Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Paula Vouk

Analiza spletnih novic s tehnikami prikaza pojavitev besed in besednih zvez

diplomsko delo

visokošolski strokovni študijski program prve stopnje računalništvo in matematika

Mentor: prof. dr. Blaž Zupan

Ljubljana, 2016

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

V diplomski nalogi raziščite uporabo vizualizacij tipa Circos in Sieve za analizo pojavitev besed oziroma besednih zvez za besedila v slovenskem jeziku. Besedila za izbrano časovno obdobje pridobite iz odprtih, spletnih virov. Besedila primerno predobdelajte (npr. lematizacija). Uporabnost vizualizacijskih tehnik ocenite preko izbranih primerov časovne pojavitve ključnih besed ali pa zanimivih besednih zvez.

Zahvaljujem se mentorju prof. Blažu Zupanu, Ajdi Pretnar, Andreju Čoparju in ostalim članom Laboratorija za bioinformatiko za ideje in pomoč pri izdelavi diplomske naloge.

Viljana, Mia, nona, Eva in Peter hvala, da ste mi stali ob strani v času študija, mi nudili nasvete kadarkoli sem jih potrebovala. Veselili ste se z mano uspehov in me bodrili, ko sem podvomila vase. Posebna zahvala gre še Istoku, čigar znanja so mi tudi tokrat prišla še kako prav.

Noni Mariji

Kazalo

Povzetek

Naslov: Analiza spletnih novic s tehnikami prikaza pojavitev besed in besednih zvez

Na voljo imamo ogromne količine literature v slovenskem jeziku, iz katere lahko s preprostimi algoritmi veliko izvemo o naši družbi in njeni kulturi, znanosti, politiki ter drugih področjih. V diplomski nalogi smo izbor zožili na novičarske članke, ki so bili med letoma 1998 in 2006 objavljeni na spletni strani časopisa *Dnevnik*. S pomočjo grafov frekvence pojavitev določenih besed in besednih zvez smo želeli prikazati vpliv nekaterih pomembnih dogodkov v svetovnem in slovenskem merilu na poročanje slovenskih medijev. Ugotovili smo, da se povečane frekvence pojavitev besed kronološko ujemajo s pripadajočimi fenomeni. Preučevali smo tudi sopojavitve nekaterih poznanih imen s pojmi ter jih na ta način umestili v tematsko okolje. Na številnih primerih smo preizkusili kako sta za predstavitev rezultatov takšne vrste primerni sievovi in circos diagrami. Dobljene povezave med besedami so smiselne in do neke mere pričakovane, kljub temu pa nastali diagrami prikazujejo in poudarjajo zanimiva in presenetljiva razmerja.

Ključne besede: Circos, sievov diagram, n-gram, sopojavitev besed, frekvenca besed.

Abstract

Title: Online news analysis with the techniques of word occurrence visualization

There is an enormous amount of publications in Slovenian language waiting to be analysed. With simple algorithms we can reveal interesting facts about our society and it's culture, science, politics as well as many other aspects. In this thesis we focused on online articles that were published by newspaper *Dnevnik* between 1998 and 2006. By evaluating word-usage frequency graphs we wanted to investigate the influence of some important phenomena on Slovenian press. We found that higher usage frequencies of specific words chronologically match with associated phenomena. We also studied how the names of well-known people co-occur with words that pertain to a specific topic. With several examples we examined how appropriate Sieve and Circos diagrams are to visualising these types of results. Word connections presented with selected visualization tools are meaningful and expected but on the other hand the diagrams bring forward some interesting and unexpected relations.

Keywords: Circos, Sieve diagram, n-gram, word co-occurrences, word frequency.

Poglavje 1

Uvod

V Sloveniji je samo leta 2015 izšlo 4.941 knjig, vsak dan pa izide najmanj 11 časopisov in revij. Na voljo imamo ogromne količine podatkov, ki iz dneva v dan še rastejo, vendar ostajajo slabo izkoriščene. Vprašajmo se, kaj vse bi lahko počeli s takšno količino besedil, ki je na nek način obraz slovenske kulture in družbe. Vse te literature seveda ni mogoče prebrati, lahko pa poskusimo poiskati drugi način, da izluščimo željene informacije in morda dobimo grob vpogled ali nekakšen povzetek vsega napisanega.

Z zelo preprostimi tehnikami obdelave besedil, kot so štetje besed oz. besednih zvez, ki se pojavljajo v literaturi, lahko pridemo do zanimivih odkritij. Novinarski članki, s katerimi smo se ukvarjali v tem diplomskem delu, zajemajo širok nabor tem kot so politika, gospodarstvo, šport, kultura in znanost na enem mestu. Z analizo pojavitev primernih besed se lahko vsaj dotaknemo vsake izmed njih. Raziščemo lahko, kako se dogodki odražajo na spremembi uporabe besed na časovni osi. O čem mediji več poročajo, kateri družbeni fenomeni v zgodovini pustijo za seboj večji pečat, o čem govorimo še leta in kaj ter na koga takoj pozabimo. Analiza vsebine besedil pa še zdaleč ni vse kar lahko počnemo. Če lahko dostopamo do malce starejših besedil, lahko preučujemo, kako se naš jezik spreminja, ali se pojavljajo nove besede in katere vse bolj izginjajo in na koncu izumrejo.

Diplomska naloga, s katero smo želeli odgovoriti vsaj na nekaj zgornjih

vprašanj sestoji iz treh delov. Pri prvem smo želeli ponoviti poskuse, ki so bili že narejeni na veliko večjem in precej drugačnem korpusu besedil. S pomočjo merjenja frekvence pojavitev besed smo poskušali prikazati nekaj dogodkov v svetovnem in pa slovenskem merilu, ki so za seboj pustili močan pečat. V drugem delu se sprašujemo ali obstajajo zanimive povezave med določenimi besedami in ali se slednje odraža na tem, da besede v člankih pogosto tičijo blizu druga drugi. Pri tem smo se osredotočili predvsem na imena oseb v kombinaciji s pojmi, ki sodijo v sorodno tematsko polje. Tretja naloga pa je bila poiskati orodja, s katerimi bi na nazoren način prikazali rezultate analiz. Iskali smo način prikaza, ki bi bil dovolj preprost, dovolj pregleden in vendar ne povsem običajen. Preiskusili smo vizualizacijsko orodje Circos, ki tipično ni uporabljen na podatkih te vrste ter prikaze primerjali s prepoznavnejšimi sievovimi diagrami.

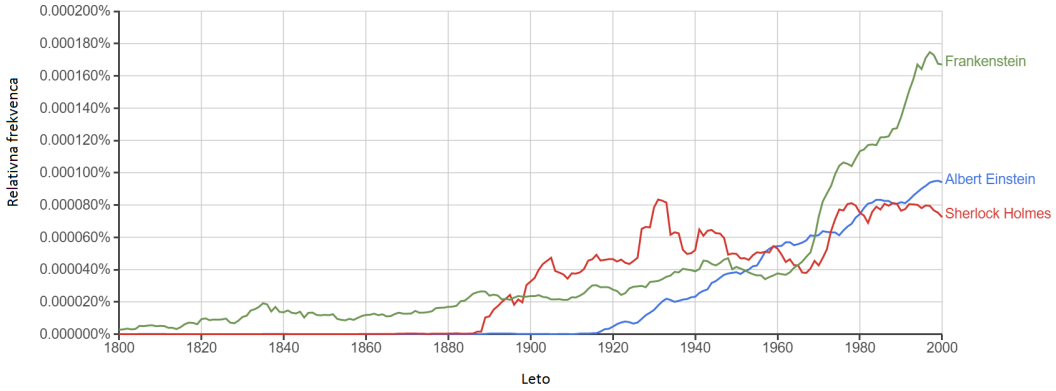
1.1 Sorodna dela

Idejo za osrednjo temo diplomskega dela smo črpali iz študije kvantitativne analize kulture z uporabo digitaliziranih knjig, ki sta jo vodila raziskovalca s Harvarda Erez Lieberman Aiden in Jean-Baptiste Michel [1].

Raziskava je bila izvedena na 4 odstotkih vseh knjig, ki so bile kadarkoli natisnjene. Tako obsežen korpus je omogočal kvantitativno analizo kulturnih trendov, ki sta jo avtorja poimenovala *Culturomics*. Avtorja sta se osredotočila na lingvistične in kulturne fenomene, ki so se izražali v angleškem jeziku med 1800 in 2000. Študija je pokazala, da s takšnimi prijemi lahko pridobimo vpogled na tako raznolika področja kot so leksikografija, evolucija slovnice, kolektivni spomin, cenzura, zgodovina epidemiologije in številne druge. Človek ne more prebrati celotnega korpusa. Če bi želeli prebrati le knjige v angleščini iz leta 2000, bi za to potrebovali 80 let brez premorov. Lahko pa poiščemo odgovor na vprašanje, kako pogosto je bil določen n-gram (zaporedje n-tih besed) uporabljen v času. Raziskovalci so odgovore podali z grafi pojavitvenih frekvenc n-gramov. Primerjali so nastale krivulje

in utemeljili rezultate z zgodovinskimi dejstvi.

Na podlagi zgoraj opisane študije je nastala spletna stran (*Google Books Ngram Viewer*), kjer si vsak lahko izbere poljuben nabor besed oz. besednih zvez in izrišejo se krivulje pojavitvenih frekvenc izbranih besed za poljubno obdobje med letoma 1800 in 2000. Na sliki 1.1 je primer grafa pojavitvenih frekvenc, ki je nastal na tak način.

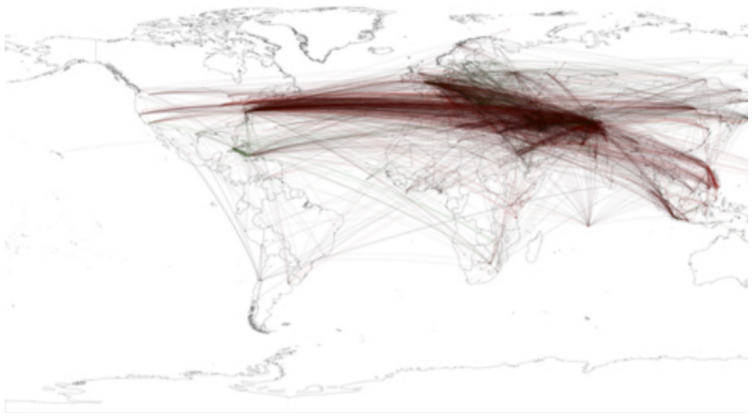


Slika 1.1: Primer grafa pojavitevenih frekvenc imen *Frankenstein, Albert Einstein in Scherlock Holmes*

Sledile so številne študije, ki analizirajo besedila knjig in pa predvsem novinarskih člankov. Spodaj omenjenim raziskavam je skupno, da opazujejo vpliv različnih družbenih faktorjev na preiskovana besedila oziroma se povsem fokusirajo na nekaj vsebovanih sociooloških in političnih fenomenov ter jih preučijo v globino. Rezultati so predstavljeni z diagrami in grafi, ki tradicionalnim načinom prikaza dodajajo inovativne vložke.

