Izvješće o radu projektnog zadatka završnog rada

Student: Nola Čumlievski

Kolegij: Upravljanje znanjem

1. Priprema podataka

Za izradu ovog projektnog zadatka korišten je skup web scrapanih podataka prikupljenih sa portala „*baranjainfo*“ (iz prvog projektnog zadatka). Nakon učitavanja podataka

```
podaci = pd.read_csv("C:/Users/38591/AppData/Local/Programs/Python/Python36/baranjainfopravi.csv", sep=";")
```

uklanjamo nepotrebne varijable (autora, kategoriju,...) s obzirom da navedene nisu važne za zadanu analizu. Čistimo skup podataka tako da u skupu podataka ostanu varijable *Datum, Naslov* i *Tekst*. Također, potrebno je varijablu *Datum* pretvoriti u *datetime* vrstu podataka te sva slova unutar tekstova pretvoriti u mala slova.

```
podaci = podaci.drop(columns = ["Broj", "Link", "Autor", "Kategorija"])
podaci["Datum"] = pd.to_datetime(podaci["Datum"], format="%Y-%m-%d", errors='coerce')

podaci['Naslov'] = podaci['Naslov'].str.lower()
podaci['Tekst'] = podaci['Tekst'].str.lower()
```

Nakon uklananja nepotrebnih varijabli, filtriramo skup podataka prema ključnim riječima (covid, korona, cjepivo, capak, novooboljeli,...) kako bi dobili članke koji se odnose isključivo na tematiku korona virusa.

```
rijeci_korona = ["koron\w+", "koronavirus\w*", "virus\w*", "capak", "epidemij\w+", ...]
ukupno_korona = podaci[(podaci["Naslov"].str.contains('|'.join(rijeci_korona))) & (podaci["Tekst"].str.contains('|'.join(rijeci_korona)))]
```

Ukupan broj članaka koji se odnose na korona tematiku iznosi 763. S obzirom da je zadatak podijeliti članke u 4 grupe, koje rangiraju od datuma 1.1.2020 do konačno 25.8.2020 uklanjamo članke unutar skupa podataka koji datiraju nakon konačnog datuma. Nakon uklananja nepotrebnih redova, imamo 463 članka.

```
podaci = podaci[podaci.Datum < "2020-08-26"]
```

Zatim, kako bi imali sve tekstove na jednom mjestu, spajamo riječi iz naslova i teksta članka u jedan ujedinjeni stupac, *Text.*

```
podaci["Text"] = podaci["Naslov"] + podaci ["Tekst"]
```

Nakon stvaranja „zajedničkog“ stupca, možemo ukloniti stupac s naslovom i tekstom članka s obzirom da su sada redundantni.

```
podaci = podaci.drop(columns = ["Naslov", "Tekst"])
```

Nakon što smo strukturno pripremili skup podataka za analizu, slijedi „pročišćavanje“ zadanog skupa od specijalnih znakova, zaustavnih riječi, zamjenica, veznica i sličnih vrsta riječi. Napravljene su tri liste, jedna sa zaustavnim riječima, jedna sa veznicima, česticama, zamjenica i sličnim te treća sa specijalnim znakovima i skup podataka se pročišćava od navedenih na isti način.

```
zaustavne = ["a", "ako", "ali", "bi", "bih", "bila", "bili", "bilo", ...]
rijeci = ["ovoliko", "toliko", "onoliko", "nekoliko", "zadnja", ...]
spec_chars = ["!",'\"', "#", "%", "\$", "&","\'","\(", ...]
pattern = r'^\w+{}$'.format('|'.join(spec_chars))
podaci['Text'] = podaci["Text"].str.replace(pattern, '')
```

U ovom bloku koda prikazan je postupak sa listom specijalnih znakova, no u analizi je ponovljen isti za svaku navedenu listu. Navodimo regex (regularni izraz) u kojem spajamo svaki objekt unutar navedene liste i iste zamjenjujemo sa praznim stringom unutar varijable koja sadrži

naslove i tekstove članaka. Naknadno, izvedena su dodatna pročišćavanja znakova za koje je primjećeno da su ostali unutar teksta nakon provedbe ovog procesa (brojevi, strani znakovi kao što su emotikoni, strelice i slično).

Nakon pročišćavanja skupa podataka slijedi grupiranje istih u 4 navedena perioda. Odlučeno je isto napraviti tako da se za svaki navedeni period napravi poseban dataframe koji sadrži članke navedenog perioda.

```
pocetak = "2020-01-01"
kraj = "2020-02-24"
posli = podaci["Datum"] >= pocetak
prije = podaci["Datum"] <= kraj
izmedu = posli & prije
prvi_period = podaci.loc[izmedu]
```

Kao što je navedeno u pdf-u projektnog zadatka, periodi su grupirani na slijedeći način:

1. Skupina: svi članci nastali u periodu 1.1.2020.-24.2.2020.
2. Skupina: svi članci nastali u periodu 25.2.2020.-13.3.2020.
3. Skupina: svi članci nastali u periodu 14.3.2020.-11.5.2020.
4. Skupina: svi članci nastali u periodu 12.5.2020.-25.8.2020.

U bloku koda iznad prikazan je način izrade prve grupe, no proces je isti za sve periode, izuzev početnog i završnog datuma koji su prilagođeni periodu. Tražimo redove koje pripadaju unutar specifičnog raspona te na kraju pomoću metode *loc* pridružujemo skupu podataka redove koji imaju vrijednost *True* s obzirom na zadani uvjet filtracije.

Kada smo napravili grupaciju i 4 različita data frame-a možemo ukloniti varijablu datum iz svake s obzirom da je u ovom trenutku sam datum redundantna informacija.

```
prvi = prvi_period.drop(columns = ["Datum"])
```

Nakon izrade 4 data frame-a, veličine istih su redom:

- 1. grupa – 6 članaka
- 2. grupa – 12 članaka
- 3. grupa – 217 članaka
- 4. grupa – 228 članaka

Možemo primjeniti veliku razliku između grupa. Prva i druga sastoje se tek od nekoliko članaka sve skupa dok treća i četvrta grupa sadrže većinu članaka vezanih uz koronu u navedenom periodu.
Za daljni rad ideja je, da se svaki stupac s tekstom pojedinog data frame-a pretvori u listu, a zatim i u string kako bi pravilno mogli generirati oblik datoteke koji je pogodan za izradu mreže iz *pandas* data frame-a. U nastavku je opisano korak po korak kako je postignuta konačna struktura podataka.

1. Pretvorba pandas stupca u listu

```
prvi_list = prvi['Text'].tolist()
```

S obzirom da je rezultat ove operacije dvodimenzionalna lista (lista listi) čiji svaki element odgovara retku data frame-a potrebno je napisati funkciju pomoću koje ćemo „spljoštiti“ navedenu listu tako da dobijemo jednu listu koja se sastoji od riječi u člancima.

2. Pretvorba liste u jednodimenzionalnu listu

```
def flatten_list(lista):
    flat_list = []
    # iteracija kroz vanjsku listu
    for element in lista:
        if type(element) is list:
            # ako je objekt unutar liste isto lista
            for item in element:
                flat_list.append(item)
```

```
    return flat_list

lista_prva = flatten_list(prvi_list)
```

Inicijalizirana je prazna lista u koju ćemo nadopunjavati riječi. Obavlja se iteracija kroz vanjsku listu i provjerava da li je objekt unutar liste također lista te ako je, gledamo pojedine elemente unutarnjih listi koje zatim pridružujemo inicijalnoj praznoj listi. Funkcija vraća jednodimenzionalnu listu sa svim elementima podlisti vanjske liste.

3. Pretvorba liste elemenata u string s riječima članaka

```
string_prvi = " ".join(str(x) for x in lista_prva)
print(string_prvi)
string_prvi = " ".join(string_prvi.split())
```

Pridružujemo (engl. *join*) elemente unutar liste tipa string u inicijalni string. Zatim elemente istog pomoću funkcije *split* i *join* spajamo kako bi imali samo jedan razmak (*whitespace*) između riječi s obzirom da se u inicijalno stvorenom stringu nalazilo previše razmaka između pojedinih riječi. Ovaj postupak ponovljen je za sva 4 perioda (4 data frame-a).

4. Stvaranje data frame-a sa čvorovima i težinama

```
def broj_parova(string):
    rijeci = re.findall("\w+", string)
    parovi = zip(rijeci, rijeci[1:])
    return collections.Counter(parovi)
```

Za početak je napisana funkcija za pronalaz parova riječi i frekvencije pojavljivanja istog unutar teksta. Pronalaze se sve riječi unutar stringa, te se svaka pomoću funkcije *zip* spaja sa svojim susjednom i pomoću funkcije *Counter* prebrojava se ponavljanje postavljenog para.

Key	Type	Size	Value
('agencija', 'plaćanja')	int	1	3
('aktivnosti', 'dokumentirane')	int	1	1
('aktivnosti', 'pratećih')	int	1	3
('aktualne', 'podatke')	int	1	1
('alkohola', 'krvi')	int	1	7
('alkohola', 'prekršaja')	int	1	1
('automobila', 'godišnji')	int	1	1

Slika 1. Rezultat funkcije broj_parova

S obzirom da je rezultat gore navedene funkcije *Counter* objekt, koji je ujedno i rječnik (sadrži ključeve i vrijednosti) gdje su ključevi navedeni parovi, a vrijednosti sama težina (frekvencija) para, koristimo funkcije *DataFrame.from_dict* kako bi rječnik pretvorili u *pandas* data frame.

```
lista_1 = broj_parova(string_prvi)
df = pd.DataFrame.from_dict(lista_1, orient='index').reset_index()
df.columns = ["rijeci", "tezina"]
tezina = df["tezina"]
df1 = pd.DataFrame(df['rijeci'].tolist(), index=df.index)
df1 = df1.join(tezina)
df1.columns = ["source", "target", "tezina"]
```

Pretvaramo rječnik u data frame pomoću navedene funkcije i preimenujemo stupce. Zatim izdvajamo težinu kao zaseban *pandas series objekt* i stvaramo novi data frame u kojem pomoću funkcije *tolist* razdvajamo vrijednosti do tada jednog stupca u dva zasebna (kako bi imali zaseban stupac za source i za target riječ - čvor). Nakon toga, novom data frame-u je pridodan težina *pandas series* objekt i kao rezultat imamo data frame sa tri stupca, izvornom riječ (*source*), završnom riječi (*target*) te težinom navedenog para.

