UNIVERZITET U NOVOM SADU
PRIRODNO-MATEMATIČKI FAKULTET
DEPARTMAN ZA MATEMATIKU I INFORMATIKU

Analiza mreže reddit zajednica

- seminarski rad -

Ime i prezime: Ognjen Kulić

Broj indeksa: 53m/19

Sadržaj

1. Uvod

Ovaj rad prikazuje rezultate istraživanja mreže reddit[1] zajednica. reddit je popularan website na kojem korisnici mogu da postavljaju, ocenjuju i komentarišu razni sadržaj. Korisnici mogu da naprave zajednice za specifične teme sadržaja, poznate kao subreddit-i. U postavljenom sadržaju mogu da se nađu i link-ovi ka nekom drugom subreddit-u. Glavni cilj istraživanja je da pronađe zajednice subreddit-a u kojima se subreddit-i međusobno link-uju. Kako sentiment link-a ka drugom subreddit-u može da bude pozitivan, neutralan ili negativan, potrebno je posebno pronaći zajednice subreddit-a gde su odnosu između subreddit-a pozitivni ili neutralni, a posebno za grupe subreddit-a koji imaju česte negativne odnose.

U okviru ovog istraživanja se takođe postavljaju i pitanja koji subreddit-i najviše link-uju druge subreddit-e u zavisnosti od sentimenta link-a, koji subreddit-i su najviše link-ovani od stane drugih subreddit-a u zavisnosti od sentimenta link-a, da li subreddit-i sa sličnim interesima pripadaju istim zajednicama.

2. Prikupljanje podataka

Za ovo istraživanje su preuzeti podaci sa dva različita izvora. Podaci o subreddit-ima su preuzeti sa pushshift.io[2], sajta koji se specijalizuje u big data analizu podataka sa socijalnih mreža. Ovaj skup podataka ima podatke o 763058 najpopularnijih subreddit-a. Za svaki subreddit postoje 62 atributa, među kojima su naziv subreddit-a, broj prijavljenih korisnika, datum kreiranja i jezik. Podaci o link-ovima između subreddit-a su preuzeti od Stanford Network Analasys Project-a[3]. Podaci su prikupljeni između januara 2014. godine i aprila 2017. godine za istraživanje o iterakcijama i konfliktima između web zajednica[4]. Ovaj skup podataka sadrži 858490 instanci link-ovanja između dva subreddit-a, i ima podatke o tome koji subreddit je postavio link, ka kojem subreddit-u vodi link, ID post-a u kojem je link, kada je postavljen link, koji je sentiment link-ovanja i vektor sa raznim podacima o tekstu post-a u kojem je postavljen link.

3. Obrada podataka

Prilikom prvog pregleda skupa podataka subreddit-a, utvrđeno je da postoje velike količine nedostajućih vrednosti za određene atribute. Zbog toga prvi koark čišćenja podataka je bio ukljanjanje tih kolona. Izbačena je 21 kolona. Pošto su svi atributi učitanog skupa podataka bili tipa **String**, drugi korak se sastojao iz konverzije tih kolona u odgovarajući tip podataka. Kolone **created_utc** i **retrieved_on**, koje predstavljaju kada je subreddit kreiran i kada su informacije o tom subreddit-u prikupljene , su pretvorene u vremenski tip podataka, dok je kolona **subscribers** koja predstavlja broj prijavljenih korisnika na subreddit pretvorena u brojevni tip podataka.

U pregledu podataka o link-ovima između subreddit-a nijedan atribut nije imao nedostajući vrednost, pa nije bilo potrebe da se izvrši uklanjanje atributa ili instanci. Pošto su podaci o link-ovima razdvojeni u dva skupa podataka, jedan gde su link-ovi u naslovu a drugi gde su link-ovi u telu postavljenog sadržaja, trebalo je da se oba skupa podataka spoje u jedan. Kako

su atributi oba skupa podataka bili potpuno isti, urađena je prosta unija skupova podataka. Isto kao i kod podataka o subreddit-ima, bilo je potrebno da se određeni atributi konvertuju u odgovarajući tip podataka. U ovom slučaju je konvertovana samo jedna kolona, **TIMESTAMP** koja označava kad je postavljen sadržaj sa link-om ka drugom subreddit-u.

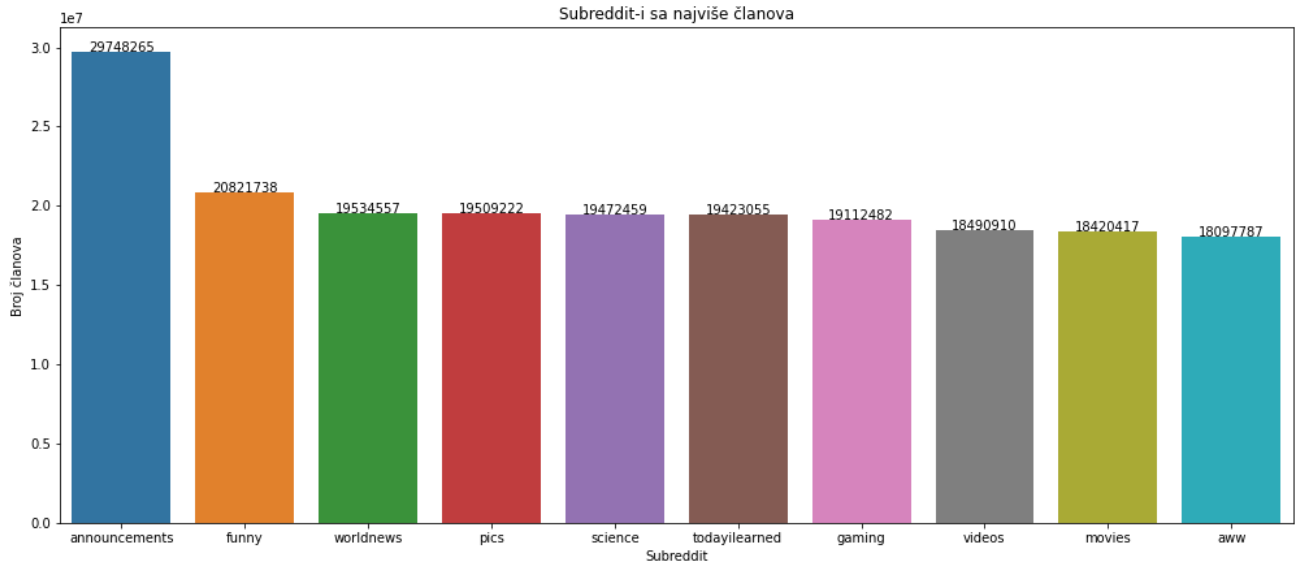
Pošto se analiza mreže radila pomoću GraphFrames-a[5], potrebno je pronaći i preimenovati kolone koje određuju naziv čvora, izvorni čvor i ciljni čvor kako bi GraphFrames mogao korektno da napravi graf. Kolona **display_name** iz podataka o subreddit-ima je preimenovana u **id**, dok su kolone **SOURCE_SUBREDDIT** i **TARGET_SUBREDDIT** preimenovane u **src** i **dst**.

