za zajem, ohranjanje in nudenje dostopa uporabnikom do zajetih spletnih publikacij, spletnih mest, vsebin. Z namenom ohranitve stanja strukture slovenske spletne domene *.si*, poskusa analize sestave podatkov, ki se tam nahajajo, in tvorjenja posnetka slovenskega spleta v nekem trenutku smo v NUK-u med letoma 2015 in 2016 z zajemom vsebin iz celotne slovenske spletne hrbtenice ***.si*** pridobili približno 2 TB podatkov. Uporabili smo seznam spletnih domen registrarja Arnes, kar je predstavljalo nabor približno 105.000 spletnih domen. Tudi za okolje ***.si*** je značilen visok delež spletnih domen, ki so zgolj »parkirane« ali pa rezervirane. Zajem je trajal približno pol leta, pri čemer so bili ključni cilji zajema naslednji:

– preveriti delovanje/aktivnost spletne domene,
– arhivirati slovenski splet v določenem časovnem trenutku,
– pridobiti podatke o razvejenosti in količini informacij na spletni domeni (glej primer na Sliki 5) in
– pridobiti informacijo o poddomenah na določeni spletni domeni.



Slika 5: *Primer razvejenosti spletne domene uni-lj.si*

Poleg tega pa smo želeli **predvsem**:
- opraviti analizo besedišča spletnih domen in slovenskega spleta ter
- poiskati sorodnosti, korelacije in razmerja med spletnimi domenami oziroma besedili na njih.

Na Sliki 5 podajamo primer razvejenosti spletne domene uni-lj.si. Zajem podatkov na osnovni strani te domene nam prikazuje različne poddomene, na katere kaže domena uni-lj.si. Zaradi zahteve, da bi zajem opravili v tekočem letu, in omejenih računalniških kapacitet smo obiskali zgolj prvi nivo globine vsake spletne domene. Če bi šli globlje, bi bila Slika 5 ustrezno kompleksnejša, prikazovala bi več spletnih poddomen, saj bi robot pri zajemanju znova in znova odkrival nove. Kot primer naj navedemo spletno poddomeno nuk.uni-lj.si. V kolikor bi robot raziskal globlje, bi dobili podatek tudi o vseh poddomenah naše, nuk.uni-lj.si domene, v naslednjih korakih pa poddomene poddomen itd.

Opazili smo, da je veliko spletnih domen atomarnih, torej takšnih, ki ne vsebujejo poddomen; večinoma so to spletne trgovine, predstavitve podjetij, oglasi itd. Nasprotno je v primeru domen, kot sta **gov.si** in **uni-lj.si**, jedro informacij v poddomenah; gre za razvejen sistem, ki nam vsebine, ki bi jo radi zajeli, arhivirali in analizirali, ne ponuja zgolj na vstopnih straneh. To pomeni, da bi za potrebe

izgradnje celotne slike večje, razvejene domene, kot so uni-lj.si, gov.si in podobne, morali opraviti globlji, kompleksnejši zajem kot za večino spletnih domen, ki imajo podatke zgolj na zgornjem nivoju ter so – plitke.

Za potrebe analize besedišča spletnih domen in slovenskega spleta ter iskanja povezav med besedili na spletnih domenah je treba gradivo, besedišče, ki nastopa v podatkih, ustrezno pripraviti. To storimo v več korakih ali fazah. Najprej smo besedišče, ki smo ga pridobili z zajemom vsebine spletnih domen, lematizirali (Lematizacija, B. l) z uporabo slovarjev. Vsaki posamezni rabljeni besedni obliki smo določili osnovno, slovarsko obliko, ki jo poimenujemo lema (osnovna oblika), npr. ***elokvenca***, *elokvence, elokvenci, elokvenco, elokvenc, elokvencami, elokvencam, elokvencah, elokvencama*. Vse naštete sklanjalne oblike so dobile isto lemo – elokvenca.

V drugi fazi analize smo iz besedišča izločili vse besedne vrste, ki niso samostalnik, pridevnik ali glagol (prislov, predlog, veznik, členek, medmet). Tako smo množico besed občutno zmanjšali, poleg tega pa poskrbeli, da besede, kot so vezniki, ne bi vplivale na analizo, predvsem pa na statistiko besedila.

Pri analizi podatkov, besedišča, ki smo ga z uporabo zajema pridobili in »očistili«, smo uporabili metode tako imenovane »vreče besed« (Bag of words, 2017). »Vreča besed« je algoritem, s pomočjo katerega štejemo pojavljanje besed v dokumentu, s tem pa lahko posredno merimo sorodnost med njimi.