Index	source	target	tezina
0	objavljena	nova	1
1	nova	natječaja	1
2	natječaja	podmjeru	1
3	podmjeru	potpora	1
4	potpora	ulaganja	3
5	ulaganja	preradu	3
6	preradu	marketing	3
7	marketing	razvoj	3
8	razvoj	poljoprivrednih	3

Slika 2. Konačni data frame s parovima

Na slici 2. prikazan je data frame sa source, target i težinom prvog perioda. Prethodno opisani postupak proveden je na isti način na sva 4 data frame-a te kao rezultat imamo:

- Data frame prvog perioda – 519 parova
- Data frame drugog perioda – 1678 parova
- Data frame trećeg perioda – 26 442 para
- Data frame četvrtog perioda – 17 787 parova

Konačno, zbog potrebe izrade i velike mreže (sa svim parovima) stvoren je poseban data frame koji sadrži sve čvorove perioda od 1.1.2020 do 25.8.2020.

```
frames = [df1, df2, df3, df4]
ukupno = pd.concat(frames)
```

2. Analiza velike jezične mreže

Za početak, izrađujemo veliku jezičnu mrežu koja se sastoji od svih čvorova (sva 4 perioda).

```
graf = nx.from_pandas_edgelist(ukupno, source = "source", target = "target
", edge_attr=True, create_using = nx.DiGraph)
```

Korištena je funkcija *nx.from_pandas_edgelist* gdje navodimo data frame sa informacijama, izvorni čvor (*source*), završni čvor (*target*), postavljamo atribute veza na *True* kako bi funkcija povukla atribut težine koji postoji unutar data frame-a te navodimo da želimo usmjereni graf (*DiGraph*).

Na tablici 1. prikazane su globalne mjere i rezultati dobiveni za navedenu jezičnu mrežu *graf*.

Tablica 1. Globalne mjere

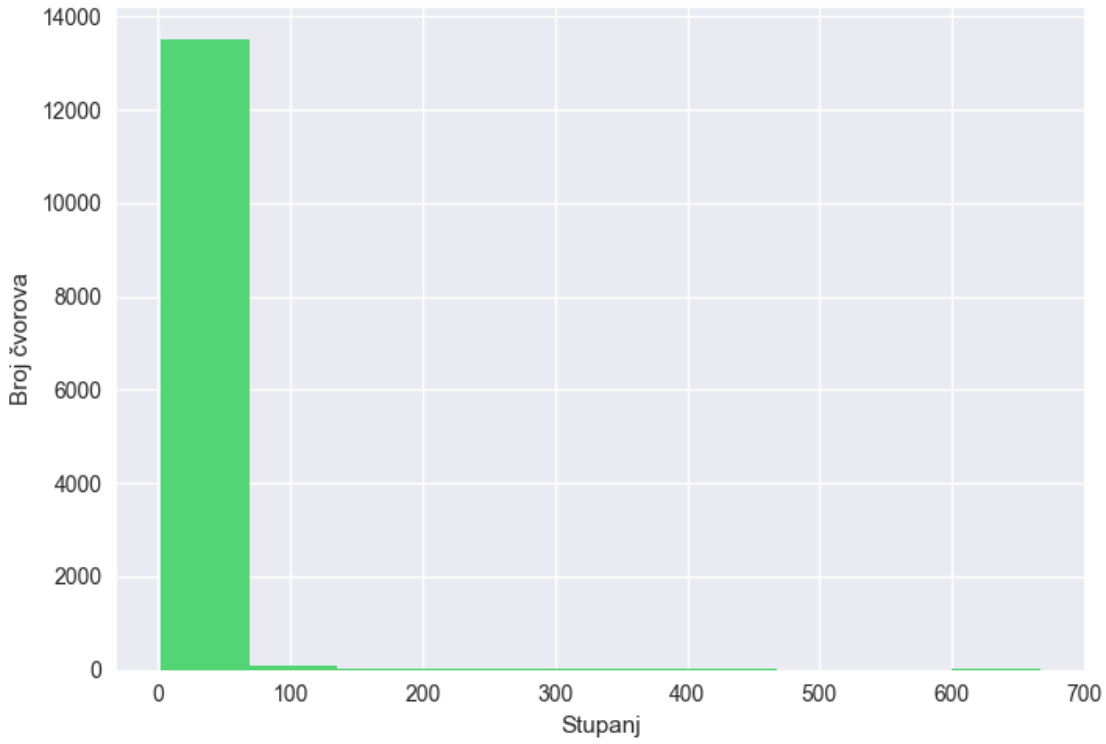
Mjera	Rezultat
Broj čvorova	13 626
Broj veza	44 101
Prosječan stupanj mreže	16.37
Gustoća mreže	0.0002375
Broj povezanih komponenti	1
Prosječna duljina najkraćeg puta	9.979
Dijametar mreže	24
Prosječni koeficijent grupiranja	0.0357
Koeficijent asortativnosti	0.016

Zatim, izrađen je grafikon distribucije stupnjeva za cijelu mrežu.

```
def graf_distribucije_stupnjeva(graf):
    stupnjevi = [graf.degree(n) for n in graf.nodes()]
    plt.style.use("seaborn")
    plt.hist(stupnjevi, color = "#53d475")
    plt.xlabel("Stupanj")
    plt.ylabel("Broj čvorova")
    plt.show()
```

```
graf_distribucije_stupnjeva(graf)
```

Za izradu grafikona distribucije stupnjeva korištena je ista funkcija kao i u drugom projektu.



Slika 3. Grafikon distribucije stupnjeva

Na slici 3. možemo vidjeti histogram distribucije stupnjeva velike jezične mreže. Vidimo da većina čvorova sadrži stupanj od 0-100 što znači da većina čvorova ima ~70-80 čvorova susjeda i manje. Postoje i čvorovi sa stupnjem od 100, 300 pa čak i više od 600 ali ih ima izrazito malo.

3. Analiza jezičnih mreža pojedinog perioda

1. period

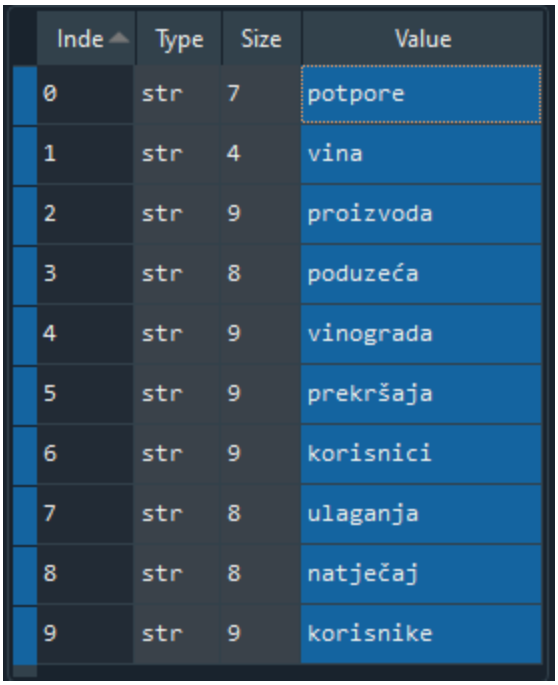
```
graf1 = nx.from_pandas_edgelist(df1, source = "source", target = "target",
 edge_attr=True, create_using = nx.DiGraph)
```

Mrežu radimo na isti način kao i prethodnu uz drugačiji navod data frame-a iz kojeg izvlačimo podatke. Mreža prvog perioda sastoji se od samo 389 čvorova.

Mjera centralnosti stupnja

```
centralnost_stupnjeva = nx.degree_centrality(graf1)
centralnost_top10 = sorted(centralnost_stupnjeva, key=centralnost_stupnjev
a.get, reverse=True)[:10]
```

Mjeru računamo pomoću *networkx* funkcije *degree_centrality*. Zatim iz navedenog rječnika izvlačimo 10 čvorova koji su sortirani od čvora s najvećom mjerom centralnosti prema onome sa najmanjom mjerom centralnosti stupnja.



Inde	Type	Size	Value
0	str	7	potpore
1	str	4	vina
2	str	9	proizvoda
3	str	8	poduzeća
4	str	9	vinograda
5	str	9	prekršaja
6	str	9	korisnici
7	str	8	ulaganja
8	str	8	natječaj
9	str	9	korisnike

Slika 4. Tablica čvorova prvog perioda sa najvećom mjerom centralnosti stupnja

Na slici 4. vidimo tablicu sa 10 čvorova sa najvećom mjerom centralnosti stupnja prvog perioda. Vidimo da prvi period (prvi mjesec i dio drugog mjeseca) nemaju ključne riječi povezane sa korona tematikom što i ima donekle smisla s obzirom da se tada još nije previše pisalo o virusu na portalima. Također mislim da to ima i veze s portalom, ovdje se radi o lokalnom portalu (za baranju) te su možda portali poput 24 sata, večernjeg lista i sličnih i imali članke o koroni u tom periodu, no to ovdje nije slučaj.