Sa obrađenim podacima o subreddit-ima i link-ovima je napravljen usmeren graf gde subreddit-i predstavljaju čvororve, a link-ovi između njih predstavljaju usmerene grane. Odmah posle pravljenja grafa su uklonjeni izolovani čvorovi. Tako je uklonjen 743761 čvor, ostavljajući 19297 čvorova u grafu.

Pošto je potrebno posebno analizirati zajednice subreddit-a koje su povezane link-ovima sa pozitivnim ili neutralnim sentimentom, a posebno zajednice povezane link-ovima sa negativnim sentimentom, potrebno je da se pronađu dva odgovarajuća podgrafa sa ovim osobinama. Podgraf sa pozitivnim ili neutralnim zajednicama se dobio tako što su filtrirane grane grafa tako je vrednost **LINK_SENTIMENT**-a bila **1**, a podgraf sa negativnim zajednicama je dobijem filtriranjem grana gde je vrednost **LINK_SENTIMENT**-a bila **-1**. Pošto je prilikom filtriranja grana moguće da neki čvorovi postanu izolovani, u oba slučaja je opet izvršeno odbacivanje izolovanih čvorova. Posle ove obrade podgraf koji se sastojao od link-ova sa pozitivnim ili neutralnim sentimentom se sastojao od 19060 čvorova i 776278 grana, dok se podgraf sa link-ovima negatinog sentimenta sadržao od 4359 čvorova i 82210 grana.

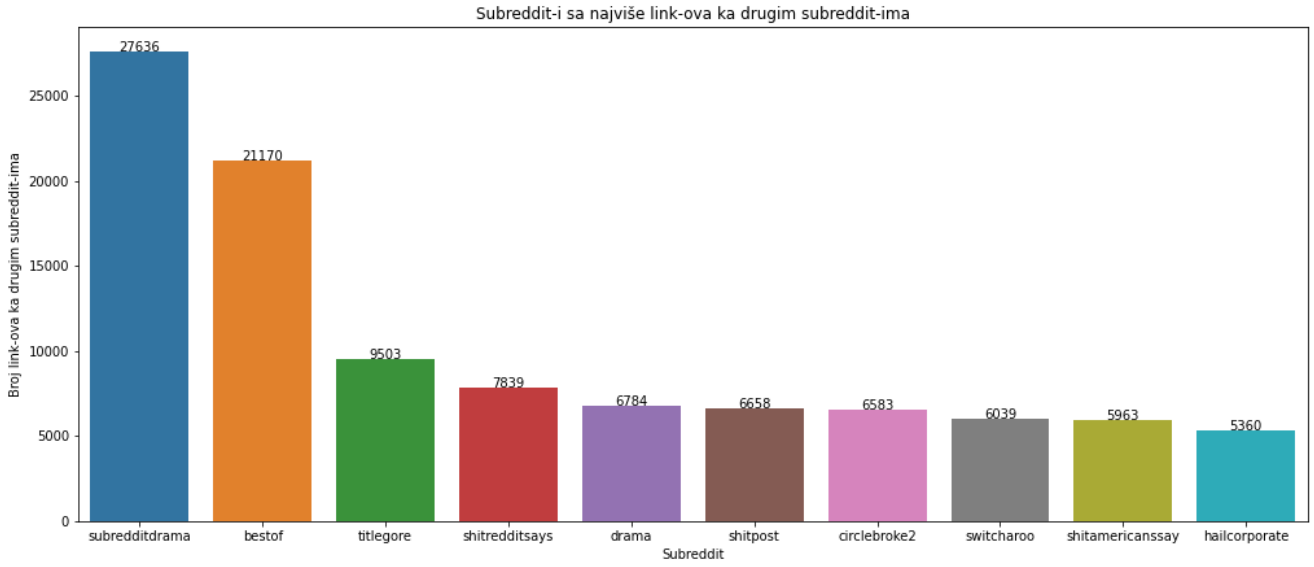
4. Eksplorativna analiza

Raspon veličine subreddit-a u mreži ide od 0 do 29748256 članova. Postoji 12 subreddit-a bez ijednog člana. Subreddit-i sa najviše članova su prikazani na slici 4.1. Prosečan broj članova subreddit-a je 47781,65. Najstariji subreddit je **nsfw**, osnovan 19.1.2006., dok je najmlađi u ovom skupi podataka **plantschemistry**, osnovan 29.4.2017.