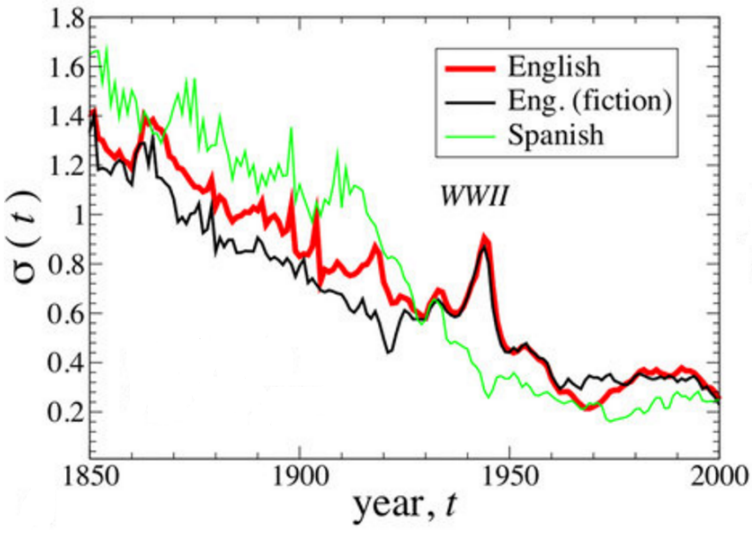
Raziskovalcem v študiji *Culturomics 2.0* je uspelo z analizo arhivov novičarskih strani napovedati številne ekonomske in politične dogodke kot so npr. revolucija v Tuniziji, Egiptu in Libiji [2]. Locirali so tudi potencialno skrivališče Osame Bin Ladna v radiju 200 kilometrov. Slika 1.2 prikazuje sopojavitve geografskih referenc z imenom *bin Laden* v člankih med letoma 1979 in 2001. Na prikazu lahko vidimo, da večina povezav vodi v severni Pakistan. Konkretneje skoraj 49 odstotkov člankov, ki omenja bin Ladna, vključuje tudi

kakšno mesto v Pakistanu.



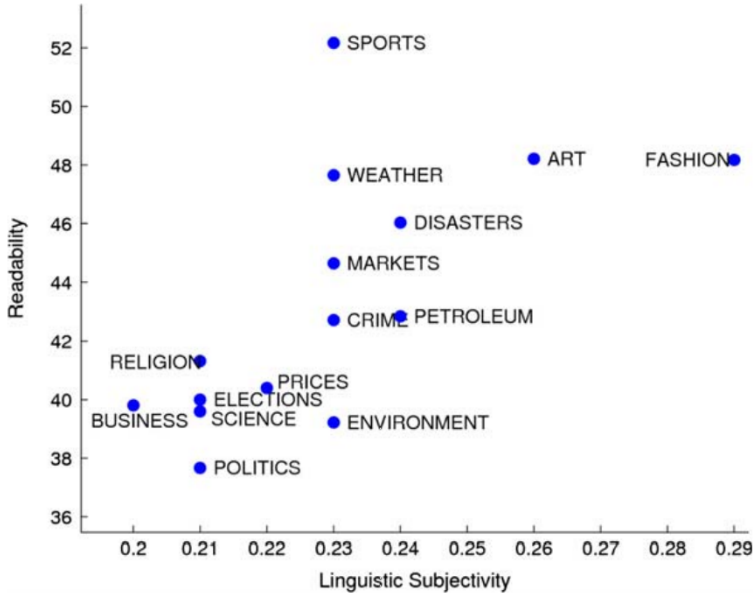
Slika 1.2: Prikaz sopojavitev imena *bin Laden* z geografskimi referencami. Povzeto po [2].

Leta 2012 je bila objavljena študija o statističnih zakonih uporabe besed od rojstva do smrti besede [3]. Analizirali so lastnosti 100 miljonov besed v angleščini, španščini in hebrejščini. Opazovali so kako se spreminja frekvenca uporabe starih in novih besed ter kako na razvoj in uporabo besed vplivajo politični, tehnološki in socialni faktorji. Na sliki 1.3 je prikazan graf spreminjanja standardne deviacije pojavitev novih besed med letoma 1850 in 2000. Krivulje prikazujejo spreminjanje španščine, angleščine ter angleškega jezika v leposlovni literaturi (v legendi *Eng. fiction*). Opazimo, da je druga svetovna vojna (*WWII*) povzročila rast krivulje angleškega jezika za razliko od španskega. To so znanstveniki utemeljili z izolacijo Španije in Južne Amerike od evropskega konflikta. V splošnem so ugotovili, da mednarodne krize lahko vodijo v globalizacijo jezika, vendar le v prisotnih jezikih. Na jezike oddaljenih regij imajo takšni konflikti minimalen ali ničelen vpliv. Podobno kot na rojstvo novih besed, na izumrtje starih vplivajo predvsem tehnološki in sociološki faktorji.



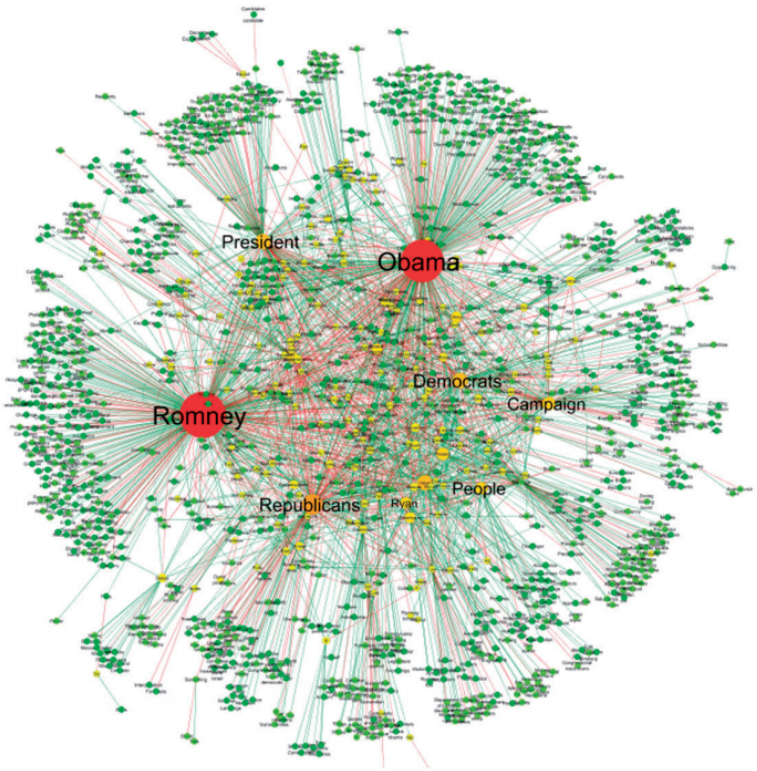
Slika 1.3: Pikaz spreminjanja španskega in angleškega jezika v času glede na pojavitev novih besed. Povzeto po [3].

Na Bristolski univerzi so istega leta analizirali 2.5 miljona člankov v angleščini iz 500 različnih novičarskih spletnih strani [4]. Zanimala jih je podobnost člankov glede na demografsko pozicijo bralcev in razmerje med spoloma pri določenih temah. Raziskovali so še, kakšna je relacija med popularnostjo in temo članka ter ali so članki določenih tem berljivejši (eden izmed kriterijev je dolžina besed in stavkov) oziroma subjektivnejši (glede na uporabljene pridevnike). Ugotovili so, da sta tematiki *šport* in *umetnost* veliko bolj berljivi kot *politika* in *ekologija* ter da so članki o *modi* najsubjektivnejši, kot je prikazano na sliki 1.4. Članki različnih tem pa se med seboj razlikujejo tudi po prevladi spola pojavljenih osebnosti. V *športnih* in *finančnih* člankih dominirajo moški, v *umetnosti* in *modi* pa je izid nevtralen. V splošnem med 1000 najbolj omenjenimi osebami v celotnem korpusu prevladujejo moški.

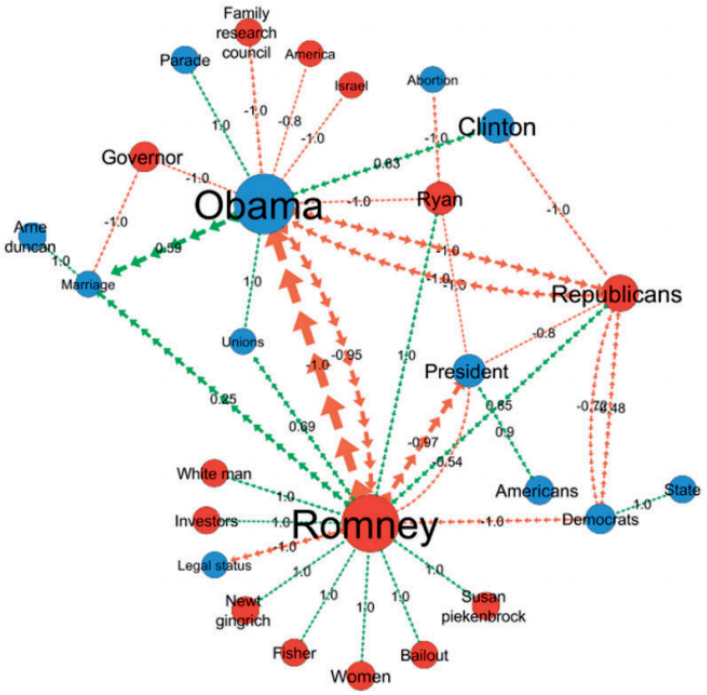


Slika 1.4: Pikaz primerjave novičarskih tem glede na stil pisanja. Povzeto po [4].

Leta 2015 je bila na 130.000 člankih opravljena analiza ameriških predsedniških volitev iz leta 2012 [5]. Zgrajeno je bilo omrežje političnih figur in problematik, ki so jih povezovale relacije podpore in nasprotovanja. Samostalniške besedne zveze predstavljajo vozlišča, povezujejo jih glagoli, ki predstavljajo akcijo enega vozlišča nad drugim. Raziskovalci so odkrili, da se ločnico med republikanskim in demokratskim taborom na enostaven način poišče s particijo grafa in identificirali najbolj centralna vozlišča obeh političnih strani. Študija je pokazala še, da je imel Clinton pomembnejšo vlogo med demokrati kot Biden, da se je v predvolilni kampaniji vse vrtelo okrog ekonomije in pravic ter da so mediji o demokratih poročali pozitivneje kot o republikancih. Na sliki 1.5 je eden izmed grafov, ki so nastali tekom študije. Rdeče črte predstavljajo negativne relacije, zelene pa pozitivne povezave oziroma konkretneje glagole. Sledi še slika 1.6, kjer je prikazan del omrežja, pobarvan s particijo grafa. Modre entitete pripadajo demokratom, redeče pa republikancem.



Slika 1.5: Pikaz grafa političnih figur predsedniških volitev leta 2012. Povzeto po [5].



Slika 1.6: Prikaz particije grafa političnih figur na demokrate in republikance. Povzeto po [5].

1.2 Pregled poglavij

V nadeljevanju pričujočega dela sledi poglavje Metode (poglavje 2), kjer je razloženo, s kakšno vrsto podatkov smo imeli opravka in na kakšen način smo jih pripravili za nadaljno obdelavo. Pojasnjeni so nekateri pojmi kot so n-gram, pojavitvena frekvenca in sopojavitev besed ter predstavljena je njihova vloga pri pridobivanju končnih rezultatov. Poglavje smo sklenili s predstavitvijo orodij, ki smo jih uporabili za vizualizacijo rezultatov.