Page rank

Page rank mjera centralnosti, s obzirom da nije objašnjena u prošlom projektu, će biti ukratko objašnjena ovdje. Navedena mjera predstavlja prilagodbu Katz centralnosti (opisane u prošlom seminaru) koja uzima u obzir 3 faktora:

1) broj linkova (poveznica) koju prima određeni čvor

2) sklonost povezivanju poveznica čvora

3) mjeru centralnosti poveznica čvora

U prijevodu, što više poveznica čvor privlači, to je čvor važniji. Isto tako, linkovi iz važnijih čvorova, s većom mjerom centralnosti važniji su od onih koji dolaze iz slabijih čvorova niže mjere centralnosti.

```
page_rank = nx.pagerank(graf1, weight = "tezina")
page_rank_top10 = sorted(page_rank, key=page_rank.get, reverse=True)[:10]
```

Networkx ima funkciju za izračun mjere page ranka. U navedenu funkciju, navodimo mrežu za čije čvorove želimo obračunati mjeru centralnosti te težinu čvorova. Konačno, sortiranjem i izvlačenjem čvorova najveće mjere page ranka dobijemo slijedeće:

Inde	Type	Size	Value
0	str	7	potpore
1	str	8	alkohola
2	str	10	prisutnost
3	str	9	korisnike
4	str	9	proizvoda
5	str	4	krvi
6	str	4	vina
7	str	5	mjera
8	str	8	ulaganja
9	str	9	vinograda

Slika 5. Page rank mjera za prvi period

Na slici 5. prikazani su čvorovi sa najvećom mjerom page ranka za prvi period. Možemo primjetiti sličnost sa 10 čvorova s najvećom mjerom centralnosti, što bi značilo da većian tih čvorova, ne samo da ima najveći broj susjeda već prima poveznice (linkove) od čvorova koji također imaju visoku mjeru centralnosti stupnja.

Vizualizacije

Za vizualizaciju mreže prema navedenim lokalnim mjerama centralnosti korištena je slijedeća funkcija:

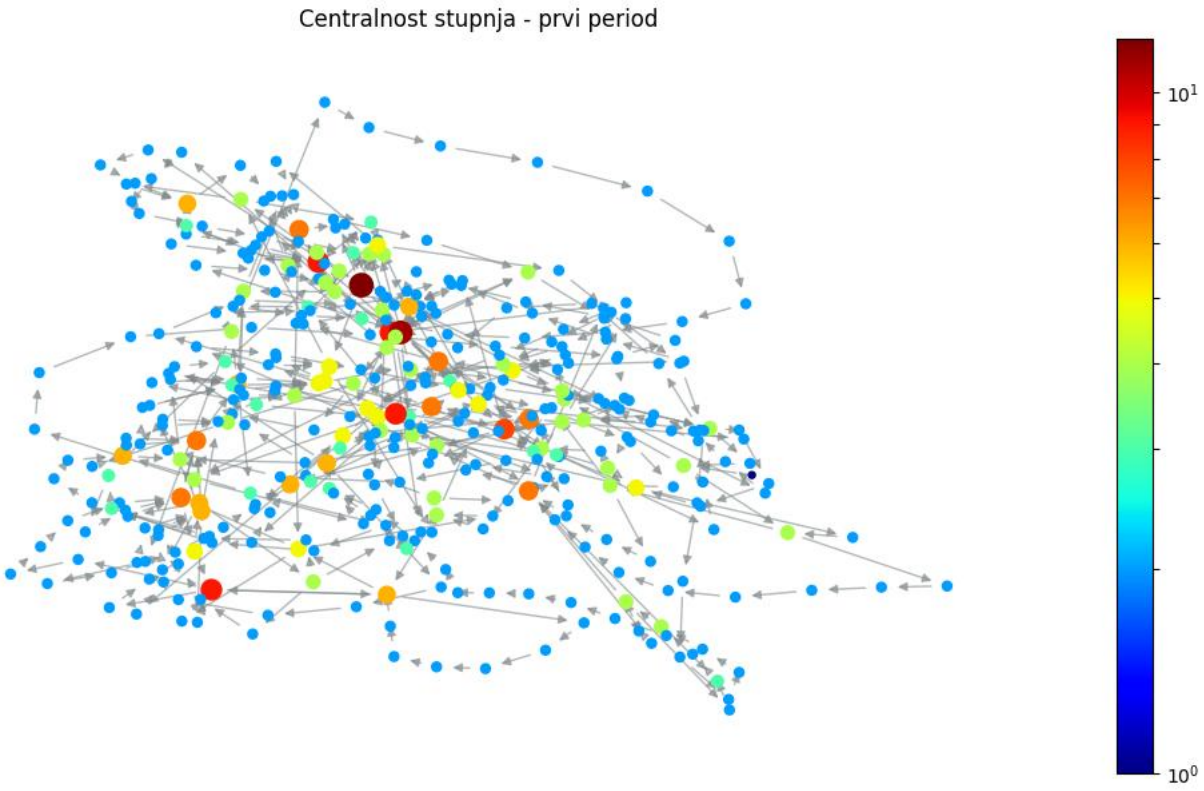
```
def draw(G, pos, node_size, measures, measure_name):
    # grafovi prema centralnosti
    nodes = nx.draw_networkx_nodes(G, pos, node_size=node_size, cmap=plt.c
m.jet, node_color=list(measures.values()),nodelist=measures.keys())
    nodes.set_norm(mcolors.SymLogNorm(linthresh=0.01, linscale=1, base=10)
)
    edges = nx.draw_networkx_edges(G, pos, edge_color='#818a8c', alpha=0.6
, width=1)
    plt.title(measure_name)
    plt.colorbar(nodes)
    plt.axis('off')
    plt.show()
```

Ova funkcija opisana je u seminaru drugog projektnog zadatka. Ukratko, kao parametre, prima mrežu, poziciju koju također moramo izračunati za pojedini čvor, veličinu čvorova te mjeru i naziv iste prema kojoj se vizualizira boja i veličina dotičnog čvora.

```
plt.figure(3,figsize=(200,200))
node_sizes = []
for n in centralnost_stupnjeva.values():
        node_sizes.append( 5000 * n )
pos = nx.spring_layout(graf1, seed=1)
draw(graf1, pos, node_sizes, dict(graf1.degree), 'Centralnost stupnja - pr
vi period')
```

Postupak za vizualizaciju ove mreže, kao i svake što slijedi, ide na slijedeći način: kalkulira se veličina čvorova iz prethodno dobivenog rječnika (kada je izračunata određena mjera centralnosti za čvorove) te se veličine za sve čvorove spremaju u praznu, prethodno iniciranu listu. Računamo poziciju čvorova pomoću funkcije *spring_layout* te u samu funkciju

vizualizacije navodimo mrežu, poziciju, listu veličine čvorova, te mjeru i naziv mjere koja se koristi za vizualizaciju.

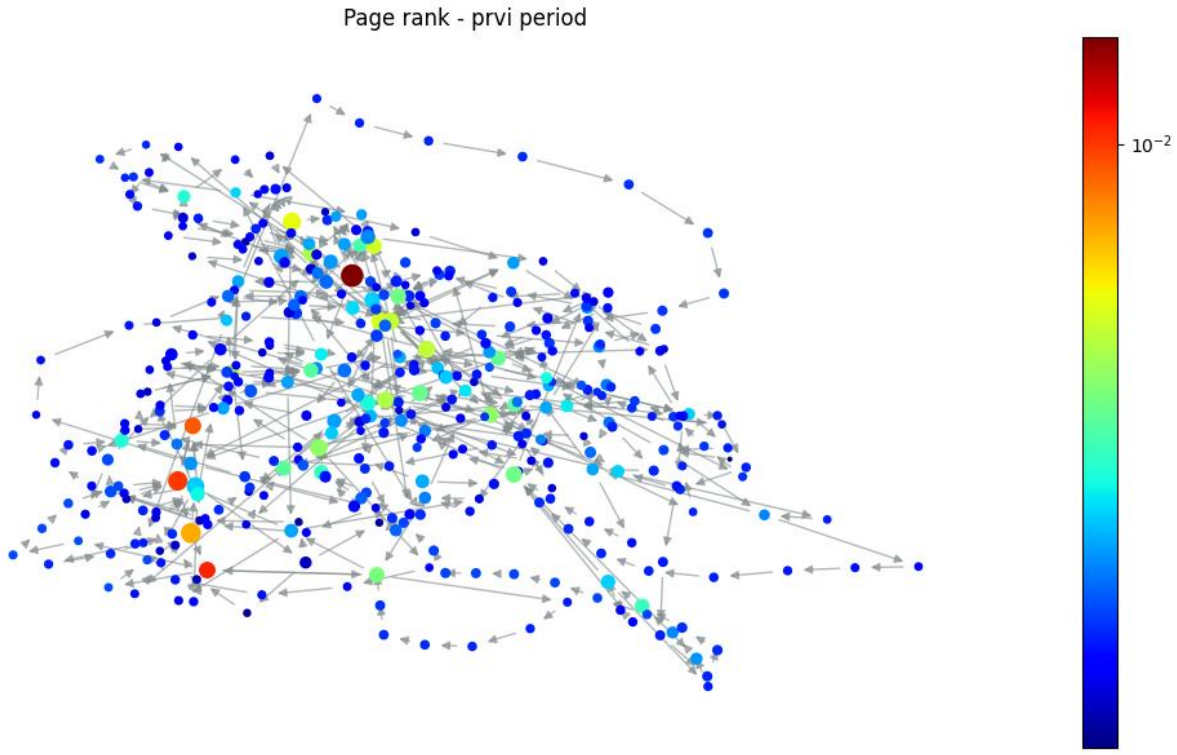


Slika 6. Vizualizacije mreže prvog perioda prema mjeri centralnosti stupnja

Na slici 6. prikazana je vizualizacija jezične mreže prema centralnosti stupnja čvorova. Vidimo da se većina čvorova veće centralnosti nalazi upravo u središtu same mreže, te isto objašnjava da upravo ti čvorovi imaju najviše susjeda unutar mreže.

```
plt.figure(3,figsize=(200,200))
node_sizes = []
for n in page_rank.values():
        node_sizes.append( 10000 * n )
draw(graf1, pos, node_sizes, nx.pagerank(graf1), 'Page rank - prvi period'
)
```

Kao što je vidljivo prema kodu, na isti način vizualiziramo mrežu prema drugoj mjeri centralnosti. Sukladno mjeri, izrađujemo novu listu veličine čvorova te nadodajemo pripadajuću mjeru za izračun centralnosti.