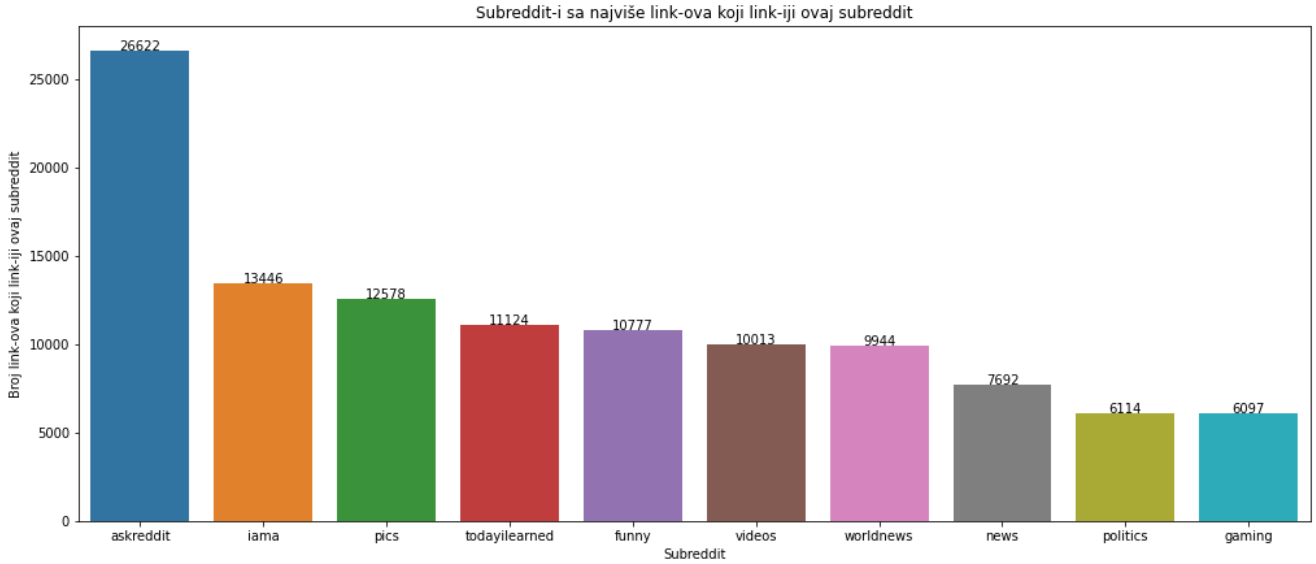


Slika 4.1. – Subreddit-i sa najviše članova

U periodu u kojem su prikupljani podaci o link-ovima, subreddit u proseku ima 15,37 link-ova prema drugim subreddit-ima i 24,83 link-ova ka sebi. Subreddit-i sa ukupno najviše link-ova ka drugim subreddit-ima su prikazani na slici 4.2. Nije iznenađujuće da se među njima nalaze subreddit-i kao što su **subreddit drama**, **bestof** i **shitredditsays** gledajući da takvi subreddit-i služe da sakupe ili diskusuju određene vrste sadržaja postavljenog na druge subreddit-e. Na slici 4.3. su prikazani subreddit-i koji imaju najviše link-ova ka sebi. Ovde mogu da se vide subreddit-i kao što su **pics**, **videos**, **worldnews**, **news** i **politics** koji imaju sadržaj koji je pogodan za dalje širenje i diskusiju na drugim subreddit-ima.