V poglavju 3 se nahaja nekaj najbolj zanimivih prikazov spreminjanja frekvence pojavitev n-gramov v času, ki so rezultat naše raziskave na spletnih člankih. Sledi še poglavje, kjer smo z sievovim in circos diagramom predstavili sopojavitve besed, ki pripadajo nekaterim pogostim temam v novičarskih člankih (poglavje 4).

V zadnjem poglavju (poglavje 5) smo še enkrat poudarili pomembnejše ugotovitve, navedli nekaj idej kako bi se dalo delo še izboljšati in kaj vse se na tem področju še da storiti.

Poglavje 2

Metode

Na poti do diagramov in grafov, ki jih najdemo v poglavjih 4 in 5 smo se srečevali z različnimi izzivi, ki smo jih odpravili z uporabo sledečih tehnik in metod. Zajemu podatkov iz baze je sledila predobdelava teksta, za njo pa štetje pojavitev besed in besednih zvez ter sopojavitev posameznih besed. Dobljene rezultate smo prilagodili zahtevam vizualizaciskih orodij. Z njihovo pomočjo smo pridelali diagrame, ki na zanimiv način ponazarajo naše ugotovitve.

2.1 Pridobivanje in predobdelava podatkov

V namen raziskave spletnih novic v diplomski nalogi smo od anonimnega vira prejeli 217.000 sicer javno dostopnih člankov slovenskega časopisa Dnevnik objavljenih med letoma 1998 in 2006 na spleti strani časopisa. Ti so bili pridobljeni s spletnim luščenjem podatkov (angl. *web scraping*).

Predobdelava korpusa je zajemala tokenizacijo (delitev besedila na besede) in lematizacijo. Lematizacija (tudi geslenje) je postopek določanja osnovne (slovarske) oblike posameznim besedam, ki jih najdemo v besedilu. Osnovno obliko besede imenujemo lema[1]. Za lematizacijo smo uporabili

[1] `https://sl.wikipedia.org/wiki/Lematizacija`

knjižnico *LemmaGen*[2], ki se ni izkazala za popolnoma zanesljivo. Problematična so predvsem imena, ki se pogosto pretvarjajo v napačne besede (npr. *Bush* se pretvori v *Bus*, *Peterle* v *Petereti*). Prav ta pa so bila ključnega pomena pri nastanku številnih diagramov in grafov zato je bilo potrebno pretvorbo besed nadzorovati in razviti program za predobdelavo besedil.

2.2 Štetje pojavitev n-gramov besed

Eden izmed načinov, kako bi lahko izvedeli več o vsebinski sestavi korpusa je, da bi preverili, katere besedne zveze se znotraj besedil pojavljajo in koliko krat. Besedni 1-gram je zaporedje znakov, ki niso ločeni s presledkom oz. belim znakom tj. besede (*računalnik*), števila (*201.342*) in pa tudi tiskarski škrati (*Lubljana*) [1]. Zaporedju n-tih 1-gramov pravimo n-gram.(npr. besedna zveza *Računalništvo in informatika* je 3-gram).

2.2.1 Pojavitvena frekvenca

Predpostavimo, da nas zanima, kako se je uporaba nekega n-grama spreminjala v času. Samo štetje pojavitev n-grama nam ne bo pomagalo. Pogosto se namreč dolžine korpusev iz različnih časovnih okvirjev precej razlikujejo. Zato raje uporabimo relativno pojavitveno frekvenco t.j. število pojavitev n-grama normaliziramo s številom vseh besed v besedilu.

Grafi v poglavju 3 so nastali na sledeč način. Z oknom velikosti 7 dni smo se premikali čez korpus ter računali relativno pojavitveno frekvenco želenih besed oz. besednih zvez.

2.3 Sopojavitev besed

Za dve besedi pravimo, da se sopojavljata (angl. *co-occur*), takrat, ko je verjetnost, da se pojavita skupaj, večja od naključne oziroma od verjetnosti

[2] `http://lemmatise.ijs.si/`

sopojavitve, če bi bilo besedilo sestavljeno iz naključnega zaporedja besed. Medtem ko je n-gram tvorba sosednjih besed, tu slednje ni nujno, med njima je lahko tudi vnaprej določeno število drugih besed.

Iz besed, za katere nas je zanimalo ali se sopojavljajo, smo ustvarili dva seznama. Če smo želeli izvedeti, ali se določeni besedi pogosto v besedilu nahajata blizu, smo ta dva pojma uvrstili v različna seznama. Sopojavitve smo beležili s t.i. sopojavitveno matriko, katere stolpci predstavljajo elemente iz prvega ter vrstice elemente iz drugega seznama. S premičnim oknom velikosti 100 besed smo programsko pregledali besedila in beležili sopojavitve parov besed zotraj okna. Matriko s preštetimi sopojavitvami smo pozneje uporabili pri vizualizaciji sopojavitev s sievovim in circos diagramom v poglavju 4. Kot primer vzemimo preiskovanje sopojavitev besed med športnimi novicami. V prvem seznamu se bodo nahajala imena športnikov v drugem pa pojmi povezani s športom. Sopojavitvena matrika bo vsebovala informacijo kolikokrat se je določen pojem npr. *lovorika* pojavil v bližini določene osebe npr. *Schumacher*.

2.3.1 Točkasta vzajemna informacija

Pri sopojavitveni matriki smo vnaprej izbrali dva seznama besed in sešteli sopojavitve besed iz nasprotnih seznamov. Dodatno nas je zanimalo, s katerimi pojmi se najpogosteje skupaj pojavljajo izbrana imena (tokrat za razliko od prej izbora pojmov ne omejimo). Raziskovali smo na primer katere besede se najpogosteje pojavljajo ob politiku *Janezu Janši* in katere ob pevki *Madonni*.

Za vrednotenje sopojavitev besed x in y smo uporabili mero *točkasta vzajemna informacija* (ang. *point mutual information*)[3]:

$$\mathrm{pmi}(x;y) = \log \frac{p(x,y)}{p(x) \times p(y)}, \tag{2.1}$$

kjer je $p(x)$ verjetnost, da se v besedilu pojavi beseda x in $p(x,y)$ verjetnost,

[3] `https://en.wikipedia.org/wiki/Pointwise_mutual_information`

da se besedi pojavita v istem oknu.

Recimo, da je x v naprej izbrana beseda, y pa vsaka beseda, na katero v besedilu naletimo. Znova smo se s premičnim oknom premikali čez besedilo, vendar tokrat šteli sopojavitev besede x z vsemi ostalimi in izračunali vrednost po enačbi (2.1). Izmed vseh besed smo izbrali tiste, ki so se glede na izračunano vrednost $\mathrm{pmi}(x; y)$ najbolj sopojavljale z besedo x (vrednosti so bile najvišje). Na ta način smo pridobili seznam besed, ki so se pojavljala v besedilih zelo redko, vendar skoraj vedno v bližini besede x. Konkretneje, imena znanih osebnosti so se sopojavljala z drugimi imeni, ki javnosti niso tako poznana. Mediji so o njih poročali malokrat in sicer le ob specifičnih in unikatnih dogodkih kot so politične afere.

2.4 Vizualizacijski pristopi

Vizualizacija rezultatov je eden izmed ključnih problematik tega diplomskega dela. Grafi pojavitve frekvence besed so nastali z uporabo Pythonove knjižnice *matplotlib* in knjižnice *seaborn*[4], ki je omogočila detaljne olepšave. Za ponozoritev sopojavitev besed smo uporabili orodje *Circos*[5], nastale prikaze pa smo primerjali še s sivovimi diagrami.

2.4.1 Vizualizacija circos

Circos[5] je orodje za vizualizacijo podatkov. Uporablja krožno obliko prikazovanja podatkov, kar olajša prikaz relacij med objekti. Prvotno je izdelan za uporabo prikaza podatkov iz molekularne biologije, vendar ga je zaradi njegove prilagodljivosti možno uporabiti tudi na podatkih drugih vrst.

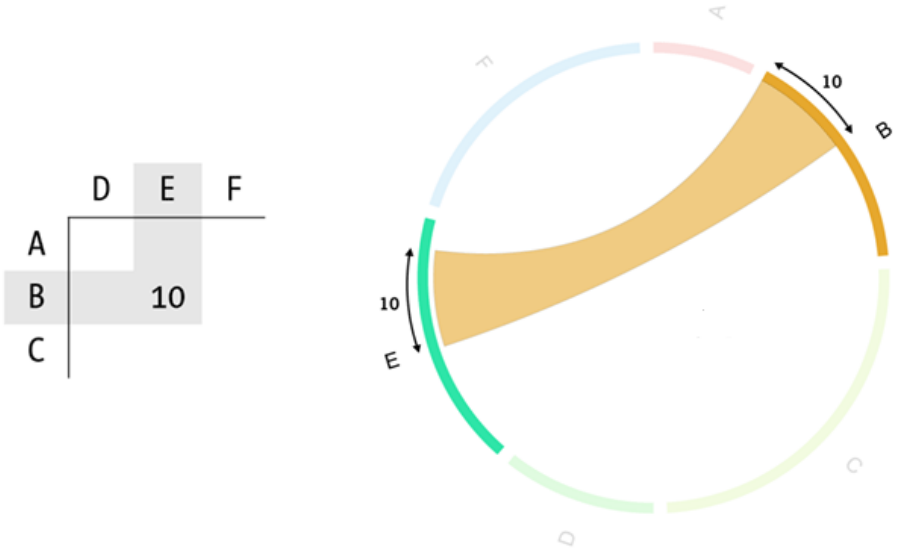
Med drugim lahko z orodjem circos prikazujemo tabelarne podatke[6], kot je prikazano na sliki. Elemetom na robu kroga pravimo segmenti, pobarvanim območjem, ki povezujejo segmente pa trakovi (angl. *ribbons*). Vrstice in

[4] `https://stanford.edu/~mwaskom/software/seaborn/`

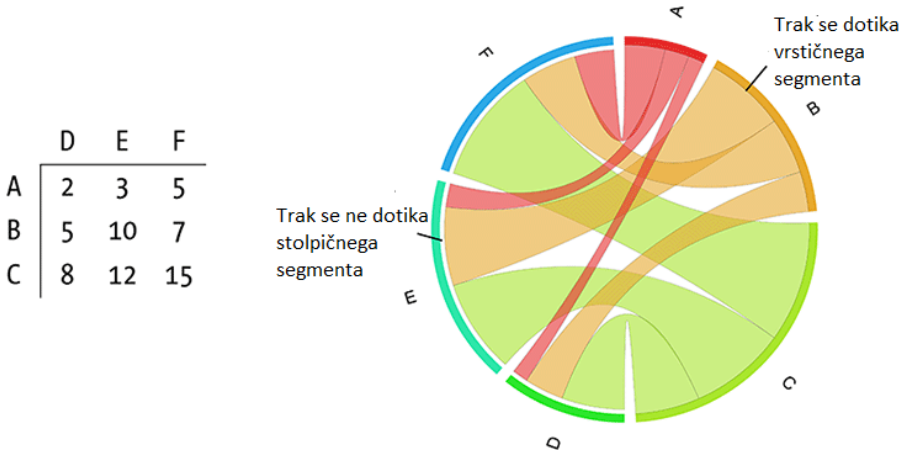
[5] `http://circos.ca/`

[6] `http://circos.ca/presentations/articles/vis_tables1/`

stolpci v tabeli so prikazani kot segmenti na krožnici, trakovi pa ponazarjajo celice tabele. Oranžen trak na sliki 2.1 predstavlja celico B-E v tabeli.