Slika 7. Vizualizacija mreže prema mjeri page ranka

Na slici 7. nalazi se vizualizacija mreže prvog perioda prema mjeri centralnosti page ranka. Vidimo da se čvorovi koji primaju najveći broj poveznica od važnijih čvorova nalaze lijevo dolje i prema sredini jezične mreže. Možemo primjetiti logiku i povezanost ove dvije mjere za ovu mrežu. S obzirom sa se najveći broj čvorova s najvećom mjerom centralnosti stupnja nalazi u središtu mreže, možemo vidjeti da su čvorovi smješteni relativno blizu tih čvorova te ako pogledamo strelice veza, možemo primjetiti da dosta poveznica prema čvorovima najvećeg ranka dolazi iz središta mreže (što potvrđuje da su važniji čvorovi prema kojima dolaze poveznice čvorova veće mjere centralnosti).

2. period

```
graf2 = nx.from_pandas_edgelist(df2, source = "source", target = "target",
 edge_attr=True, create_using = nx.DiGraph)
```

Za izradu jezične mreže drugog perioda koristimo istu funkciju, uz navođenje pripadajučeg data frame-a. S obzirom da su za izračun mjera i izradu vizualizacije korišteni isti postupci (što će se i vidjeti u programskoj skripti priloženoj uz ovo izvješće) neću opisivati dalje kod (da se ne ponavljam) te će se poglavlja fokusirati na rezultate i do sad neviđeni kod.

Mreža drugog perioda sastoji se od ukupno 1 160 čvorova.

Inde	Type	Size	Value
0	str	7	stožera
1	str	9	hrvatskoj
2	str	5	osoba
3	str	5	osobe
4	str	9	oboljelih
5	str	8	županije
6	str	13	koronavirusom
7	str	8	području
8	str	7	zaštite
9	str	11	koronavirus

Slika 8. Čvorovi s najvećom mjerom centralnosti stupnja

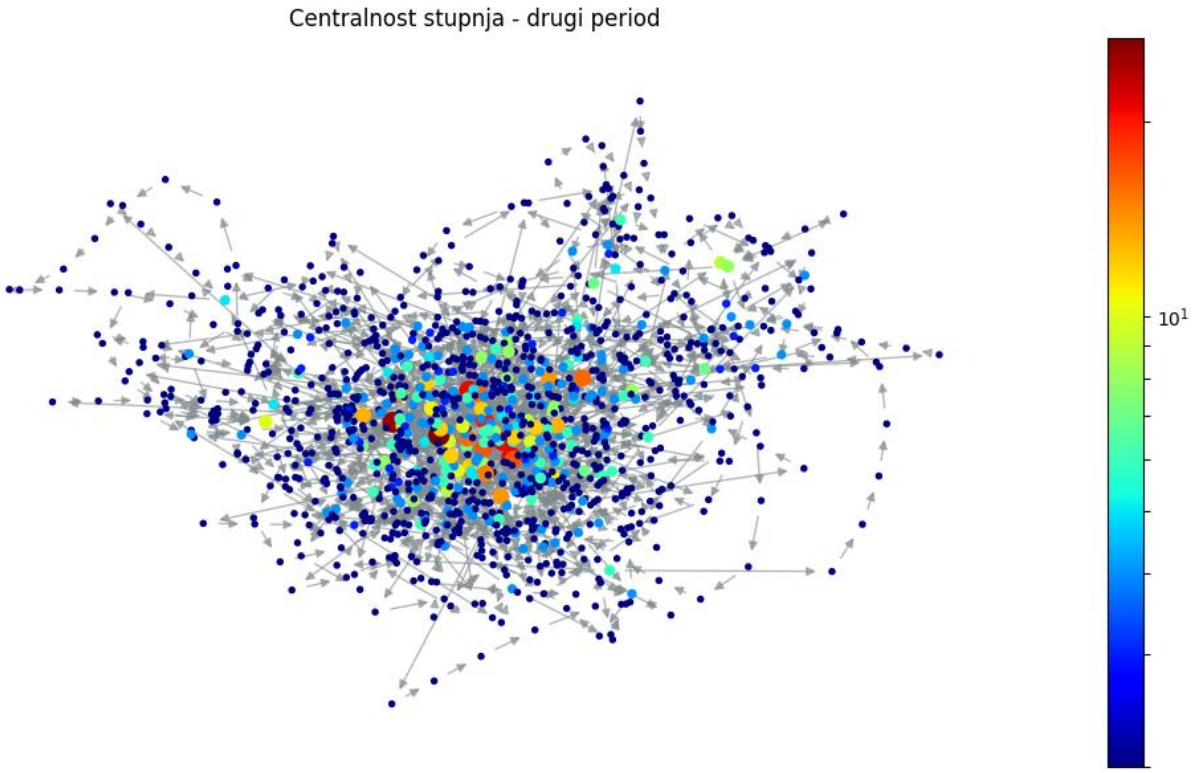
Na slici 8. nalazi se 10 čvorova drugog perioda s najvećom mjerom centralnosti stupnja. Za razliku od ključnih riječi prvog perioda, možemo primjetiti da se u drugom periodu (krajem drugog do sredine trećeg mjeseca) već pojavljuju riječi (čvorovi) s izrazitom povezanosti s korona tematikom, što ima smisla kad pogledamo kako se počelo u hrvatskoj govoriti i pisati o korona virusu (ovaj period predstavlja „prijelomni“ period kada je virus došao u hrvatsku te su se počeli zatvarati razni objekti).

Inde	Type	Size	Value
0	str	7	stožera
1	str	8	županije
2	str	9	hrvatskoj
3	str	8	području
4	str	9	oboljelih
5	str	7	zaštite
6	str	5	osoba
7	str	7	civilne
8	str	5	osobe
9	str	3	kbc

Slika 9. Čvorovi s najvećom mjerom page ranka drugog perioda

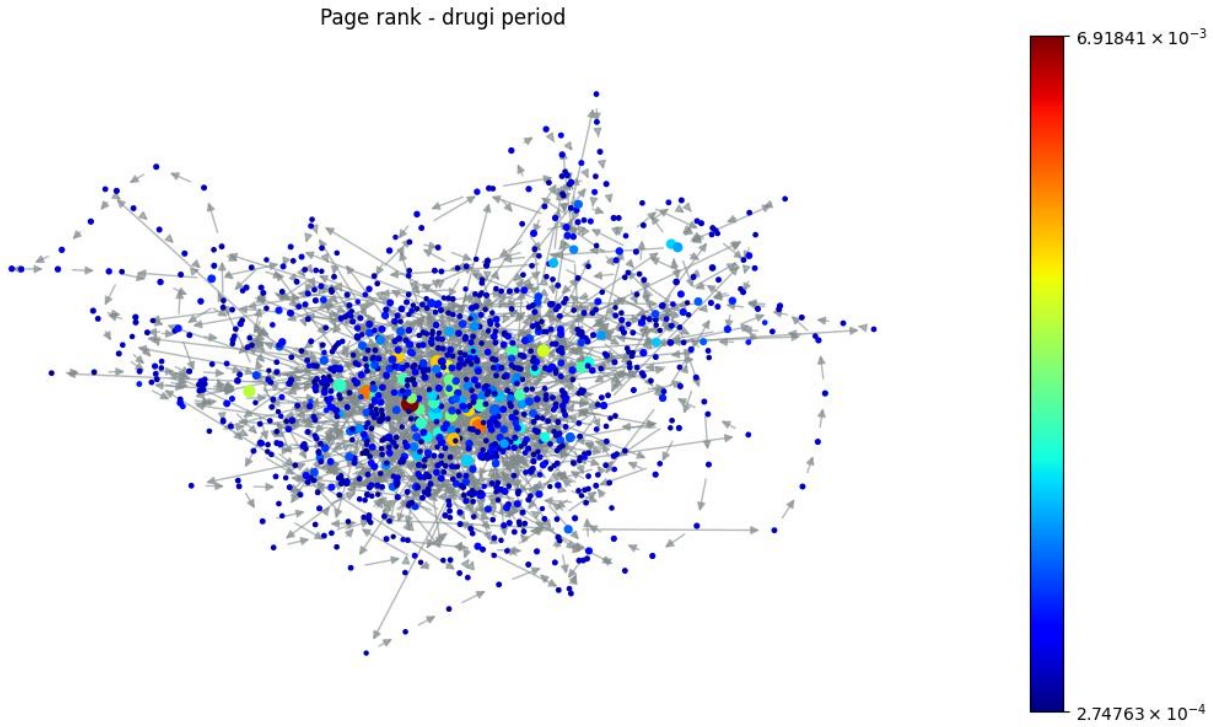
Na slici 9. nalazi se deset čvorova s najvećom mjerom centralnosti page ranka drugog perioda. Možemo primjetiti izrazitu sličnost sa deset čvorova iste mreže s najvećom mjeron centralnosti stupnja, točnije, razlika je u samo dvije riječi („civilne“ i „kbc“) te je ostatak riječi identičan s blago drugačijim poretkom.

Vizualizacije



Slika 10. Vizualizacija mreže drugog perioda prema mjeri centralnosti stupnja

Na slici 10. možemo vidjeti vizualizaciju mreže prema mjeri centralnosti stupnja čvorova. Kao i kod vizualizacije prema centralnosti stupnja prvog perioda, možemo primjetiti kako se čvorovi s najvećom mjerom centralnosti stupnja nalaze u središtu mreže, što su čvorovi navedeni u prethodnoj tablici (uz ostale).



Slika 11.Vizualizacija mreže prema stupnju centralnosti page ranka

Na slici 11. možemo vidjeti vizualizaciju mreže prema mjeri centralnosti page ranka drugog perioda. Vidimo da su otprilike isti čvorovi, koji su imali najveću vrijednost mjere centralnosti stupnja upravo većinom isti čvorovi koji također imaju i najveću vrijednost ove mjere, što i potvrđuju prethodno napravljene tablice koje prikazuju prvih deset čvorova za obe mjere.