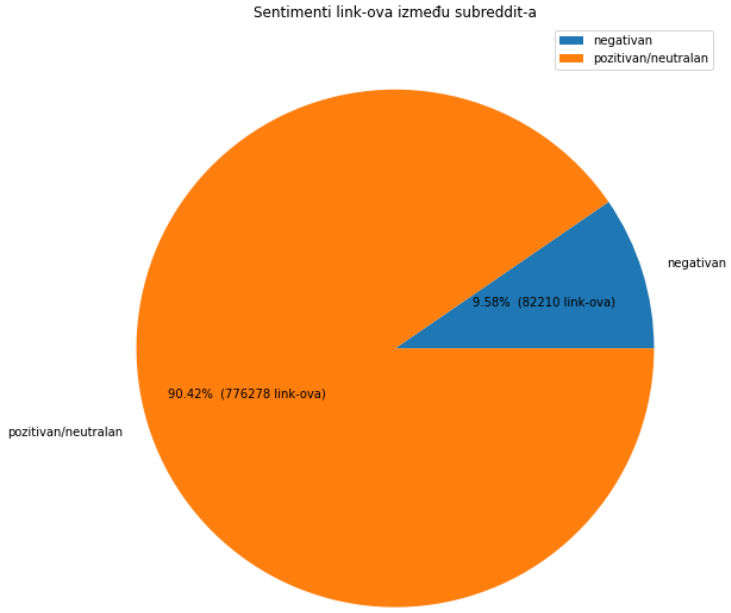


Slika 4.2. – Subreddit-i sa najviše link-ova ka drugim subreddit-ima



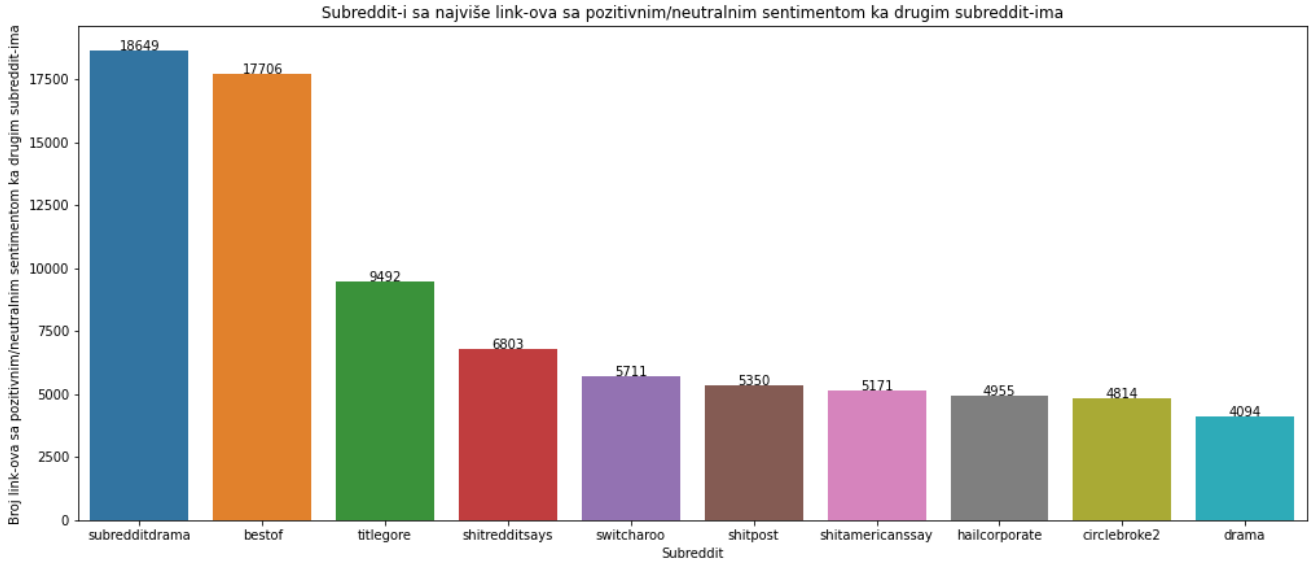
Slika 4.3. – Subreddit-i sa najviše link-ova ka sebi

Kada su u pitanju link-ovi između subreddit-a, većina link-ova je postavljena sa pozitivnim ili neutralnim sentimentom. Ovo može da se vidi na slici 4.4.

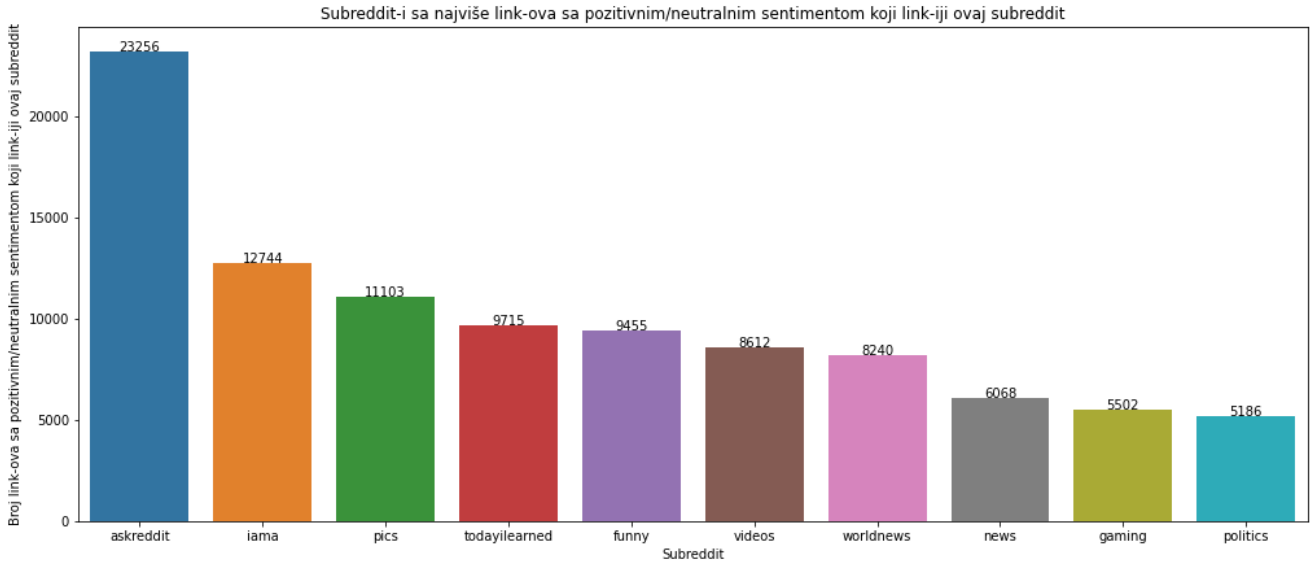


Slika 4.4 – Sentimenti link-ova itmeđu subreddit-a

Od subreddit-a koji imaju bar jedan pozitivan ili neutralan link, prosečan broj pozitivnih ili neutralnih link-ova ka drugim subreddit-ima je 14,19, a prosečan broj pozitivnih link-ova ka njemu je 22,86. Subreddit-i sa najviše pozitivnih ili neutralnih link-ova prem drugim subreddit-ima i ka sebi su isti, ali sa malo drugačijim rasporedom (slike 4.5 i 4.6).