Teža termina, ki je vsebovan v dokumentu, je enaka deležu pojavitve tega termina v razmerju do preostalega besedišča (Luhn, 1957).

$$E(t) = \begin{cases} 1, & \text{če } t = \text{»igra«} \\ 2, & \text{če } t = \text{»lutkoven«} \\ 3, & \text{če } t = \text{»grški«} \\ 4, & \text{če } t = \text{»gledališče«} \\ \ldots & \ldots \\ \ldots & \ldots \\ m, & \text{če } t = \text{»čitalnica«} \end{cases}$$

V množici $E(t)$ hranimo različne besede, ki se pojavijo v dokumentih, pridobljenih iz spletnega zajema. Statistika za določanje pogostosti pojavitve se imenuje »frekvenca termina« (angl. term frequency), njen avtor je Hans Peter Luhn, prvič je omenjena leta 1957. Zapišemo jo kot:

$$tf(t, d) = \sum_{x \in d} fr(x, t)$$

kjer je $fr(x, t)$ preprosta funkcija, definirana kot:

$$fr\,(x,\ t) = \begin{cases} 1, & \text{če } x = t \\ 0, & \text{sicer} \end{cases}$$

Funkcija *tf* (*t, d*) nam vrne število pojavitev termina *t* v dokumentu *d*.

Uporabnost omenjenega algoritma je omejena, saj nam najpogosteje rabljene besede ne pomagajo pri iskanju sorodnosti med publikacijami, članki itd.

Za primer vzemimo nekaj spletnih dokumentov (639) na spletnem portalu ***rtvslo.si*** ter nekaj (223) na spletnem mestu ***delo.si*** s področja »kultura«.

Prvih deset besed po frekvenci pojavitve na ***rtvslo.si*** je: *»biti, svet, novic, novica, filmski, kitajska, kitajski, nov, majhen, slovenski«*.

Prvih deset besed po frekvenci pojavitve na ***delo.si*** je: *»glasba, dober, biti, nov, dobro, dohodek, dokazovanje, dober, dom, novica, leto«*.

Opazimo lahko, da so nekateri termini sicer skupni, pojavljajo se tako na enem kot drugem spletnem mestu, vendar na njihovi podlagi, ravno zaradi njihove splošnosti, ni mogoče sklepati, za kakšen tip informacij gre v člankih. Sama frekvenca pojavitve besed v besedilu očitno še ni informacija, ki bi nam bila v pomoč. Kot smo zapisali, smo v naboru besed izmed besednih vrst obdržali zgolj polnopomenske besedne vrste: samostalnike, pridevnike in glagole. Če besedil ne bi predhodno obdelali in bi ostala neizluščena, bi bil rezultat še slabši, kajti nastopale bi vse besedne vrste (npr. vezniki, predlogi itd.) v vseh oblikah. Po tej statistiki bi bile prav nepolnopomenske besedne vrste najfrekventnejše. Z namenom pridobitve boljših, reprezentativnejših terminov iz besedišča zajetih dokumentov smo vpeljali drugo metodo izračunavanja statistike, angleško imenovano **Term frequency-inverse document frequency *»tf – idf«*** (B. l.). Sestavljena je iz dveh delov, in sicer iz zdaj že znane ***tf*** (frekvence pojavitve terminov) in ***idf***, v angleščini ***»inverse document frequency«***. Gre za obratno frekvenco pojavitve terminov v dokumentih. Po tej statistiki je **pomembnejša beseda** ali termin, ki **se v različnih besedilih pojavlja manjkrat**. S tem ima večji vpliv na vsebino oziroma specifiko dokumenta.

$$idf\,(t) = log\ \frac{|D|}{1 + |\{d{:}t{\in}d\}|}$$

Kjer je |D| = N (**število vseh dokumentov** korpusa, ki ga pregledujemo) 1 + |{*d*:*t*∈*d*}| **število dokumentov**, v katerih se termin pojavlja. Zaradi nevarnosti deljenja z ničlo je dodana vrednost +1 v imenovalcu.

Končna enačba za izračun numerične statistike ***tf – idf*** je tako:

$$tf - idf\,(t) = tf\,(t, d) \times idf\,(t)$$

Kot rezultat dobimo besede, razvrščene glede na težo termina – višja številka označuje besedo z višjo težo, ki je torej pomembnejša za analizirano besedišče.

S primerom ponazorimo razliko nabora reprezentativnih besed na različnih spletnih mestih:

Arhiv spletnega mesta **uni-lj.si** (401.064 besed, od tega 13.213 različnih, analiziranih 2239 dokumentov):
- **tf**: *mesec, trikotnik, nov, biti, let, leto, seja, copy, mail, geslo, ime, zapomniti* ...
- **tf-idf**: *odsev, vsota, lik, mladinski, trikotnik, teleskop, baleten, kolokvij, teologija, poklic, patenten farmacevtski, observatorij, družbosloven* ...