Treći period

Inde	Type	Size	Value
0	str	5	osoba
1	str	5	mjere
2	str	5	mjera
3	str	5	osobe
4	str	12	koronavirusa
5	str	8	županije
6	str	9	hrvatskoj
7	str	9	zaraženih
8	str	9	bolesnika
9	str	9	oboljelih

Slika 12. Deset čvorova s najvećom mjerom centralnosti stupnja

Na slici 12. prikazana je tablica sa deset čvorova s najvećom centralnosti stupnja u trećem periodu. Možemo primjetiti da se, kao i u drugom periodu, ključne riječi odnose velikim djelom na korona tematiku (ovo je bio također i period prvog lockdown-a, zatvaranja kafića, škola, fakulteta i slično) te možemo primjetiti znatan porast spominjanja riječi kao što su „mjere“, „zaraženi“, „koronavirus“ i sličih riječi, te također i broja samih članaka vezanih uz korona virus (broj samih članaka u odnosu na prva dva perioda uspoređen je na početku ovog izještaja).

Inde	Type	Size	Value
0	str	5	osoba
1	str	5	mjera
2	str	5	mjere
3	str	5	osobe
4	str	12	koronavirusa
5	str	8	županije
6	str	6	stožer
7	str	7	civilne
8	str	7	stožera
9	str	7	zaštite

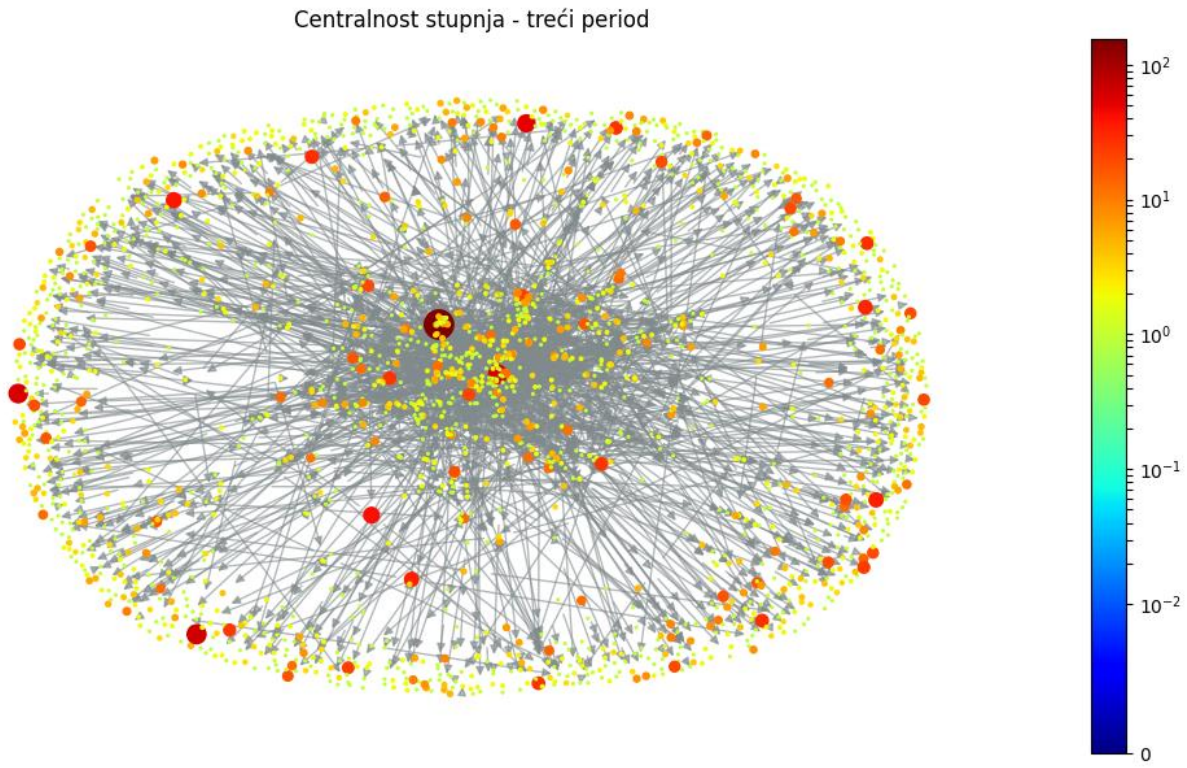
Slika 13. Deset čvorova s najvećom mjerom centralnosti page ranka

Na slici 13. nalazi se 10 riječi (čvorova) trećeg perioda s najvećim brojem centralnosti page ranka. Možemo također, kao i u primjeru drugog perioda, primjetit kao prvo povezanost riječi sa korona tematikom i samu sličnost čvorova sa čvorovima s najvećom mjerom centralnosti stupnja.

S obzirom da se mreža sastoji od 9 556 čvorova ukupno, vizualizacija cijele mreže nije bila moguća: pyplot je štekao (not responding), restart kernela prilikom pokretanja i slični problemi (laptop je star i ne izrazito jak) te je odlučeno, za potrebe vizualizacije (mjere su provedene na cijeloj mreži), napraviti „podmrežu“ navedene mreže koju možemo vizualizirati.

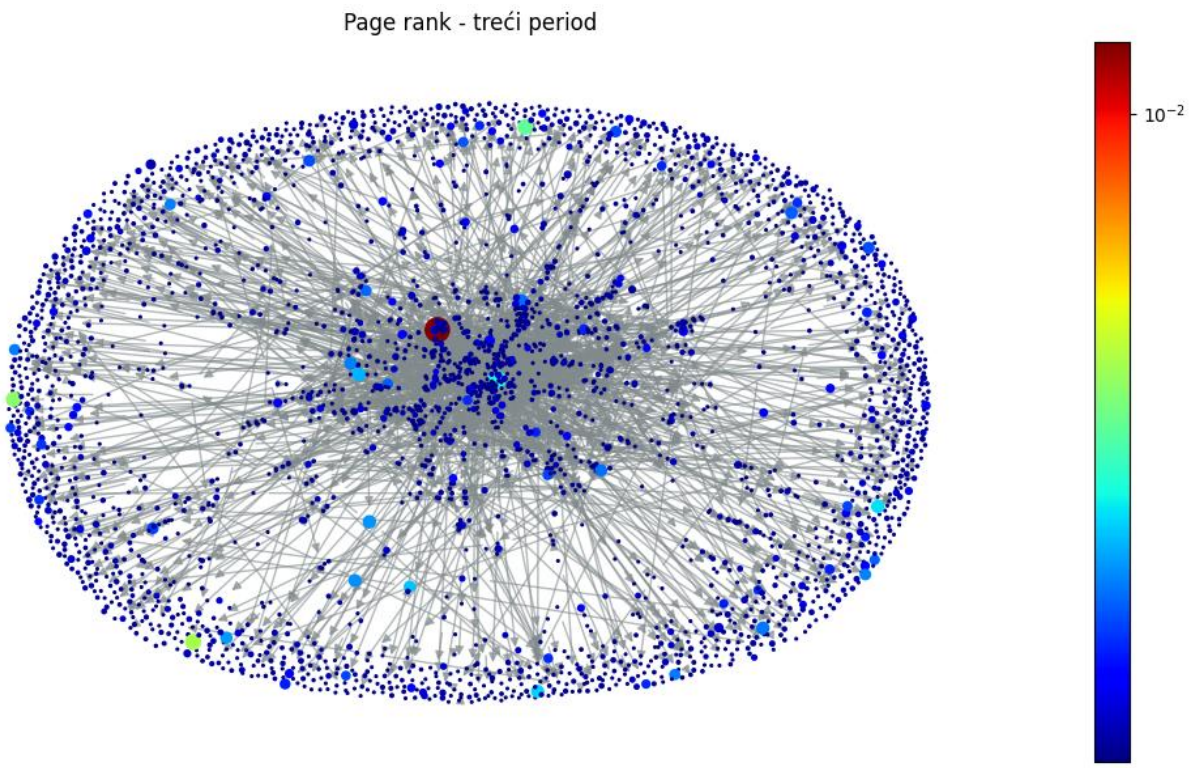
```
random_nodes = sample(list(graf3.nodes()), 3000)
graf3_1 = graf3.subgraph(random_nodes)
```

Korištena je funkcija *sample* pomoću koje se odabire 3000 nasumičnih čvorova iz liste čvorova treće mreže. Nakon izrade liste nasumičnih čvorova, istu listu prenosimo u funkciju *subgraph* pozvanu nad inicijalnom mrežom.



Slika 14. Vizualizacija mreže prema centralnosti stupnja

Na slici 14. prikazana je vizualizacija mreže prema centralnosti stupnja trećeg perioda (barem dijela). Možemo primjetiti da nasumično odabrani čvorovi imaju visok stupanj centralnosti stupnja.



Slika 15. Vizualizacija mreže prema mjeri centralnosti page ranka

Na slici 15. prikazana je vizualizacija mreže prema mjeri centralnosti page ranka za treći period. Ovdje vidimo totalnu suprotnost u ove dvije mjere centralnosti (za razliku od prva dva perioda). Dok su navedeni čvorovi imali visoku vrijednost centralnosti stupnja, isti imaju izrazito nisku vrijednost mjere centralnosti page ranka (osim jednog čvora u sredini koji ima i visoku mjeru centralnosti page ranka – riječi „osoba“). Ove dvije vizualizacije su samo dio mreže i možda ne predstavljaju točnu, veću sliku prave mreže no daju uvid u dio nje.

Četvri period

Četvrta mreža sastoji se od 7035 čvorova.