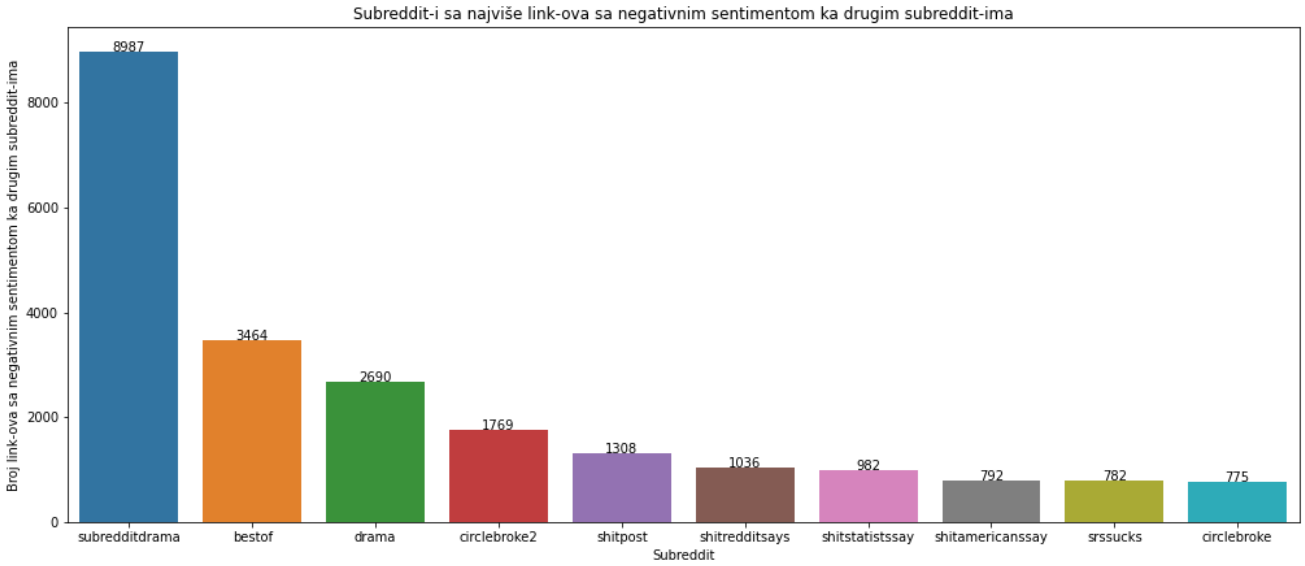


Slika 4.5. – Subreddit-i sa najviše link-ova sa pozitivnim ili neutralnim sentimentom ka drugim subredditima

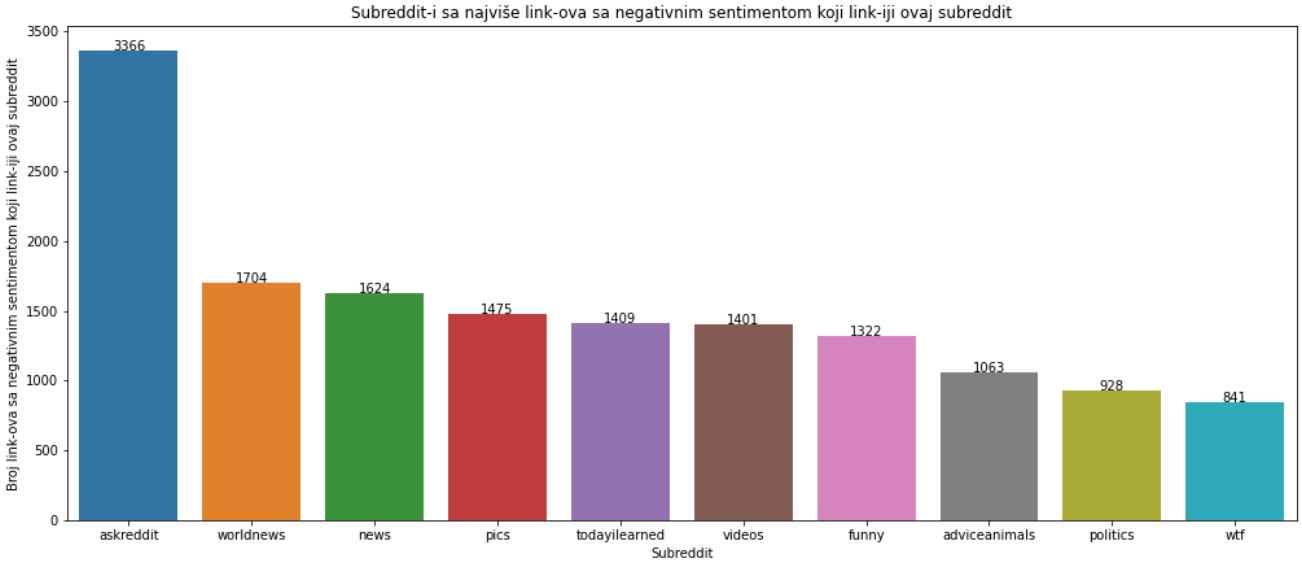


Slika 4.6. – Subreddit-i sa najviše link-ova sa pozitivnim ili neutralnim sentimentom ka sebi

Kad su u pitanju subreddit-i koji imaju bar jedan link sa negativnim sentimentom ka drugim subreddit-u, u proseku subreddit ima 9,45 link-ova sa negativnim sentimentom ka drugim subreddit-ima i 12,39 link-ova ka sebi. Među subreddit-ima sa najviše link-ova sa negativnim sentimentom ka drugim subreddit-ima se nalaze subreddit-i kao **subredditdrama**, **drama**, **shitredditsays** i **shitamericanssay**, koji se bave agregacijom i diskusijom drame i sadržaja drugih subreddit-a koji korisnici ovih subreddit-a smatraju kao loš ili "glupav" sadržaj (slika 4.7). Što se tiče subreddit-a koji imaju najviše link-ova sa negativnim sentimentom ka sebi, tu se opet stvar ne menja putno u odnosu na to kad se gledaju svi link-ovi (slika 4.8).