Slika 2.1: Preslikava celice iz tabele na Circos diagram[6].



Slika 2.2: Primer prikaza dvodimenzionalne tabele s Circosom[6].

Na upodobitvi na sliki 2.2 lahko opazimo, da se vrstični segmenti dotikajo trakov, stolpični pa ne. Na tak način na grafu ločimo dve vrsti segmetov.

Barva traku pripada enemu izmed segmentov, ki ju povezuje. Na primeru s slike je to vedno vrstični segment.

Razmerja med širinami trakov se ujemajo z razmerji med števili v tabeli. Na prikazu s slike 2.2 ima polje D-F največjo vrednost v tabeli, posledično je trak, ki povezuje D in F, najširši.

V določenih primerih nas ne zanima, kateri segment prevladuje v tabeli, temveč hočemo prikazati le razmerja znotraj segmentov (npr. kakšen delež pripada rdečemu traku v segmentu F v primerjavi z segmentom E). V ta namen lahko diagram normaliziramo. Vsi segmenti na krožnici bodo tako enako veliki.

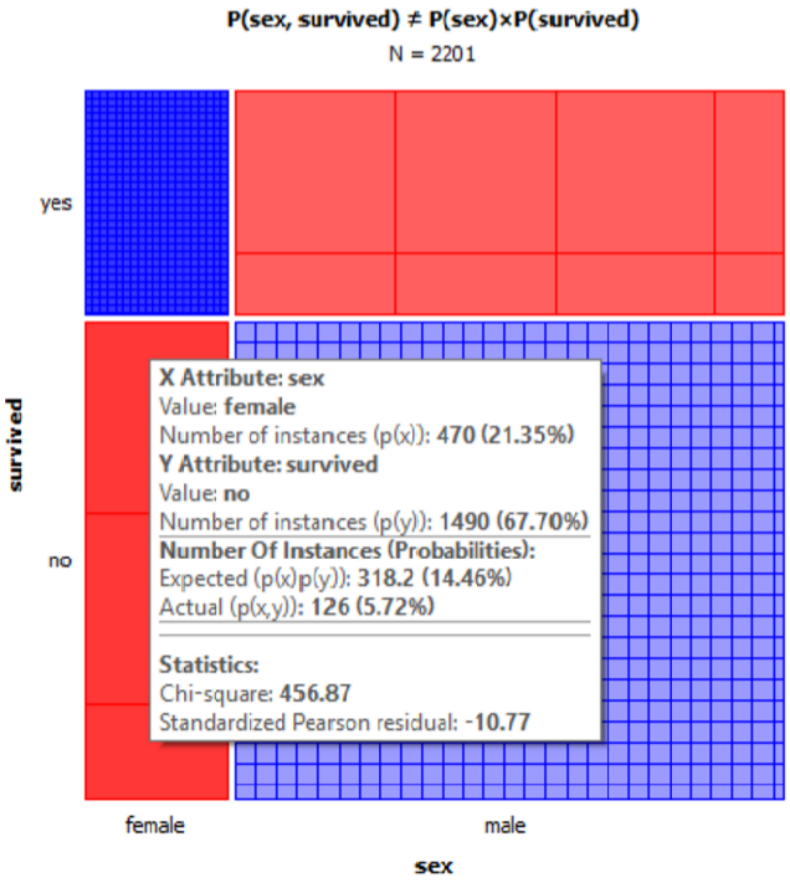
2.4.2 Sievov diagram

Sievov ali parquetov diagram[7] je grafična metoda za vizualizacijo frekvenc v dvodimenzionalni kontingenčni tabeli. Z njim primerjamo pričakovane frekvence s pojavitvenimi, pri čemer predpostavljamo neodvisnost atributov. Ploščina pravokotnikov je proporcionalna pričakovani frekvenci, ta pa je sorazmerna s Pearsonovo porazdelitvijo Hi-hvadrat, medtem ko je pojavitvena frekvenca sorazmerna s številom kvadratkov znotraj pravokotnika. Razlika med obema frekvencema je prikazana kot intenziteta pobarvanega pravokotnika. Barva je modra, če je deviacija od neodvisnosti pozitivna in rdeča, če je negativna.

Na sliki 2.3 je primer sievovega diagrama v orodju Orange[8], ki prikazuje razmerje potnikov na ladji Titanik, ki so oziroma niso preživeli nesreče. V spodnjem levem pravokotniku je vidno, da je žensk, ki niso preživele, veliko manj kot bi to pričalovali glede na število vseh žensk in število vseh žrtev nesreče. Deviacija od neodvisnosti je negativna, zato je pravokotnik osenčen z rdečo barvo.

[7] `http://docs.orange.biolab.si/3/visual-programming/widgets/visualize/sievediagram.html`

[8] `http://orange.biolab.si`



Slika 2.3: Primer sievovega diagrama[7] v orodju Orange[8], ki prikazuje razmerje preživelih in nepreživelih potnikov na ladiji Titanik.

Poglavje 3

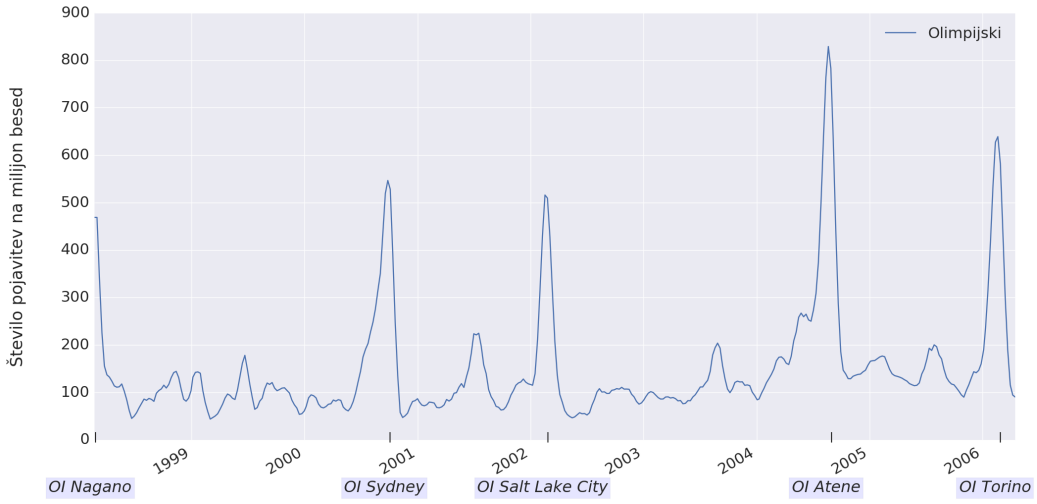
Grafi pojavitvenih frekvenc

Sledi devet grafov, s katerimi smo se skušali prikazati kakšen vpliv so imeli na poročanje slovenskih medijev nekateri dogodki med letoma 1998 in 2006 ter kakšen pečat so za seboj pustila velika svetovna in slovenska imena. Grafi ponazarjajo spreminjanje frekvence uporabe določenih besednih zvez v času. Nekateri so dodatno opremljeni z interpretacijo posameznih vrhov (na grafih se pojavljajo oznake dogodkov, ki so eksaktno časovno umeščeni). Upoštevati je potrebno, da so besede lematizirane (npr. pod oznako *terorist* spadajo tudi besede *teroristi, teroristični itd.*)

Graf na sliki 3.1 prikazuje kako se v času spreminja frekvenca besede *olimpijski*. Opazimo, da se vrhovi ujemajo z datumi olimpijskih iger. Po grafu sodeč mediji konstantno pišejo o olimpijskih igrah, saj se krivulja nikoli ne spusti na ničlo. Od prikazanih olimpijskih iger so največ poročali o poletnih olimpijskih igrah v Atenah.

Na sliki 3.2 lahko spremljamo medijsko pozornost namenjeno *Bushu* in *Clintonu*. Že preden *Bush* zamenja *Clintona* v Beli hiši *Bush* glede na pojavitveno frekvenco dominira.

Na prikazu slovenskih politikov (slika 3.3) je presenetljivo, da med letoma 2003 in 2005, ko je bila na oblasti 7. vlada brez *Janeza Janše* v koaliciji, temu ni občutno upadla frekvenca kot bi lahko pričakovali. Poleg tega pa je zanimiva *Erjavčeva* krivulja. Ta je nižja v času, ko je postal minister za



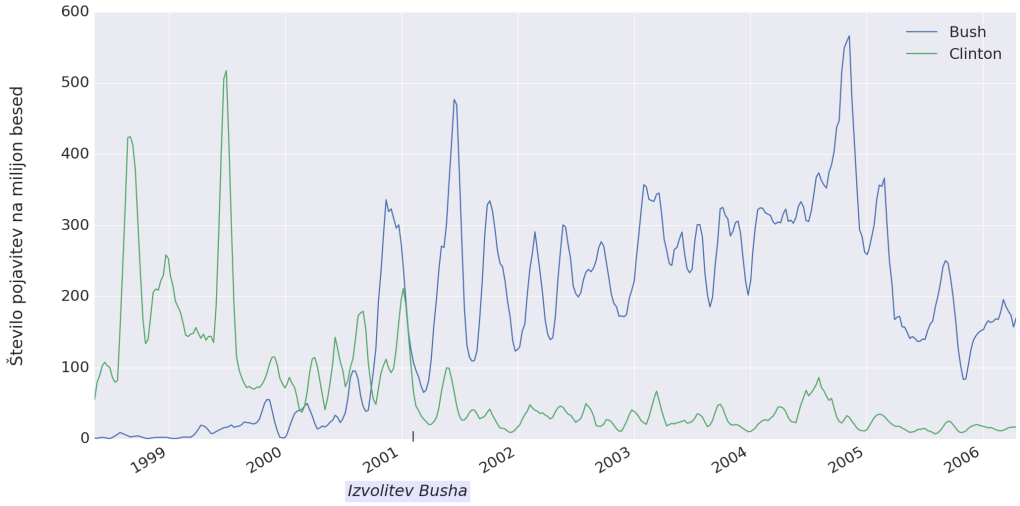
Slika 3.1: Prikaz pojavitvene frekvence besede *olimpijski*

obrambo (2004), kot leta 2005, ko je postal še predsednik stranke.

Na sliki 3.4 lahko vidimo, kako sta Ameriški invaziji na *Irak* in *Afganistan* vplivali na pojavitveno frekvenco teh dveh bližnjevzhodnih držav. Zanimivo je, da krivulja *Afganistana* kmalu po začetku vojne naglo pade. Izgleda, kot da je vso medijsko pozornost prevzelo dogajanje v *Iraku*.

Na sliki 3.5 lahko primerjamo medijsko odmevnost epidemij *ptičje gripe* in *norih krav (BSE)* v primerjavi z vedno prisotnim *virusom HIV*. Slednji ima konstantno gledano višjo frekvenco, ki pa se z izbruhom obeh epidemij seveda ne more primerjati. Pri virusu *ptičje gripe* opazimo, da je največ prahu dvignilo, ko je bil februarja leta 2006 potrjen prvi primer okužbe na slovenskih tleh. Vrh je precej višji kot pri prvi pojavitvi *norih krav* v Sloveniji novembra 2001.