Inde	Type	Size	Value
0	str	5	osoba
1	str	5	osobe
2	str	9	hrvatskoj
3	str	8	županije
4	str	5	mjere
5	str	9	oboljelih
6	str	5	mjera
7	str	6	stožer
8	str	12	koronavirusa
9	str	13	novooboljelih

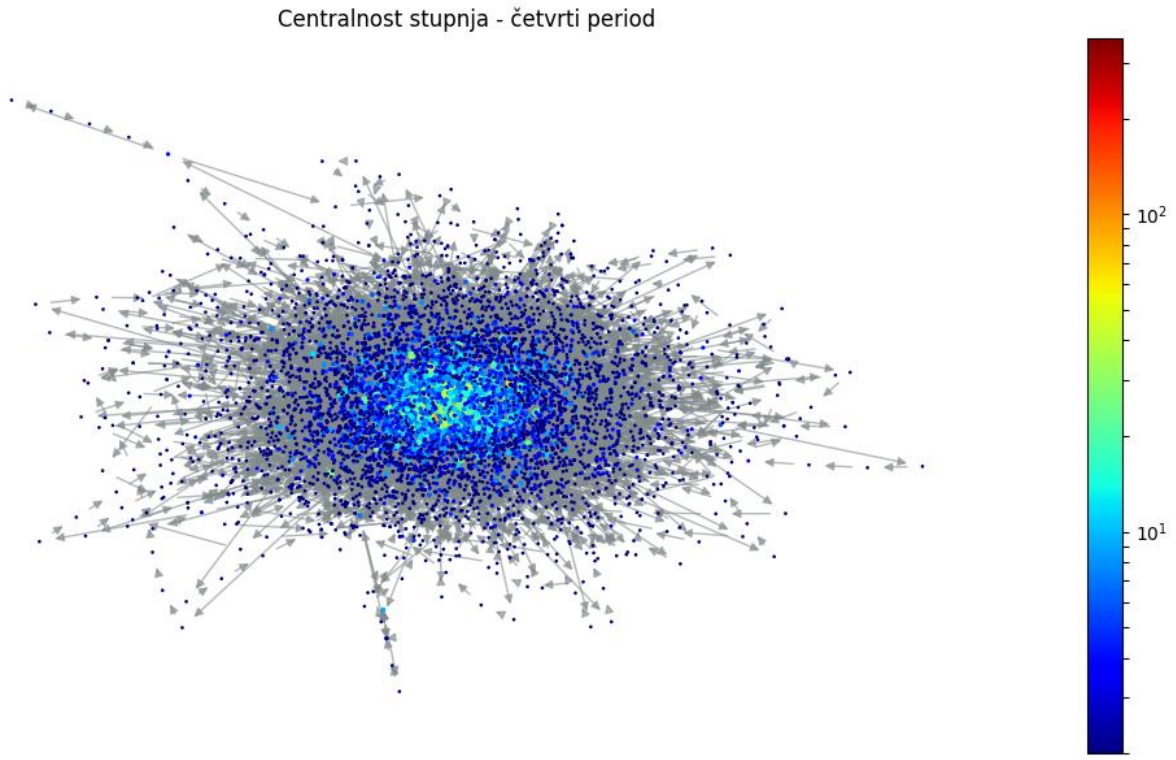
Slika 16. Čvorovi prema centralnosti stupnja

Na slici 16. nalazi se prikaz čvorova sa najvećom mjeron centralnosti stupnja za četvrti period. Možemo primjetiti, kao i kod prošla dva perioda, da se ključne riječi odnose izrazito na korona tematiku te da se iste riječi ponavljaju među naredna tri perioda.

Inde	Type	Size	Value
0	str	5	osoba
1	str	5	osobe
2	str	8	županije
3	str	6	stožer
4	str	9	hrvatskoj
5	str	9	oboljelih
6	str	5	mjere
7	str	8	području
8	str	12	koronavirusa
9	str	6	osijek

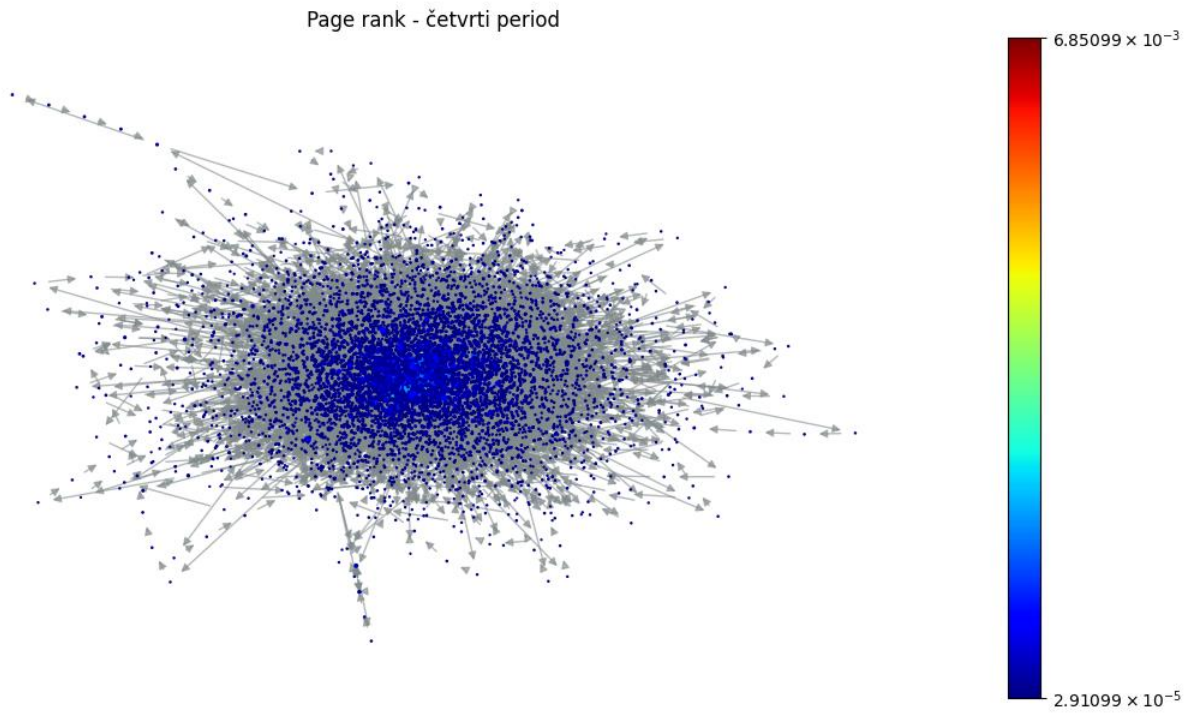
Slika 17. Čvorovi s najvećom mjerom centralnosti page ranka

Na slici 17. nalazi se 10 čvorova s najvećom mjerom centralnosti page ranka za četvrti period (sredina petog do kraja osmog mjeseca). Kao i kod čvorova drugog perioda, može se primjetiti izrazita sličnost u ključnim riječima sa riječima najveće mjere centralnosti stupnja.



Slika 18. Vizualizacija mreže prema mjeri centralnosti stupnja

Na slici 18. nalazi se vizualizacija jezične mreže četvrtog perioda prema mjeri centralnosti stupnja. S obzirom da mreža ima velik broj čvorova, nije moguće točno prikazati tok i same čvorove, no možemo primjetiti da veći broj čvorova ima mjeru centralnosti stupnja manju od 10, dok određen dio čvorova u središtu mreže ima veću vrijednost mjere centralnosti stupnja.



Slika 19. Vizualizacija mreže prema mjeri centralnosti page ranka

Na slici 19. prikazana je vizualizacija mreže četvrtog perioda prema mjeri centralnosti page ranka. Kao i kod prošle vizualizacije (što je potvrđeno i ključnim riječima u tablicama) vidimo da postoji sličnosti između vrijednosti čvorova za ove dvije mjere.

4. Analiza zajednica jezičnih mreža prema periodima

```
communities = community.asyn_lpa_communities(graf1, weight = "tezina")
communities = list(communities)
```

Za izradu zajednica odabran je algoritam asinkronog propagiranja oznaka. Nakon inicijalizacije oznake za svaki pojedini čvor, algoritam kroz više navrata postavlja oznaku čvora ovisno o tome koja je oznaka čvorova koji okružuju navedeni čvor. Algoritam se zaustavlja kada svaki čvor ima oznaku koja se pojavljuje najviše među njegovim susjedima. Prilikom proračuna oznake, uzima se u obzir i težina. *Networkx* ima određen broj algoritama za izračun iste, no većina nije implementirana za usmjerene mreže te izbor nije bio izrazito velik pri odabiru algoritma.

Prvi period

Inde	Type	Size	Value
0	set	54	{'objavljivati', 'informacije', 'vezane', 'zemlje', 'unijiizvor', 'zdr ...
1	set	118	{'proizvođača', 'troškove', 'možete', 'obavještavamo', 'baranjske', 'm ...
2	set	35	{'gdje', 'izvora', 'svibnja', 'operacije', 'proizvodnju', 'republike', ...
3	set	91	{'djelovanja', 'mikro', 'kunskoj', 'mesa', 'obavijest', 'početnike', ' ...
4	set	8	{'upisani', 'upisane', 'pravne', 'osobe', 'vinogradarski', 'vinogradar ...
5	set	83	{'prometne', 'školskoj', 'vikenda', 'školama', 'vozilom', 'osobnog', ' ...