Arhiv spletnega mesta **gov.si** (826.605 besed, od tega 14.652 različnih, analiziranih 4790 dokumentov):
- **tf**: *lina, podelitev, utrinek, stran, postavitev, obstajati, beseda, odziv, digitalen, severozahod* ...
- **tf-idf**: *državljanski, lasten, demokrat, mina, brestanica, shod, izdelovalec, slabost, parlament, krven* ...

Zaradi velike količine podatkov, tj. besed znotraj posamezne domene, je težko pričakovati, da bomo s postopkom krnitve besedišča na zgolj deset, petnajst besed natančno opisali množico reprezentativnih besed spletne domene oziroma da bi jo s pomočjo teh nekaj besed uspeli umestiti v neko klasifikacijo. Opazimo lahko, da uporaba frekvence pojavitve besede v dokumentu (ali spletni domeni), torej statistika ***tf***, ne nudi dovolj uporabnih informacij. Veliko bolje vsebino na spletni domeni opiše statistika ***tf-idf***.

Poskusimo še z enim primerom.

Tokrat se bomo osredotočili na spletna portala **rtvslo.si/kultura** in **delo.si/kultura**. Z obeh spletnih portalov smo za poskus analizirali zgolj nekaj dokumentov in analizi med seboj primerjali (iz vsake domene po štiri).

Arhiv spletnega mesta **delo.si/kultura**:
- **tf**: *biti, gledališče, opera, slovenski, let, leto, delati, kazati, izbor, dan, ocena* ...
- **tf-idf**: *komedija, knjiga, opera, slovenski, kriminalka, britanski, delati, komedijant, kazati, izbor, smrt* ...

Arhiv spletnega mesta **rtvslo.si/kultura**:
- **tf**: *biti, predstava, igra, lutkoven, slovenski, leto, let, knjiga, delati, ga, nov ...*
- **tf-idf**: *igra, lutkoven, grški, gledališče, knjižnica, inšpektor, kriminalen, mors, oxford, pisec, smrt ...*

Za boljšo ponazoritev navedimo še en primer, tokrat na tematiko športa:

Arhiv spletnega mesta **delo.si/sport**:
- **tf**: *biti, let, leto, velik, čudovit, nov, pet, stan, igra, človek, dober ...*
- **tf-idf**: *velik, Federer, izkušen, masters, zmaga, želeti, razmišljati, ameriški, atp, finalen, melbourne ...*

Arhiv spletnega mesta **rtvslo.si/sport**:
- **tf**: *biti, igra, zmaga, dvoboj, dober, obračun, niz, turnir, velik, dober, drug, dvoboj, forma ...*
- **tf-idf**: *koleno, vrh, forma, Federer, pomeriti, odpovedati, nastop, legendaren, Djoković, lestvica, finale, atp ...*

Pri preverjanju sorodnosti besedišča dveh spletnih medijev (rtvslo.si in delo.si) smo za področje kulture z vsakega spletnega mesta analizirali po štiri članke. Pri tem smo zanemarili, ali gre za isto temo, torej enake novice na drugem spletnem mestu. Opazimo lahko, da je izluščeno besedišče mogoče umestiti na pomensko polje kulture, ne gre pa za velik odstotek preseka besed med spletnima domenama, saj gre za različne novice na enem ter drugem spletnem mestu. Z večanjem množice novic o neki tematiki (npr. kultura) se statistično ugotovljene razlike med besediščem enega in drugega spletnega mesta manjšajo, kar smo lahko opazili na analiziranem primeru. To velja za statistiko ***tf-idf***.

Pri primeru športa smo se posvetili nekaj teniškim novicam. Izbrali smo po tri športne teniške novice s spletnega mesta rtvslo.si in delo.si (poudariti je treba, da je na spletnem mestu delo.si neregistriranim uporabnikom omogočeno pregledovanje zgolj začetnih delov člankov). Tu lahko opazimo precej večjo podobnost med uporabljenim besediščem v novicah (spletnih dokumentih), kar je posledica dejstva, da novinarji, čeprav iz različnih medijskih hiš, uporabljajo soroden jezik in terminologijo pri ustvarjanju člankov na isto tematiko.