Primerjali smo tudi medijsko popularnost nekaterih svetovnih osebnosti (slika 3.6). *Bush* je gledano v celoti najbolj popularen, vendar pa *papeževa* krivulja ob njegovi smrti aprila 2005 preseže *Bushevo* maksimalno frekvenco. *Armstrongova* frekvenca lepo niha in svoje vrhove dosega ob njegovih zmagah na tekmovanju Tour de France. Krivulji pevke *Madonne* in pa še bolj igralca *Brada Pitta* sta tik ob ničli, mediji te vrste jima ne namenjajo večje



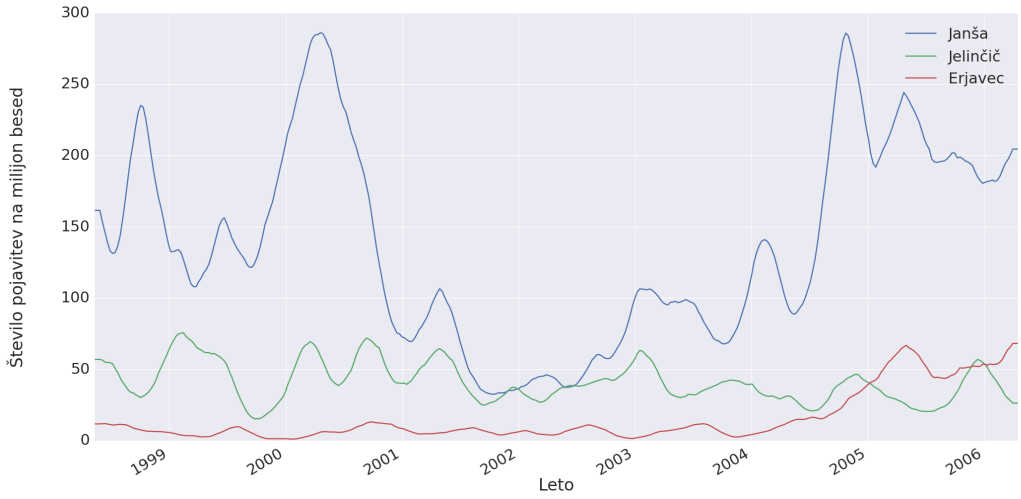
Slika 3.2: Prikaz pojavitvene frekvence besed *Bush* in *Clinton*

pozornosti v primerjavi z ostalimi omenjenimi.

Na grafu s slike 3.7 nas je zanimalo kateri ekipni šport je v Sloveniji najpopularnejši. Pričakovano se največ piše o *nogometu*, sledi mu *košarka* in *hokej*. Med najbolj izstopajoče dogodke gotovo sodi svetovno prvenstvo v nogometu leta 2002, kamor se je prvič v zgodovini uvrstila tudi slovenska reprezentanca.

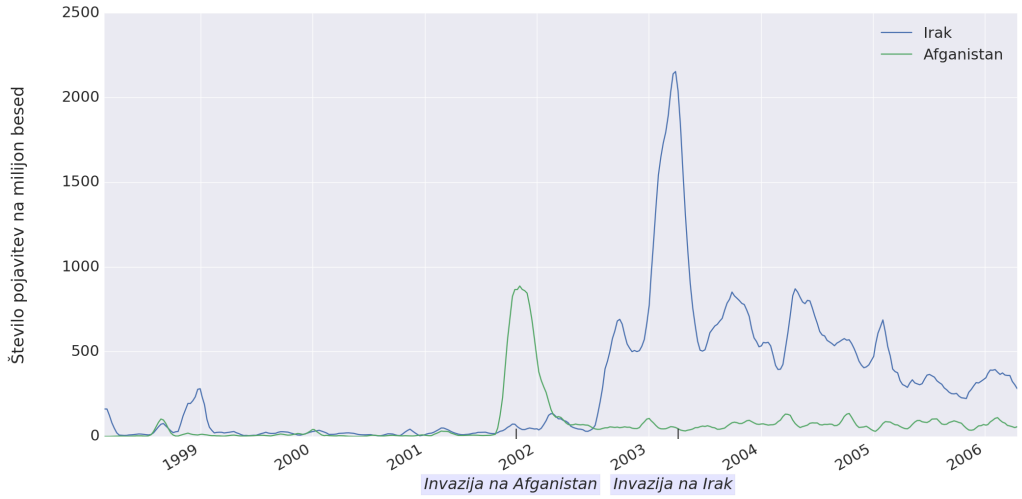
Slika 3.8 prikazuje frekvence nekaterih priznanih slovenskih umetnikov. Osebnosti so tokrat prikazane z bi-grami (oz. polnimi imeni). To se odraža v nizkih frekvencah, saj so v besedilih pogosto naslovljeni le s priimki, vendar pa je dobra stran tega gotovost, da gre res za želeno osebo. Od prikazanih imen izstopa operni pevec *Lotrič*. Pisatelj *Pahor* se mu približa leta 2003 po prejemu nagrade zlati sv. Just v Trstu. O *Mileni Zupančič*, se je največ pisalo leta 1999, ko je prejela Borštnikov prstan, o *Svetlani Makarovič* pa leta 2002 ob prejemu viktorja za življensko delo.

Na naslednjem prikazu (slika 3.9) lahko opazujemo, kako so teroristični napadi vplivali na pojav besede *terorist* v slovenskih časnikih. Zanimivo je primerjati krivuljo pred in po zloglasnim *11. septembrom*, ko sta bila porušena dvojčka v ZDA. Krivulja se je pred napadom, z izjemo posameznih

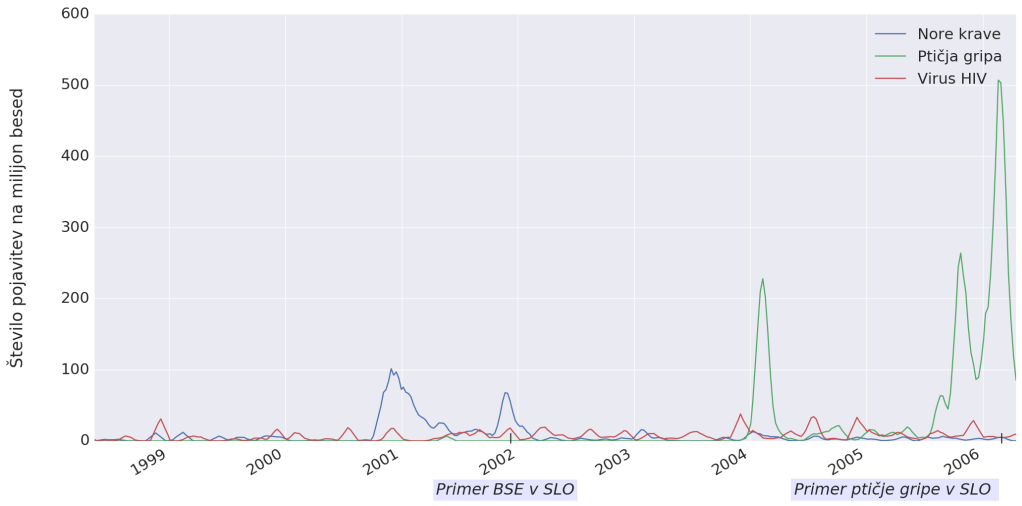


Slika 3.3: Prikaz pojavitvene frekvence besed *Janša*, *Erjavec* in *Jelinčič*

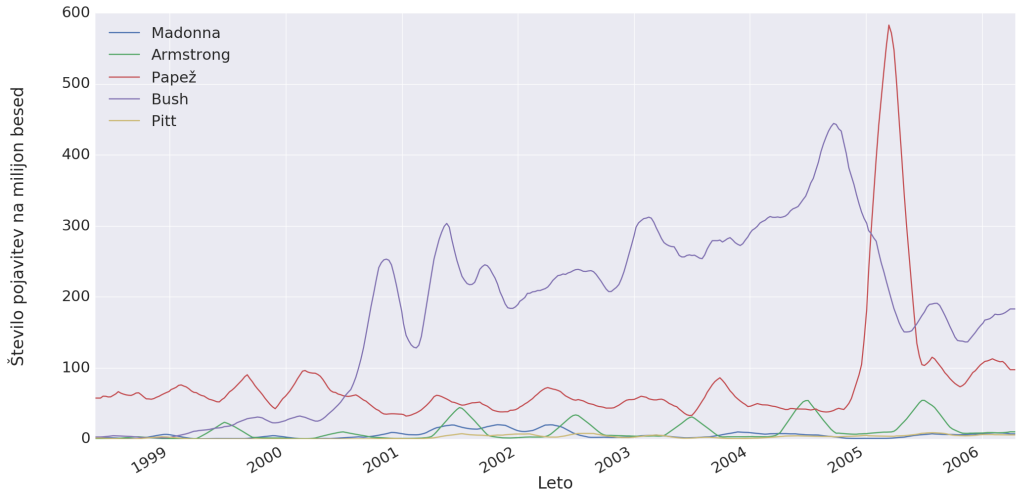
odstopanj, gibala okoli ničle, po njem pa je stacionirana veliko višje. Opazimo tudi, da na odmevnost dogodka močno vpliva njegova geografska pozicija. Vrh ob terorističnem napadu na od Evrope precej oddaljeno *Tanzanijo* je veliko nižji, kot ta ob napadu v *Londonu*, kljub temu da je ta v *Tanzaniji* terjal štiri krat več žrtev.



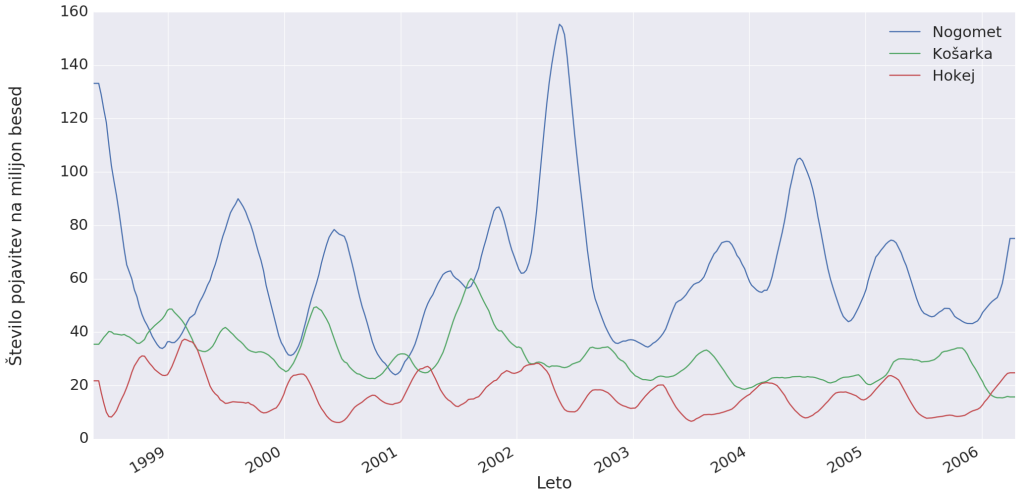
Slika 3.4: Prikaz pojavitvene frekvence besed *Irak* in *Afganistan*