Slika 20. Zajednice prvog perioda

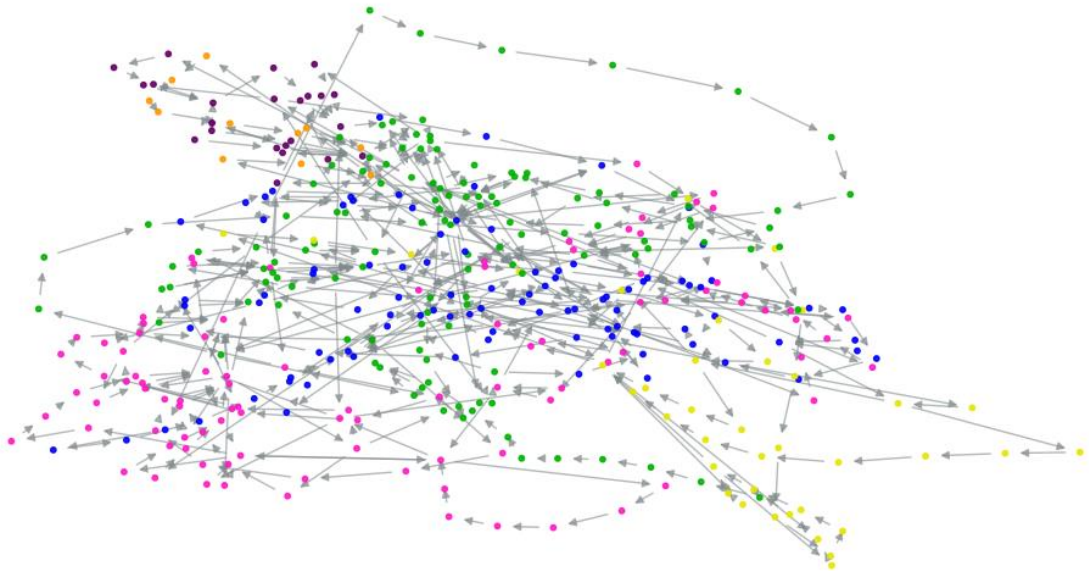
Na slici 20. prikazana je lista zajednica za mrežu prvog perioda. Vidimo da je algoritam riječi podjelio u 6 zajednica, rangirane indeksima od 0 do 5. Možemo vidjeti određeni uzorak među skupinama, npr. zadnja zajednica se odnosi na škole i promet, možda autoškole, peta zajednica na vinogradarstvo itd.

```
color_map = []
for node in graf4:
    if node in communities[0]:
        color_map.append('#5d8aa8')
    elif node in communities[1]:
        color_map.append("#f0f8ff")
    elif node in communities[2]:
        color_map.append("#e32636")
    elif node in communities[3]:
        color_map.append("#ffbf00")
    elif ...
```

Kao i u prošlom projektu, inicijaliziramo praznu listu u koju ćemo spremati boje zajednica. Iteracijom kroz čvorove mreže, ovisno o pripadnosti čvora zajednici, svakom čvoru se dodjeljuje određena boja koja nam kasnije služi za vizualizaciju mreže prema zajednicama.

```
plt.figure(3,figsize=(200,200))
sc = nx.draw_networkx_nodes(G=graf1, pos = pos, nodelist = graf1.nodes(),
alpha=0.9, node_size = 10, node_color=color_map)
nx.draw_networkx_edges(G = graf1, pos = pos, edge_color='#818a8c', alpha=0
.6, width=1)
plt.show()
```

Nakon što imamo listu boja, možemo pomoću funkcije *draw_network_nodes* i *draw_network_edges* vizualizirati mrežu prema zajednicama.



Slika 21. Vizualizacija zajednica prvog perioda

Na slici 21. možemo vidjeti vizualizaciju mreže prvog perioda prema zajednicama. Možemo uvidjeti logiku rada algoritma, s obzirom da su susjedni čvorovi većinom grupirani u iste zajednice (pod pretpostavkom da se susjedni čvorovi odnose na istu tematiku, što je prikazano u tablici zajednica).

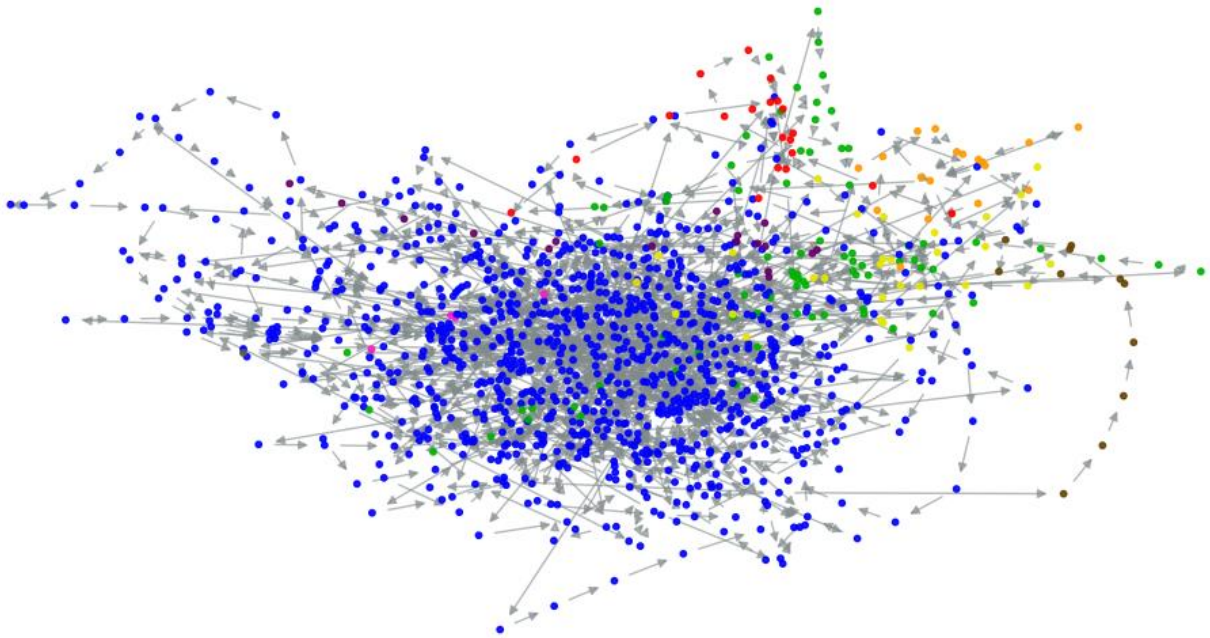
Drugi period

Proces izrade, vizualizacije i algoritam za izradu zajednica drugog perioda isti je kao i kod prvog perioda, te narednog trećeg i četvrtog perioda.

Inde	Type	Size	Value
0	set	988	{'ima', 'samoizolacije', 'velikoj', 'izolaciji', 'petoj', 'pojasnila', ...
1	set	75	{'zahtjev', 'prilaganja', 'tereti', 'predavanja', 'emaila', 'teret', ' ...
2	set	14	{'istraživanja', 'pogledati', 'budući', 'govori', 'ciljana', 'potrošač ...
3	set	4	{'ljudima', 'prvi', 'carinicima', 'doticaju'}
4	set	29	{'izgradnje', 'plaćanja', 'muškarca', 'sektoru', 'prijava', 'okviru', ...
5	set	18	{'poboljšanja', 'tržišne', 'objekta', 'računalnih', 'prezentacijskog', ...
6	set	20	{'prihvatljivih', 'velika', 'maksimalna', 'troškova', 'srednja', 'mala ...
7	set	11	{'proizvođača', 'prodajom', 'dvaju', 'sektorske', 'podnošenja', 'udruž ...
8	set	1	{'čini'}

Slika 22. Zajednice drugog perioda

Na slici 22. prikazane su zajednice mreže drugog perioda. Ovdje već možemo primjetiti, kao što će biti i dokazano kasnije, da postoji jedna zajednica sa izrazito velikim brojem čvorova dok se ostale zajednice sastoje od svega nekoliko riječi (čvorova). Ovo bi moglo ukazivati možda na neefikasnost odabranog algoritma na skupu podataka ili se jednostavno većina riječi odnosi na istu temu te su grupirane u jednu zajednicu. Promjena algoritma nije bila opcija s obzirom na izbor i brzinu/mogućnost izvođenja, no također i što smatram da je algoritam uspješno grupirao riječi iz prvog primjera tako da mislim da nije problem u algoritmu. Uglavnom, isto možemo primjetiti logiku u grupiranju zajednica, iako po meni manje vidljivu u odnosu na prvi period (ovaj i naredni periodi imaju znatno veći broj čvorova nego prvi period te je teže promatrati podjelu).



Slika 23. Vizualizacija zajednica drugog perioda

Na slici 23. prikazana je vizualizacija zajednica mreže drugog perioda. Vidimo točno da plava boja predstavlja prvu (većinsku) zajednicu, zelena drugu sa nešto manje od 100 čvorova itd. Možemo primjetiti da velika većina mreže pripada jednoj zajednici, no prema vizualizaciji i možemo primjetiti da isti i jesu susjedi (kao i čvorovi drugih zajednica).