Slika 4.7. – Subreddit-i sa najviše link-ova sa negativnim sentimentom ka drugim subredditima



Slika 4.8. – Subreddit-i sa najviše link-ova sa negativnim sentimentom ka sebi

5. Pronalaženje zajednica subreddit-a

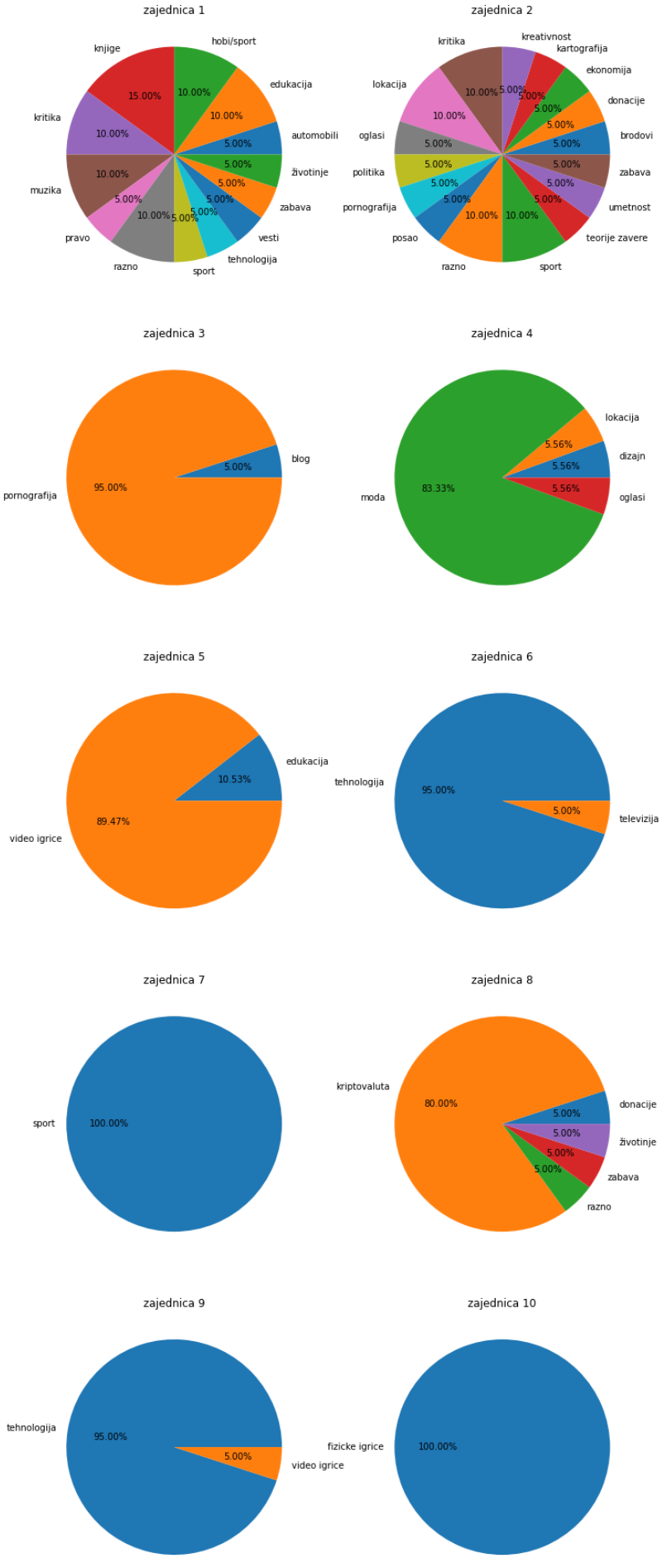
Kako bi se pronašle zajednice subreddit-a, nad grafovima je pušten Label Propagation algoritam [6] sa maksimalno 10 iteracija. U grafu sa svim link-ovima, nezavisno od sentimenta, pronađeno je 5402 zajednice. Od toga 4 875 zajednica se sastoje od samo jednog člana, što ostavlja 527 zajednica koje nisu trivijalne. Najveća zajednica se sastoji od 7733 člana, praćena zajednicama od 1614 i 153 člana. Sve ostale zajednice imaju manje od 100 članova.

Graf u kojem su link-ovi sa pozitivnim ili neutralnim sentimentom se sastoji od 5336 zajednica, gde su 4825 zajednica jednočlane i 513 zajednica sa više od jednog šlana. Najveća zajednica se sastoji od 8653 subreddit-a, druga po redu od 797 subreddit-a i treća od 154 subreddit-a. Ostale zajednice opet imaju manje od 100 članova.

U grafu gde su link-ovi sa negativnim sentimentom pronađeno je 1688 zajednica. Od toga 1527 zajednica ima samo jednog člana, a 161 zajenica se sastoji od više od jednog člana. Najveća zajednica ima 1210 članova, praćena zajednicom od 866 članova, dok ostale zajednice imaju manje od 100 članova.