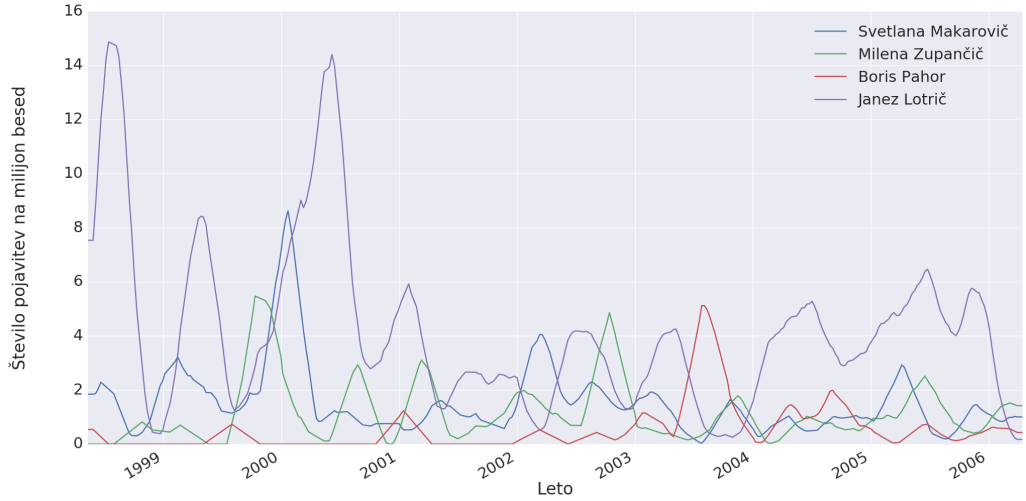
Slika 3.5: Prikaz pojavitvene frekvence besednih zvez *Ptičja gripa*, *Nore krave* in *Virus HIV*



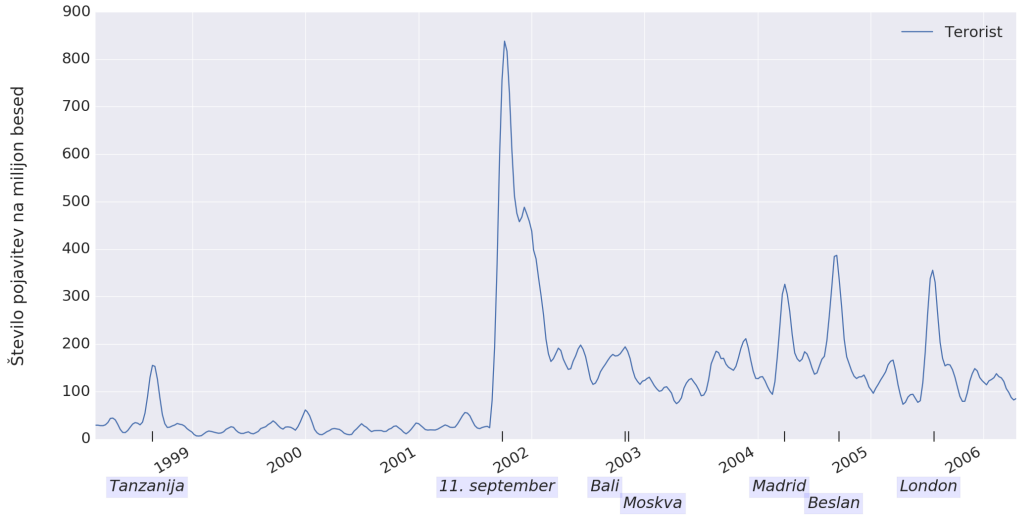
Slika 3.6: Prikaz pojavitvene frekvence besed *papež, Madonna, Pitt, Armstrong* in *Bush*



Slika 3.7: Prikaz pojavitvene frekvence besed *nogomet*, *košarka* in *hokej*



Slika 3.8: Prikaz pojavitvene frekvence imen *Milena Zupančič*, *Janez Lotrič*, *Svetlana Makarovič* in *Boris Pahor*



Slika 3.9: Prikaz pojavitvene frekvence besede *terorist*

Poglavje 4

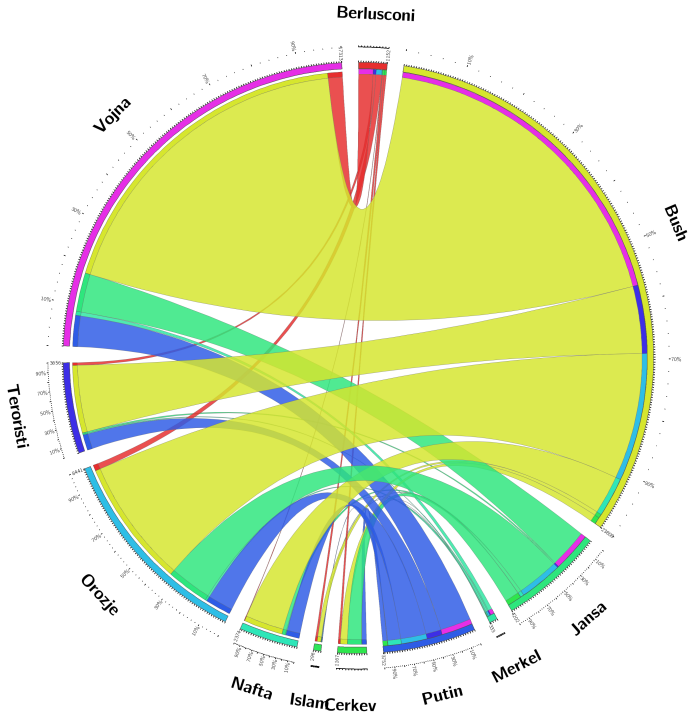
Prikaz sopojavitev besed

Pri preučevanju sopojavitev besed smo se ukvarjali predvsem z imeni in pojmi, za katere predvidevamo, da pogosto sodijo v isti kontekst z izbranimi imeni. Pri tem smo upoštevali, katera imena in dogodki so bili aktualni v letih, ko so bili objavljeni analizirani članki (1998-2006). Pri preučevanju nastalih diagramov je upotrebno upoštevati, da so besede lematizirane (beseda označena z *Šola* označuje skupek besed *šoli, šolo, šol itd.*), pri circosovih diagramih pa besede (zaradi omejitev vizualizacijskega orodja) ne vsebujejo šumnikov.

4.1 Prikazi s circos diagrami

Na sliki 4.1 smo prikazali sopojavitev nekaterih pojmov tekoče problematike s svetovnimi politiki. Takoj opazimo prevladujočo rumeno barvo, ki pripada *Bushevemu* segmentu. Slednji se je v kontekstu izbranih pojmov največkrat pojavil. Nekaterih segmetnov (npr. *islam*), pa zaradi prevlade ostalih skoraj ne opazimo. Zaradi večje preglednosti in ker nas bolj zanimajo razmerja povezav med segmenti kot pa sama velikost segmetnov (oz. skupno število pojavitev besede segmenta v izbranem kontekstu), so diagrami v nadeljevanju normalizirani.

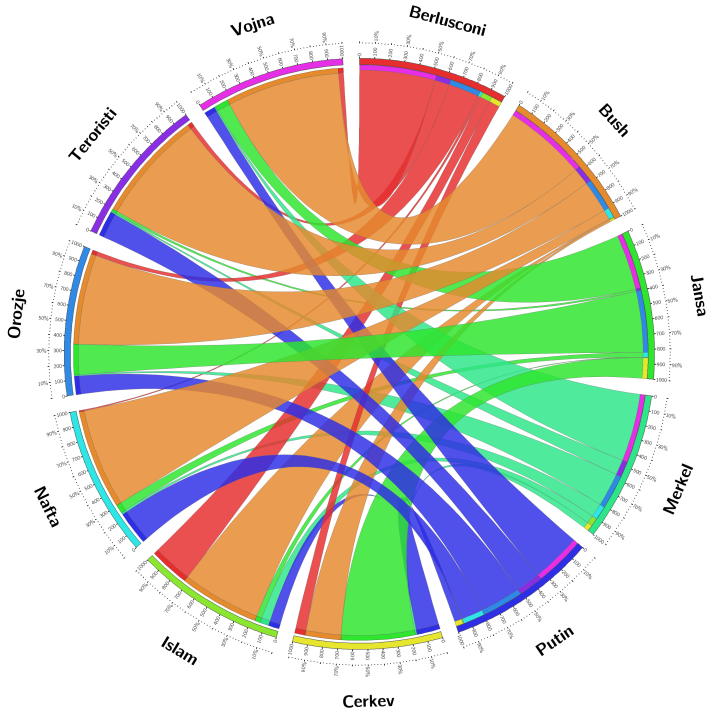
Diagram na sliki 4.2 vsebuje enake segmente kot pri prejšnjem primeru,



Slika 4.1: Prikaz sopojavitve besed na področju svetovne politike z nenormaliziranim diagramom

le da so ti tokrat normalizirani. Tudi tukaj je lepo vidno, da je s skoraj vsemi pojmi najbolj povezan *Bush*. Vendar pa so za razliko od prej razvidna tudi razmerja povezav v segmetnu *islam*. Pri tej sliki bi izpostavila segment *cerkev*, ki se od drugih razlikuje predvsem po tem, da najmočnejša povezava ne vodi k *Bushu*, temveč k slovenskemu politiku *Janezu Janši*. Predvidevam, da je vzrok enak temu, da imamo opravka s slovenskimi članki in je lahko pričakovati, da bo *cerkev* pogosteje omenjena v krogu domačih politikov.

Sledi diagram, ki vsebuje vplivnejša imena v slovesnki politiki, v povezavi s problematiko, ki je na naših tleh ves čas aktualna. Na sliki 4.3 je razvidno, da med slovenskimi politiki v besedilih dominira *Janez Janša*. To je pričakovano, saj se ta konstantno pojavlja v slovenskih medijih, o njem se govori, če je v vladi ali ne. Zanimivo je recimo, razmerje povezav v segmentu *Peterle*. *Cerkev* (rdeči del segmenta *Peterle*) zajema precej večji del pove-

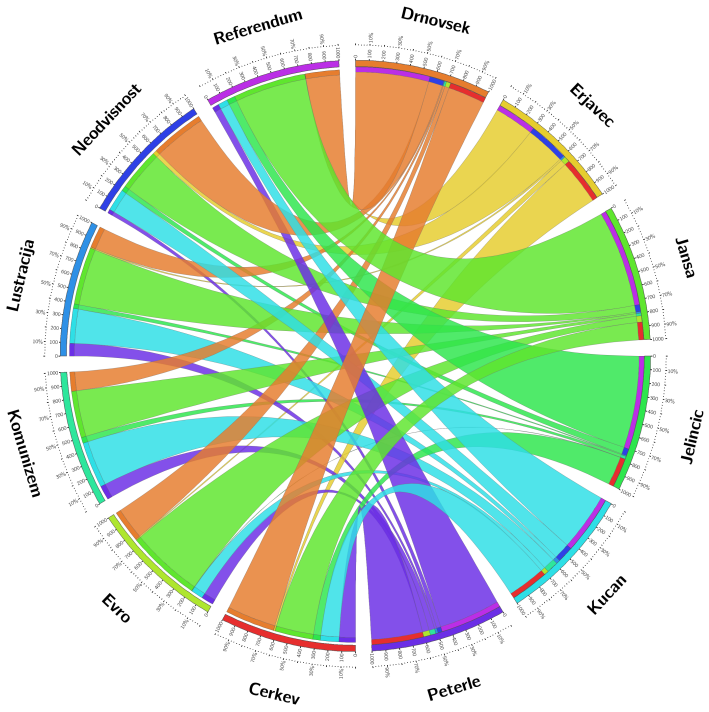


Slika 4.2: Prikaz sopojavitve besed na področju svetovne politike

zav kot pri ostalih politikih. To lahko utemeljimo s pripadnostjo Lojzeta Peterleta krščanski demokratski stranki. Zanimiva so tudi razmerja v segmetnu *neodvisnost*, kjer se *Janez Drnovšek*, eden ključnih mož pri slovenski osamosvojiti, skoraj izenači z *Janševim* deležem.