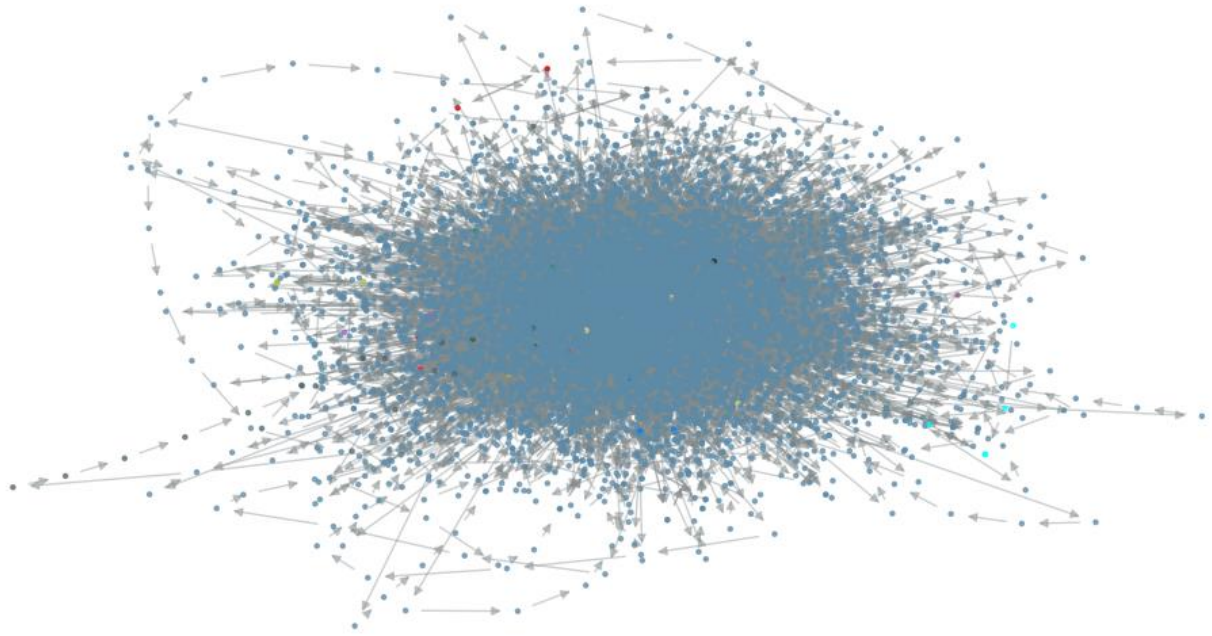
Treći period

Inde	Type	Size	Value
0	set	9502	{'utorak', 'oprezni', 'preuzeti', 'imala', 'infrastrukture', 'pokuša', ...
1	set	2	{'nacionalnih', 'parkova'}
2	set	4	{'belišće', 'mandićevac', 'petrijevci', 'samatovci'}
3	set	3	{'dozvoljen', 'lov', 'pojedinačni'}
4	set	3	{'čekati', 'propustiti', 'čekaonicu'}
5	set	3	{'šibenske', 'biskupije', 'sisačke'}
6	set	3	{'kakvi', 'trendovi', 'procijeniti'}
7	set	2	{'iznositi', 'milijarde'}
8	set	5	{'upala', 'zapanjujuće', 'slijepih', 'opstrukcija', 'crijeva'}
9	set	2	{'rokovima', 'propisanim'}
10	set	3	{'ernestinovu', 'ivanovcu', 'valpovu'}
11	set	1	{'prijave'}
12	set	2	{'vlasnici', 'plovila'}
13	set	8	{'sjemenu', 'sadnom', 'održivoj', 'pesticida', 'lovstvu', 'hrani', 'ma ...
14	set	11	{'pravcem', 'stari', 'čvor', 'čvora', 'pravac', 'obilaznim', 'goričanž ...
15	set	2	{'nadbiskup', 'đakovačkoosječki'}

Slika 24. Zajednice trećeg perioda

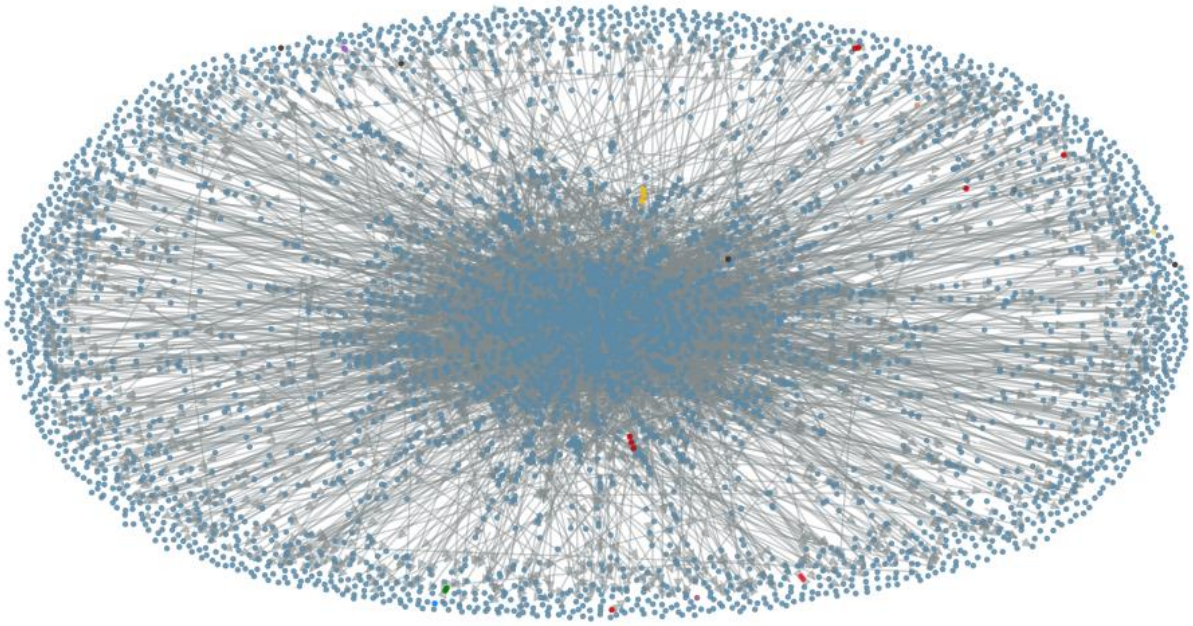
Na slici 24. prikazane su zajednice mreže trećeg perioda. Vidimo da više od 90 % čvorova pripada jednoj zajednici dok ostale imaju po par čvorova, uz izuzetak 15. zajednice koja ima 11. Podjela ima logike prema priloženom, vidimo drugu zajednicu sa riječima nacionalni park, devetu zajednicu o medicinskim simptomima i bolestima, 14. zajednica koja govori o šumarstvu, lovu i hrani itd.

Za probu, testirana je vizualizacija cijele mreže i dobiven je slijedeći rezultat:



Slika 25. Vizualizacija zajednice trećeg perioda

Na slici 25. (ne)vidimo vizualizaciju mreže trećeg perioda. Naime, čvorova ima jako mnogo te većina pripada jednoj zajednici tako da ne možemo zapravo vidjeti nekaku podjelu u zajednice nad ovom mrežom. Iz tog razloga, uzeto je 5000 nasumičnih čvorova navedene mreže i vizualizirano kako bi se mogla možda primjetiti nekakva podjela.

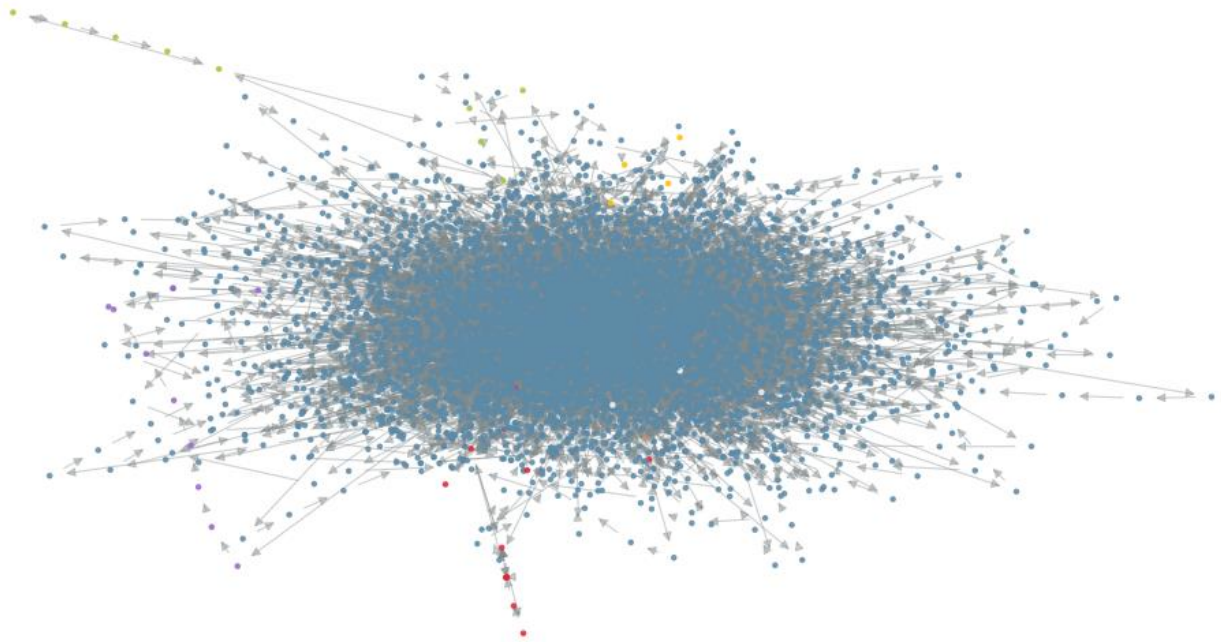


Slika 26. Vizualizacija zajednica trećeg perioda

Na slici 26. nalazi se nešto bolja vizualizacija zajednica mreže trećeg perioda. Iako stvarno većina čvorova pripada jednoj zajednici, možemo barem primjetiti par instanci drugih zajednica, koje su također same za sebe susjedi.

Četvrti period

Prilikom izračuna zajednice četvrtog perioda, izračunato je 9 zajednica, od kojih također većina čvorova pripada jednoj zajednici, što je dokazano također narednom vizualizacijom:



Slika 27. Vizualizacija zajednice četvrtog perioda

Na slici 27. nalazi se vizualizacija zajednica mreže četvrtog perioda. Možemo primjetiti prevladavanje jedne zajednice, kao i kod prošlog perioda, ali također, iako u manjoj mjeri, i drugog perioda. Riječi i teme prvog perioda bile su dosta više raznolike (nije se pisalo o koronavirusu) te su zajednice koliko toliko pravilno raspoređene prema broju čvorova, no možemo primjetit porastom broja članaka a ujedno i ključnih riječi vezanih uz korona virus da se dotična jedna zajednica povećava, dok ove druge (koje većinom ne sadrže riječi vezane uz koronu, što se moglo primjetiti po tablicama) imaju tek nekoliko čvorova vezanih uz određenu posebnu tematiku. U jednu ruku rekla bih da grupiranje i ima smisla, jer se riječi vezane uz koronu npr. ne pojavljuju u drugim grupama (zajednicama) no možda bi se uz neki komputacijski više zahtjevan algoritam mogao dobiti precizniji rezultat (u smislu raspodjele velikih zajednica u više manjih).