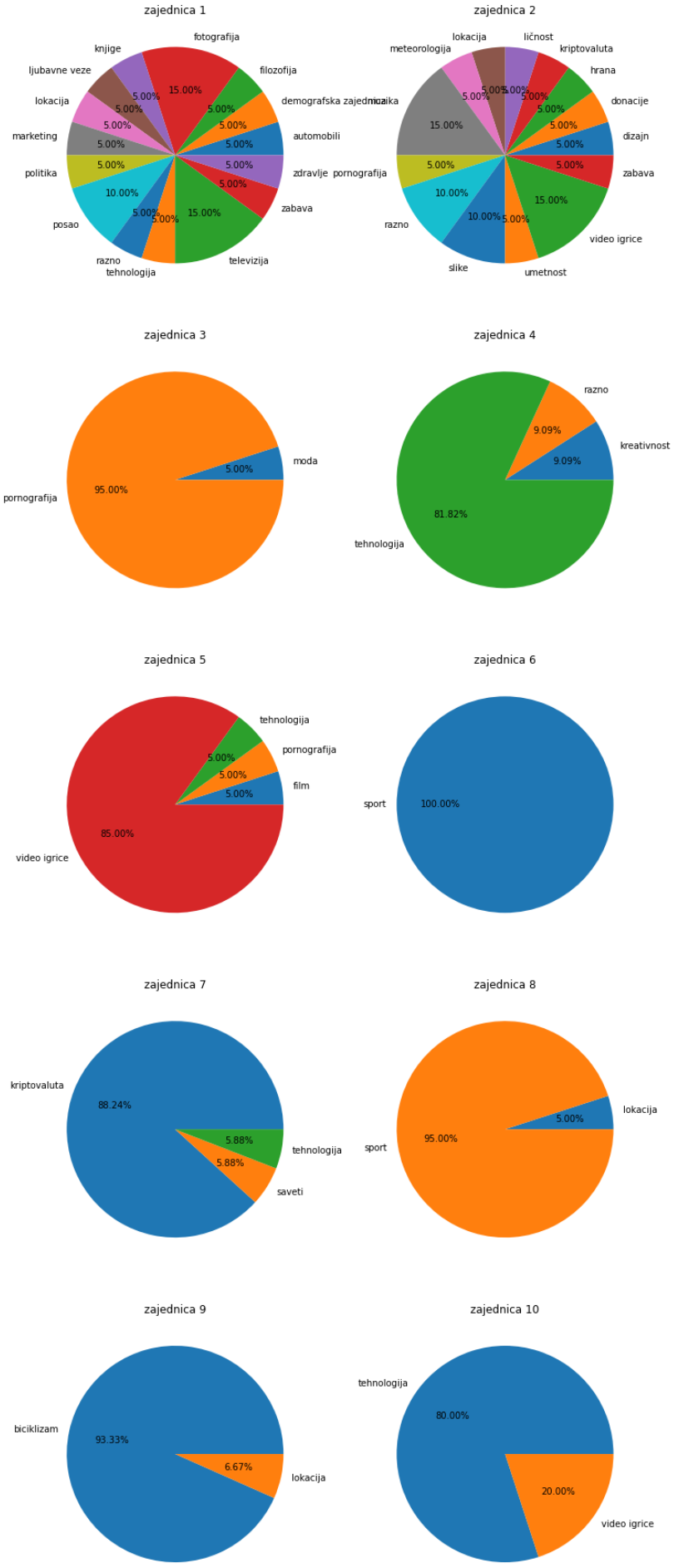
Kako bi se odgovorilo na pitanje da li subreddit-i u istoj zajednici imaju slične interese iz svakog od ova 3 grafa je uzeto 10 zajednica, 3 najveće i još 7 nasumično odabranih. Od svake od ovih zajednica je nasumično odabrano 20 subreddit-a kojima je dodeljena kategorija koja predstavlja opštu temu subreddit-a, na primer filmovi, muzika, politika. Kategorizacija se vršila ručno, pregledom opisa i sadržaja subreddit-a. Takođe je prilikom odabira opšte teme subreddit-a uzeta u obzir i zajenička tema sa ostalim subreddit-ima iz te zajednice, tako da ako na primer postoji subreddit za Japansku hranu i za Japanski jezik za oba subreddit-a je stavljeno da je opšta tema Japan. Iako postoji sajt, odnosno subreddit, koji grupiše subreddit-e preko 50000 članova po temi sadržaja subreddit-a [7], ti podaci ne postoje u obliku dostupačnog skupa podataka pa je zbog toga izvršena ručna klasifikacija odabranih subreddit-a, a pošto nije praktično da se ručno klasifikuje svih 19297 subreddit-a odlučeno je da se uzmu gorepomenuti uzorci.

Dobijeni rezultati za 10 zajednica iz grafa sa svim link-ovima su prikazani na slici 5.1. Zajednice 1 i 2, što su i najveće zajenice u grafu, se sastoje od raznolikih subreddit-a sa dosta različitim temama, tako da kod njih ne postoji jedna tema koja je zajednička za subreddit-e u tim zajednicama. Kod ostalih odabranih zajednica su stvari drugačije. U svakoj od preostalih zajednica postoji neka tema koju deli bar 80% subreddit-a u toj zajednici. U zajednici 3 to je pornografija sa 95% zajednica na tu temu, u zajednici 4 moda sa 83.33%, a u zajednici 5 video igrice (uglavnom Starcraft) sa 89.47%. Zajednice 6 i 9 se predominantno bave tehnologijom, prva više okrenuta ka software-u a druga ka hardware-u (specifično kompjuterima), sa 95% subreddit-a na tu temu u oba slučaja. Zajednica 8 se uglavnom bavi kriptovalutama sa 80% subreddit-a na tu temu, dok su zajednice 7 i 10 potpuno homogene pošto se u njima svi ispitani subreddit-i bave sportom (košarka) i fizickim igricama (Magic: The Gathering).