Na naslednjem diagramu na sliki 4.4 znova nastopajo slovenski politiki, dodaten člen je še slovenski škof *Franc Rode*. Tokrat nas zanima kako so osebnosti povezane med seboj. Zanimive so *Kučanove* povezave, med katerimi za spremembo ne prevldauje *Janševa*, temveč *Drnovškova*. Tudi *Rode* se očitno najpogosteje pojavlja ob *Drnovšku*.

Na sliki 4.5 je prikazan diagram s svetovnimi in slovenskimi vrhunskimi športniki. Ta je zelo pester, saj so razmerja med segmetni zelo raznolika. *Doping* je po pričakovanjih najbolj povezan z *Armstrongom*. Slednji je bil namreč prvič osumljen jemanja prepovedanih substanc že leta 1999 na tekmovanju Tour de France. Pri *Schumacherju* je zanimivo kako ogromen delež

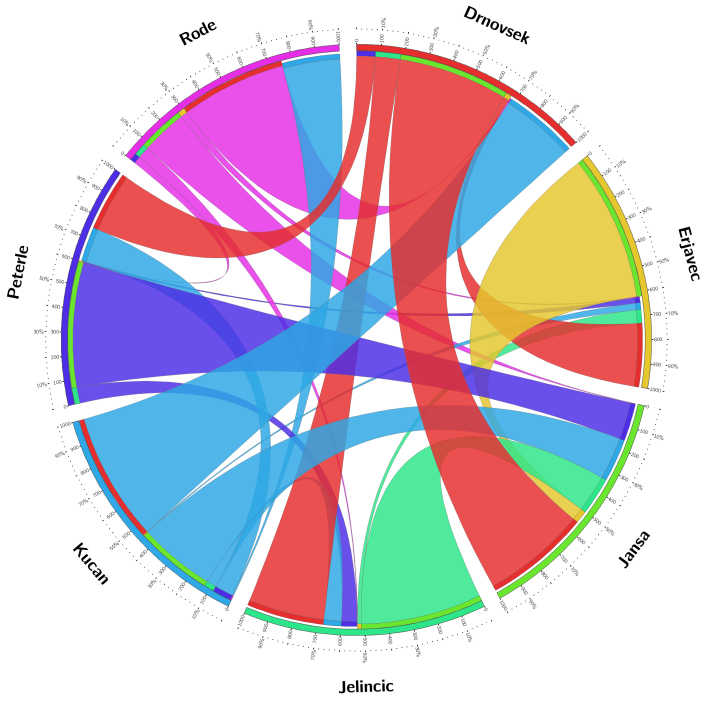


Slika 4.3: Prikaz sopojavitve besed na področju slovenske politike

povezav pripada segmentu *zmaga*. To je smiselno, saj je ta ob koncem svoje kariere (leta 2006) držal številne rekorde v Formuli 1, med drugi tudi največ zmag v eni sezoni. Njegov delež v segmentu *olimpijski* pa je pričakovano znatno majhen, saj Formula 1 ni olimpijska disciplina. Istopajoče je tudi, kakšen delež *medalje* pripada plavalcu Mankoču, ki je vrsto let dominiral v tekmovanjih v kratkih bazenih.

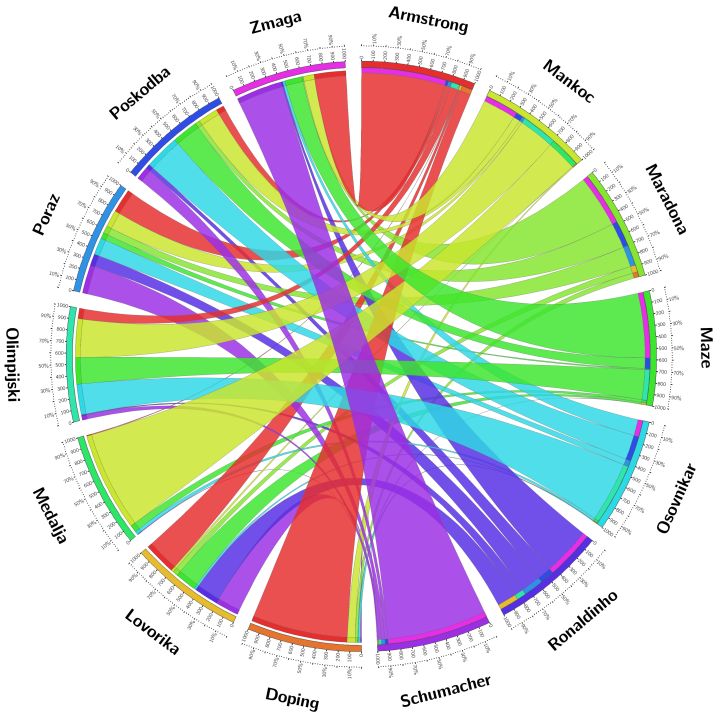
Na grafu na sliki 4.6 je mojo pozornost pritegnil rdeči del segmenta *Einstein*, ki pripada *bombi*. Velikost deleža besede, ki ni neposredno povezana s tem izjemnim fizikom je presenetljiva. Gre torej za to, da *Einstein* nehote pripomogel k nastanku atomske bobmbe in zanimivo je, da se fizik pogosteje pojavlja skupaj s pojmom *bomba* kot pa *vesolje* in *fizika*. Specifična je tudi obojestransko močna povezava *Hawking-vesolje*, ki kaže na to, da eden izmed svetovno najbolj uveljavnih fizikov pušča močan vtis tudi na slovenske medije.

Sledi prikaz osebnosti na sliki 4.7, ki so pogosto dobrodelno usmerjeni

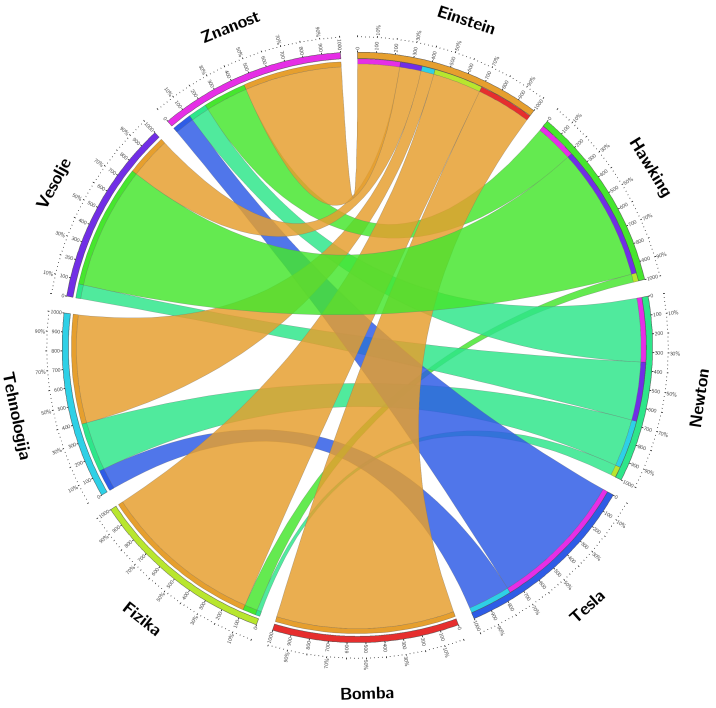


Slika 4.4: Prikaz sopojavitve osebnosti v slovenski politiki

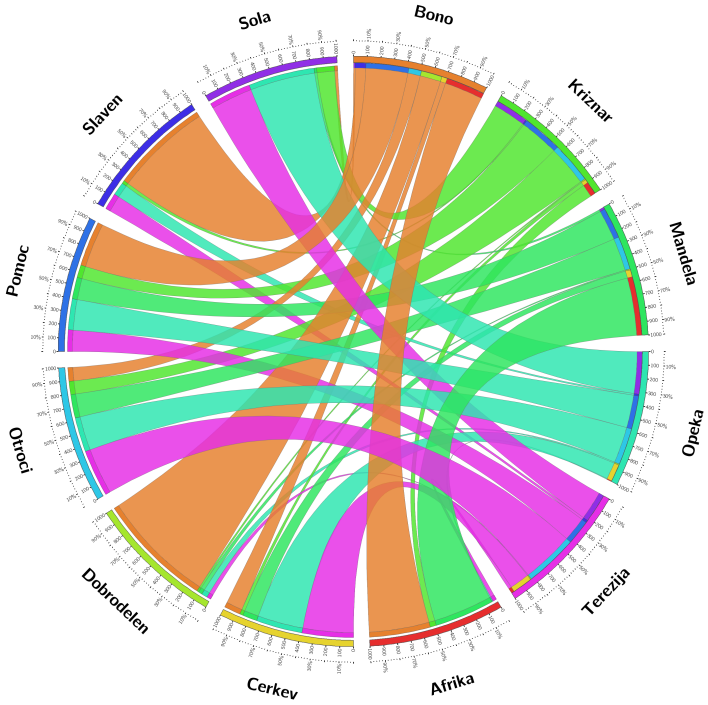
oz. pomoč drugim za njih predstavlja temelj njihovega delovanja. Čeprav so vse osebe povezane z dobrodelnostjo, pevec *Bono* prevladuje v segmentu *dobrodelen*. Predvidevamo lahko, da je to tudi posledica načina kako o sopomoči ljudem novinarji poročajo (*Bono* je velikokrat omenjen v povezavi z dobrodelnimi prireditvami in koncerti), to pa lahko vpliva na uporabo tega pridevnika. Pričakovana je močna obojestranska povezava *Mandela-Afrika* presenetljiv pa sorazmeroma majhen delež besede *cerkev* v kontekstu z slovenskim misjonarjem *Pedrom Opeko*.