Za konec poglejmo še primer novinarskega prispevka, ki so ga o isti temi pripravile tri različne medijske hiše (**rtvslo.si**, **delo.si**, **primorske.si**):
- www.delo.si/kultura/knjiga/umrl-mojster-kriminalke-colin-dexter.html
- www.rtvslo.si/kultura/drugo/poslovil-se-je-colin-dexter-avtor-kriminalk-o-inspektorju-morsu/417882

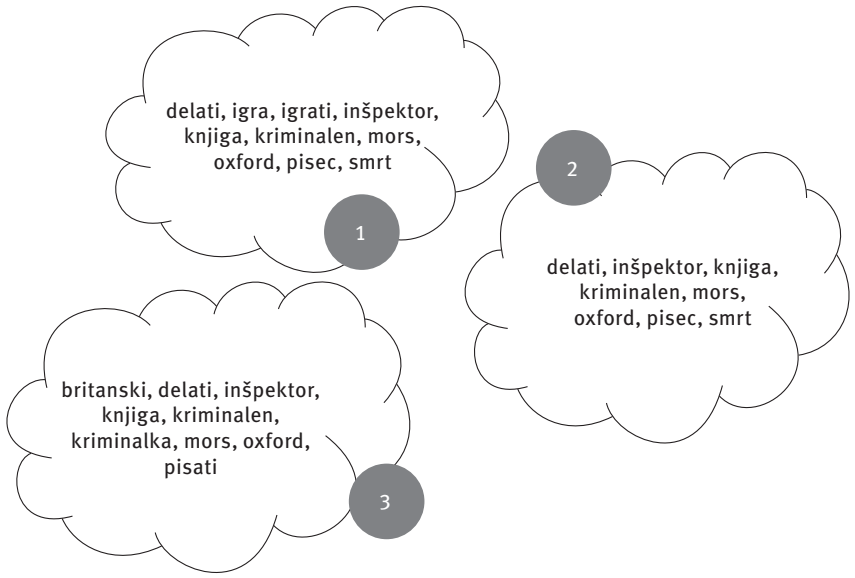
– http://www.primorske.si/Novice/Kultura/Umrl-Colin-Dexter-pisec-kriminalk-o-inspektorju-Mo

Na Sliki 6 prikazujemo sorodne besede za novinarska besedila o istem dogodku, objavljena na treh različnih spletnih mestih.

V **oblačku št. 1** prikazujemo število pomembnih besed (z najvišjo statistiko ***tf-idf)***, ki se pojavijo v prispevku na domeni **rtvslo.si** in **delo.si**.

V **oblačku št. 2** prikazujemo število pomembnih besed (z najvišjo statistiko ***tf-idf)***, ki se pojavijo v prispevku na domeni **rtvslo.si** in **primorske.si**.

V **oblačku št. 3** prikazujemo število pomembnih besed (z najvišjo statistiko ***tf-idf)***, ki se pojavijo v prispevku na domeni **delo.si** in **primorske.si**.



Slika 6: *Presek besedišča, uporabljenega za poročanje o istem dogodku na različnih spletnih portalih*

Algoritem za določanje statistike tf-idf se v praksi zelo pogosto uporablja. Najpogostejši primeri uporabe so spletni iskalniki, filtriranje in nudenje relevantnih rezultatov. V našem primeru nam algoritem sporoča podobnosti med iskanimi spletnimi viri, dokumenti. S pomočjo algoritma lahko uporabniku ponudimo boljše zadetke ter ga napotimo na tiste dokumente ali vire, ki so po iskanih terminih sorodni. Praktična uporaba pri nudenju lastnih storitev se kaže predvsem

v razvrščanju zadetkov in nudenju primernih vsebin pri nestrukturiranih, formalno neobdelanih publikacijah, kot so starejši članki na spletnem portalu Digitalne knjižnice Slovenije in arhivu spletnih vsebin. Ne nazadnje lahko povezujemo različne (lastne) spletne servise in aplikacije oziroma različne baze znanj.