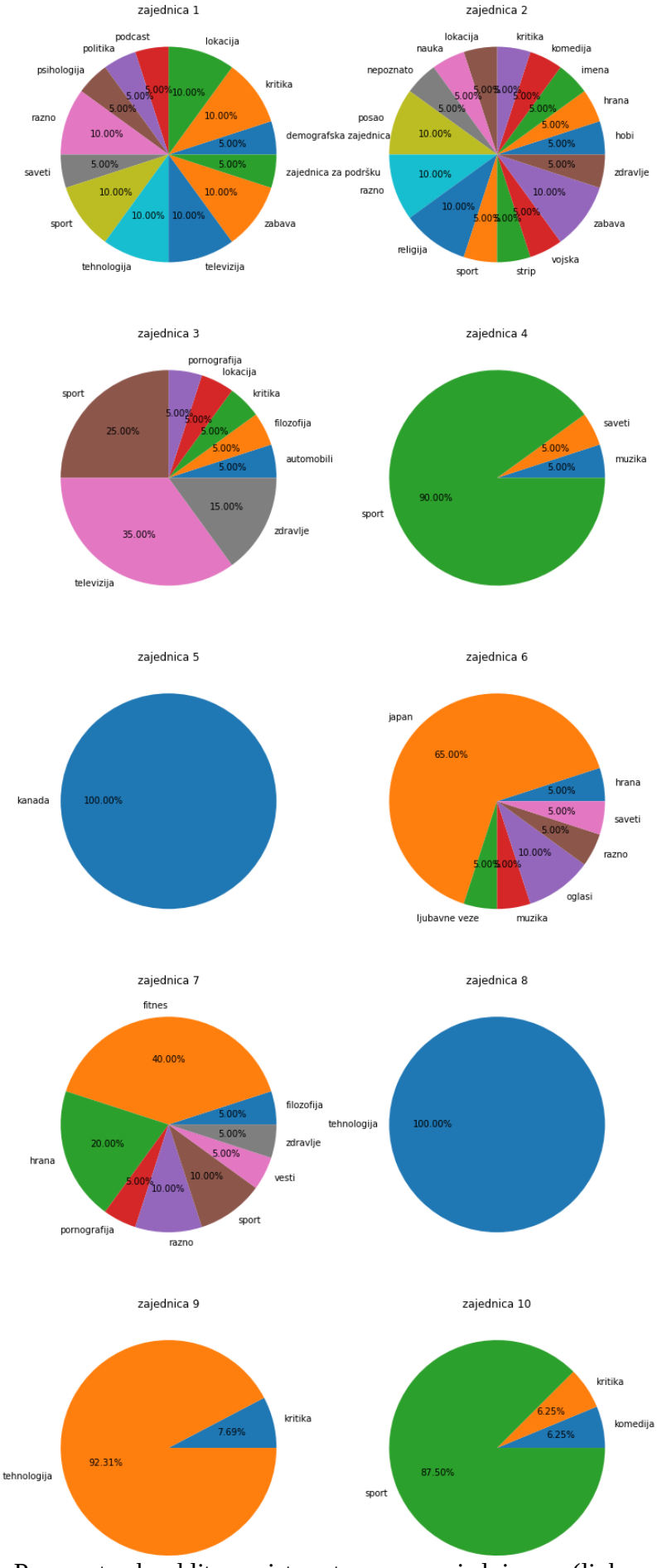
Slika 5.1 – Procenat subreddit-a sa istom temom u zajednici

Za zajednice gde su link-ovi između subreddit-a sa pozitivnim ili neutralnim karakterom, rezultati su prikazani na slici 5.2. Kao i kad su uzeti u obzir svi link-ovi, prve dve zajednice nemaju preovladavajuću zajedničku temu, dok ostalih 8 zajednica imaju zajedničku temu koja je zastupljena u bar 80% uzoraka. Zajednica 3 se opet predominantno bavi pornografijom sa 95% subreddit-a sa tom temom, zajednica 5 video igricama (World of Warcraft) sa 85%, zajednica 7 kriptovalutama sa 88.24%, a zajednica 9 rekreativnim biciklizmom sa 93.33%. Zajednice 4 i 10 se bave tehnologijom, prva više ka software-u a druga ka hardware-u (specifično kompjuterima), sa 81.82% i 80% subreddit-a na tu temu. Zajednice 6 i 8 se obe bave sportom, i to obe istim sportom – fudbalom. Zajednica 6 je potpuno homogena, do je zajednica 8 sa 95% subreddit-a na tu temu.



Slika 5.2 – Procenat subreddit-a sa istom temom u zajednicama (link-ovi sa pozitinim ili neutralnim sentimentom)

Što se tiče zajednica kod kojih su link-ovi sa negativnim sentimentom, rezultati su prikazani na slici 5.3. Za razliku od prethodna dva slučaja, rezultati ovde su malo drugačiji. Zajednice 1 i 2 opet imaju puno različitih tema. Zajednice 3 i 7 imaju nekih tema koje prate dobar deo subreddit-a u tim zajednicama, ali te teme nisu potpuno preovladavajuće u tim zajednicama. Zajednica 6 ima većinsku temu – Japan, sa 65% subreddit-a na tu temu, ali je to opet pad pošto su u prethodna dva slučaja zajednice gde potoji dominantna tema imale udeok od bar 80%. Zajednice 4 i 10 se obe bave sportom, prva hokejem a druga fudbalom, sa 90% i 87.5% subreddit-a na tu temu. Zajednica 5 je potpuno homogena i glavna tema je Kanada. Zajednice 8 i 9 se obe bave tehnologijom, prva više je više okrenuta ka mobilnim telefonima a druga kompjuterima i software-om za kompjutere, sa 100% i 92.31% subreddit-a na tu temu.



Slika 5.3 - Procenat subreddit-a sa istom temom u zajednicama (link-ovi sa negativnim sentimentom)

Od 30 ispitanih zajednica samo 8 nije imalo dominantnu temu. Na osnovu ovih rezultata se vidi da subreddit-i uglavnom link-uju subreddit-e sa zajedničkom temom. Takođe izgleda da subreddit-i više link-uju druge subreddit-e sa zajedničkom temom kada je u pitanju link sa pozitivnim ili neutralnim sentimentom (8 od 10 zajednica sa većinskom temom) za razliku od toga kada je u pitanju link sa negativnim sentimentom (6 od 10 zajednica sa većinskom temom).

Potrebno je još jednom napomenuti da se zbog nepostojanja pristupačnog skupa podataka ovaj deo istraživanja morao obaviti na relativno malom broju zajednica i subreddit-a, tako da potoji mogućnosti da se dobijeni rezultati ne mogu generalizovati na celu populaciju pronađenih zajednica.

6. Reference

[1] reddit: the front page of the internet, https://www.reddit.com/
[2] Reddit Statistics: pushshift.io, https://pushshift.io/
[3] SNAP: Stanford Network Analasys Project, https://snap.stanford.edu/
[4] SNAP: Social Network: Reddit Hyperlink Network, https://snap.stanford.edu/data/soc-RedditHyperlinks.html
[5] Overview – GraphFrames 0.8.0 Documentation, https://graphframes.github.io/graphframes/docs/_site/index.html
[6] Zhu, Xiaojin, and Zoubin Ghahramani, *Learning from labeled and unlabeled data with label propagation*, 2002.
[7] Listofsubreddits: ListofSubreddits, https://www.reddit.com/r/ListOfSubreddits/wiki/listofsubreddits