Slika 4.5: Prikaz sopojavitve besed na športnem področju



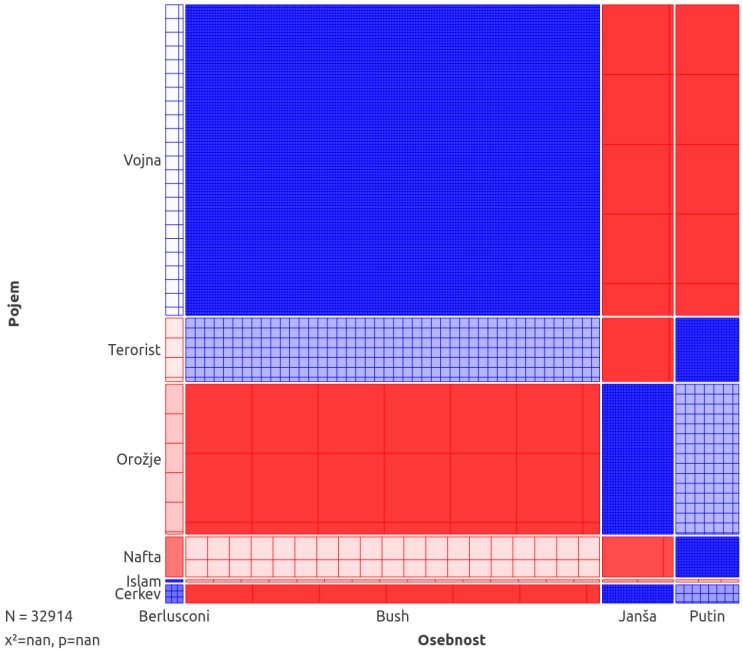
Slika 4.6: Prikaz sopojavitve besed na področju znanosti



Slika 4.7: Prikaz sopojavitve besed na področju dobrodelnega udejstvovanja

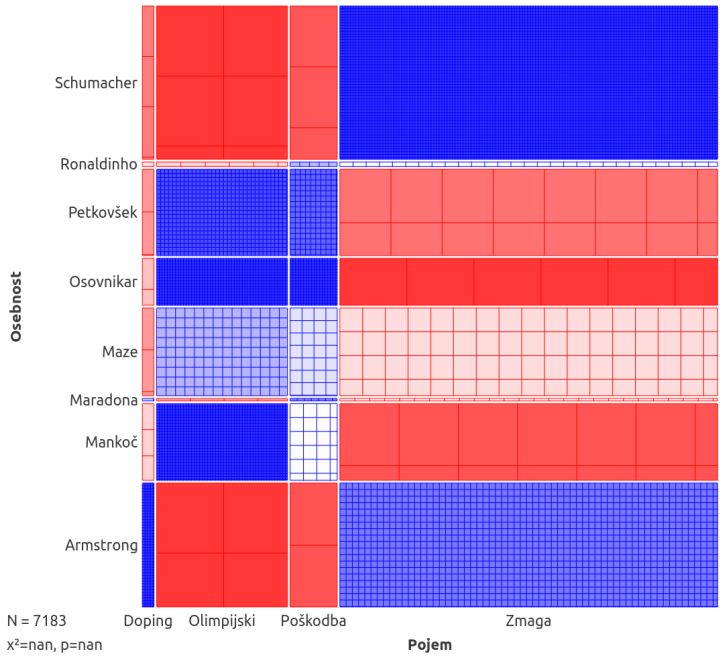
4.2 Sievov diagram

Diagrami v tem poglavju so po tematiki podobni prejšnjim circos diagramom. Nastopajo enaki pojmi, vendar v manjšem številu zaradi večje preglednosti diagramov. Sievovi diagrami nam ponujajo interpretacijo drugačne vrste. Jakost barvnega senčenja nam lepo nakaže, na kaj naj bomo še posebej pozorni in kaj mogoče ni tako pričakovano.



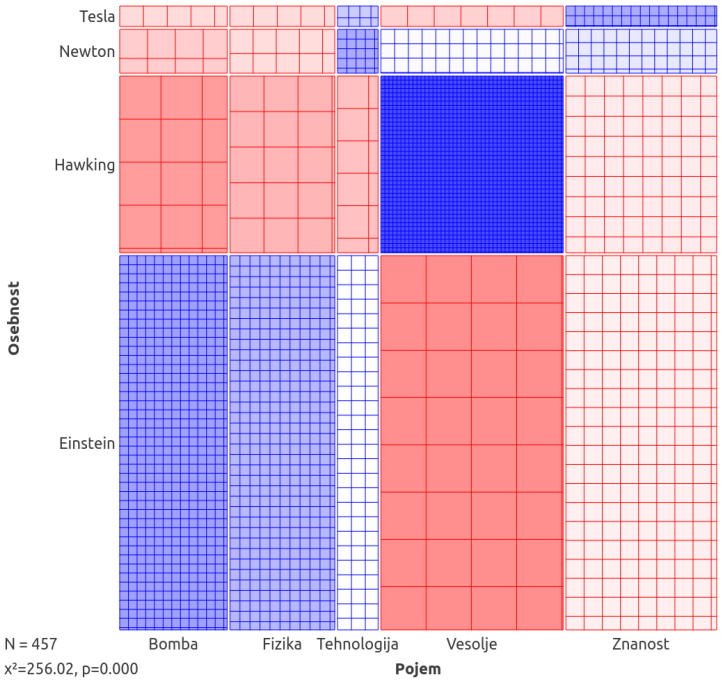
Slika 4.8: Prikaz sopojavitve besed na področju svetovne politike s sievovim diagramom

Prvi diagram na sliki 4.8 znova prikazuje svetovne politike v povezami z aktualno problematiko. Pri *Janši*, kot prej na prikazu s circosom, izstopa povezava s *cerkvijo*, dodatno pa nas temno modra barva celice *Orožje-Janša* opozarja na neko ne tako očitno povezavo. Zanimivo, da je *orožje* z *Janšo* prikazano kot veliko močneje kot z *Bushem*, za katerega bi lahko rekli, da z orožjem sploh ni povezan glede na pričakovano frekvenco. *Nafta* in *terorist* sta glede na prikaz najbolj povezana s Putinom.



Slika 4.9: Prikaz sopojavitve besed na področju športa s sievovim diagramom

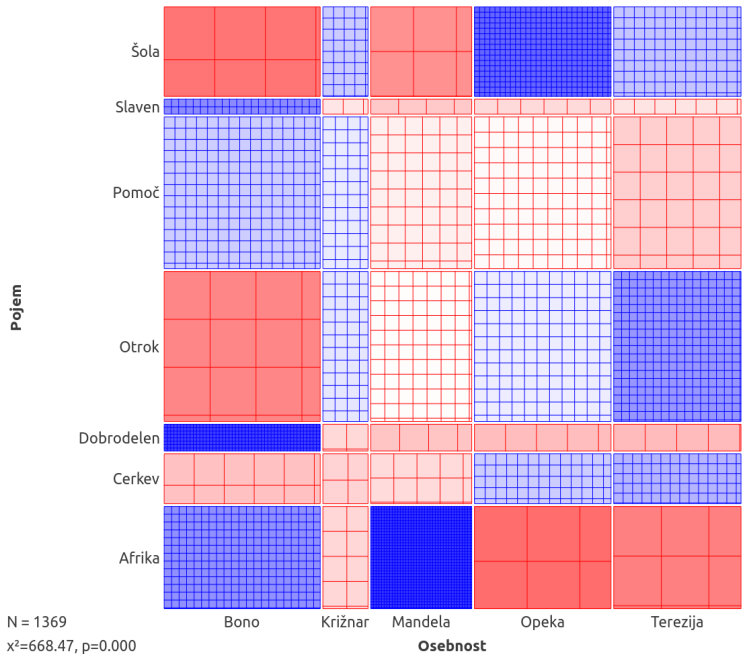
Sievov prikaz sopojavitev besed na športnem področju je tako kot na circos diagramu zelo pestra in barvita tudi tukaj (slika 4.9). V povezavi z *dopingom* se *Armstrongu* pridružuje tudi *Maradona*, ki je bil tarča ene največjih dopinških afer v zgodovini športa. Ta nogometna legenda je kot kaže precej povezana tudi z besedo *poškodba*. To bi morda lahko povzročila odmevna in izjemno groba poteza španskega igralca, ki je Maradoni zdrobil gleženj in ga prisilila počivati 8 mesecev.



Slika 4.10: Prikaz sopojavitve besed na področju znanosti s sievovim diagramom

Prikaz na sliki 4.10 znova poudarja močno povezanost *Hawkinga* z *Vesoljem*, *Einsteina* z *Bombo* ter *Tesle* s *Tehnologijo*.

Pri naslednjem diagramu (slika 4.11) preseneča rdeča barva celice *Opeka-Afrika*, saj ta slovenski misijonar že vrsto let deluje v Afriki. Je pa zanj značilna povezava s *šolo*, kar je pričakovano.



Slika 4.11: Prikaz sopojavitve besed na področju dobrodelnega udejstvovanja s sievovim diagramom

Poglavje 5

Sklepne ugotovitve

V diplomski nalogi smo s preučevanjem besed in besednih zvez v novičarskih člankih skušali pridobiti vpogled nad dogodki skozi čas ter preiskovali povezanost nekaterih osebnosti z njim sorodnimi pojmi. Rezultati so smiselni in zanimivi. Povečane frekvence pojavitev besed se kronološko ujemajo s pripadajočimi fenomeni in nam ponujajo zanimive primerjave dogodkov in oseb glede na njihovo medijsko odmevnost. Dobljene povezave med imeni in pojmi so do neke mere pričakovane, vendar kljub temu nekateri primeri prikazujejo nekoliko presenetljiva razmerja. Pokazali smo, da je mogoče študije na besedilih iz novičarskih člankov ponoviti tudi na slovenskih besedilih. Circosovi diagrami pa pričajo o tem, da vizualizacija circos na pregleden in zanimiv način lahko prikazuje tudi relacije med besedami, torej njegova uporaba ni omejena le na podatke iz bioinformatike.

V namen analize smo razvili programsko kodo v jeziku Python, ki sestoji iz treh glavnih sklopov. V prvem delu koda zajema dostop do besedil iz podatkovne baze ter predobdelavo teksta. Besedila so razčlenjena na besede (tokenizacija), te pa so po potebi lematizirane. Sledi implementacija štetja n-gramov in sopojavitev besed v premičnem oknu. V tretjem sklopu rezultate pretvorimo v ustrezno obliko za vizualizacijo s sievovim in circos orodjem ter izrišemo in ustrezno opremimo grafe pojavitvenih frekvenc besed.

Programska koda skupaj obsega 450 vrstic[1].

Uporaba tovrstnih raziskav, ki smo jo poskušali izvesti tudi sami, je zelo široka. Predvidevamo, da bi novinarske hiše zanimalo o čem statistično gledano največ poročajo, kateri dogodki so bili v časopisih bolj odmevni in kateri takoj pozabljeni. Poleg tega pa bi lahko računalničarji v sodelovanju s specialisti iz ostalih strok odgovarjali na številna vprašanja, ki se porajajo v povezavi z slovensko literaturo, kulturo in družbo.

Seveda se na tem področju, da storiti še veliko. Ena izmed izboljšav bi bila povečati časovni razpon korpusa. S tem bi lahko preučevali trenutno aktualne dogodke, nad nekatero vedno prisotno problematiko pa bi imeli še boljši vpogled. Kronološko razsežnejši korpus bi morda omogočil tudi opazovanje sprememb v slovenskem jeziku.

[1] `https://github.com/zadnjipuki/Analiza-spletnih-novic.git`

Literatura

[1] J.-B. Michel, Y. K. Shen, A.P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P.Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(10):176–182, 2010.

[2] K. H. Leetaru. Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, 16(9), 2011.

[3] A. M. Petersen, J. Tenenbaum, S. Havlin, and H. E. Stanley. Statistical laws governing fluctuations in word use from word birth to word death. *Nature*, 2(313):313–321, 2012.

[4] I. Flaunas, O. Ali, T. Lansdal-Welfare, T. D. Bie, N.Mosdell, J. Lewis, and N. Cristianini. Research methods in the ages of digital journalism. *Digital Journalism*, 1(1):102–116, 2012.

[5] S. Sudhahar, G. A. Veltri, and N. Cristianini. Automated analysis of the us presidential elections using big data and network analysis. *Big Data & Society*, 2(1):1–28, 2015.