6 Uporabnost arhiva in pomen arhiviranja

Prikazali smo možnosti in namen arhiviranja spletnih mest ter uporabo tako pridobljenega spletnega arhiva. Skladno z Zakonom o obveznem izvodu publikacij (ZoiPub) (2006) je primarna naloga NUK-a (na tem področju) ohranjanje in varovanje zajetega gradiva ter nudenje javnosti dostop do arhiva. Podobno kot se kaže tendenca v javnih katalogih želimo tudi v spletnem arhivu uporabniku na enem mestu nuditi čim bolj kakovostne informacije ter možnost povezovanja različnih (lastnih) virov razpršenih podatkov (več deset terabajtov gradiva). Zato se same po sebi kažejo tudi potrebe po uteževanju, luščenju, združevanju in drugih obdelavah podatkov. Z uporabo različnih tehnik, npr. razvrščanje po pomenskem vrednotenju posameznih besed, lahko do podatkov dostopamo hitreje, iz celotnega korpusa dokumentov pa lahko uporabniku predlagamo tisto gradivo, ki se zdi iskanemu najbolj sorodno. In to ne zgolj na enem spletnem mestu ali spletnem servisu, temveč v množici vseh, do katerih imamo dostop. Pri delu, na področju razvrščanja besedil, bomo nadaljevali z vpeljavo in uporabo sorodnih algoritmov, za potrebe razvrščanja, grupiranja (gručenje) in klasificiranja podatkov. S pomočjo teh algoritmov želimo omogočiti avtomatsko Univerzalno Decimalno Klasfikacijo (UDK) publikacij, oziroma ponuditi bibliotekarjem pomoč pri razvrščanju novih, digitalnih publikacij.

Navedeni viri

Archive-It. (2014). San Francisco: Archive-It. Pridobljeno 11. 3. 2017 s spletne strani: https://archive-it.org

Bag of words and TF-IDF [blog zapis]. (2017). S.l.: Deeplearning4j. Pridobljeno 17. 3. 2017 s spletne strani: https://deeplearning4j.org/bagofwords-tf-idf

Brown, A. (2013). *Practical digital preservation: a how-to guide for organizations of any size*. London: Facet Publishing.

Dexi.io. (2012). Copenhagen: Dexi.io. Pridobljeno 11. 3. 2017 s spletne strani: https://dexi.io

Dramowicz, K. (2016). Acquiring geographical data with web harvesting. *IOP conference series: earth and environmental science*, *34*(1), 1–8. doi:10.1088/1755-1315/34/1/012006

February 2016 web server survey [blog zapis]. (2017). Bath: Netcraft. Pridobljeno 10. 3. 2017 s spletne strani: https://news.netcraft.com/archives/category/web-server-survey/

GNU Wget. (2017). S.l.: GNU Operating System. Pridobljeno 11. 3. 2017 s spletne strani: https://www.gnu.org/software/wget

How many active sites are there? [blog zapis]. (2008). Bath: Netcraft. Pridobljeno 10. 3. 2017 s spletne strani: https://www.netcraft.com/active-sites

HTTrack website copier: version 3.49-1. (2017). S.l.: Xavier Roche and other contributors. Pridobljeno 11. 3. 2017 s spletne strani: http://www.httrack.com/

Import.io. (2017). Los Gatos, CA: Import.io. Pridobljeno 11. 3. 2017 s spletne strani: https://www.import.io

Internet archive wayback machine. (2014). San Francisco: Internet Archive. Pridobljeno 11. 3. 2017 s spletne strani: https://archive.org/web

ISO 28500:2009, Information and documentation – WARC file format. (2009). Geneva: ISO.

Jack, P. (2014). *Heritrix*. S.l.: Confluence. Pridobljeno 11. 3. 2017 s spletne strani: https://webarchive.jira.com/wiki/display/Heritrix/Heritrix

Lematizacija. (B. l.). V *Wikipedija: prosta enciklopedija*. Pridobljeno 15. 3. 2017 s spletne strani: https://sl.wikipedia.org/wiki/Lematizacija

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM journal of research and development*, *1*(4), 309–317. doi:10.1147/rd.14.0309

Parsehub. (2017). Toronto: ParseHub. Pridobljeno 11. 3. 2017 s spletne strani: https://www.parsehub.com

Perma.cc. (2013). Cambridge, MA: Harvard Law School Library. Pridobljeno 12. 3. 2017 s spletne strani: https://perma.cc

Scrapinghub. (2010). Cork: Scrapinghub. Pridobljeno 11. 3. 2017 s spletne strani: https://scrapinghub.com

Term frequency-Inverse document frequency. (B. l.). V *Wikipedia: the free encyclopedia*. Pridobljeno 10. 3. 2017 s spletne strani: https://en.wikipedia.org/wiki/Tf%E2%80%93idf

WebCite. (B. l.). Toronto: WebCite Consortium. Pridobljeno 11. 3. 2017 s spletne strani: http://www.webcitation.org/

Zakon o obveznem izvodu publikacij (ZOIPub). (2006). *Uradni list RS*, št. 69/2006 in 86/2009.

Matjaž Kragelj

Narodna in univerzitetna knjižnica, Turjaška 1, 1000 Ljubljana
e-pošta: matjaz.kragelj@nuk.uni-lj.si

Mitja Kovačič

Narodna in univerzitetna knjižnica, Turjaška 1, 1000 Ljubljana
e-pošta: mitja.kovacic@nuk.uni-lj.si