Zadatak 1. Analiza mreže riječi s Twittera

Analiza mreže

1. Odredite broj čvorova i grana takve mreže.

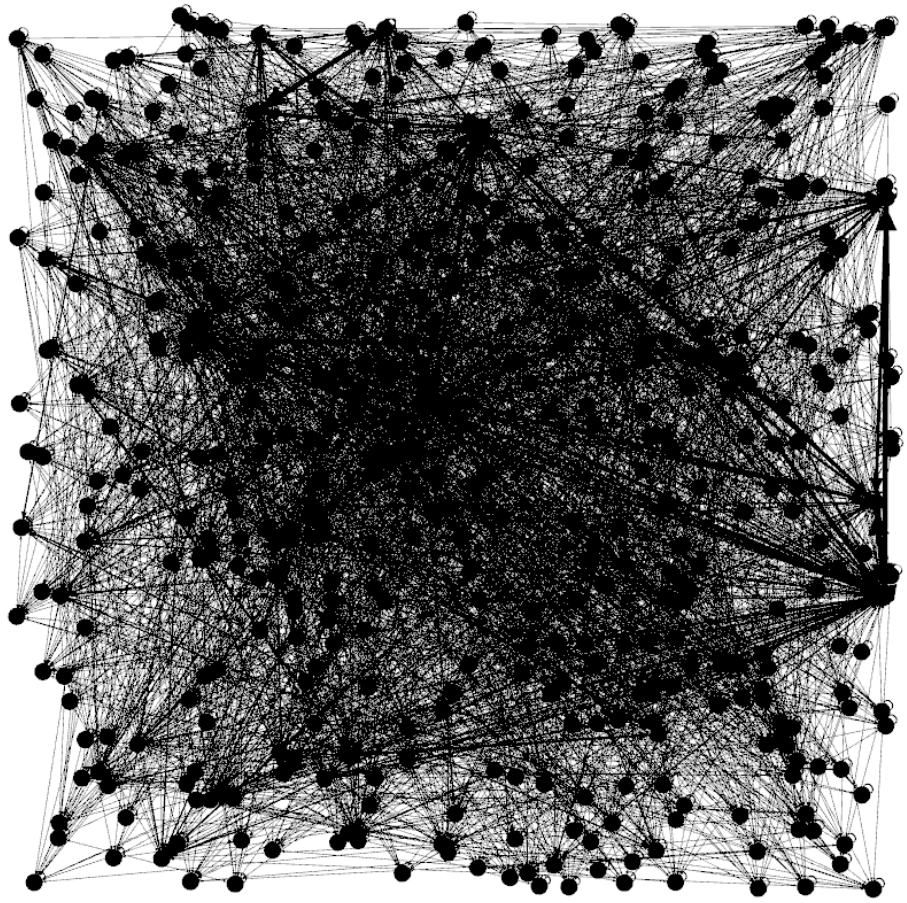
Broj čvorova i grana vidljiv je u alatu Gephi, u odjeljku *Context*

Broj čvorova: 499

Broj grana: 7693

2. Prije početka analize prikažite početnu sliku mreže.

Slika početne mreže vidljiva je u alatu Gephi, u odjeljku *Graph*.



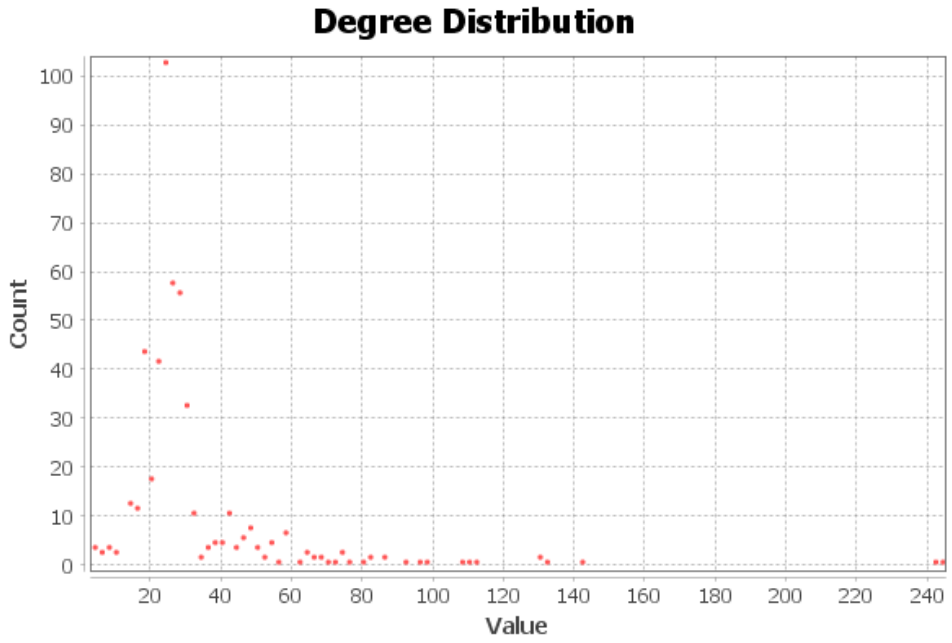
3. Nakon toga odredite prosječni stupanj čvora u mreži i prikažite distribuciju stupnjeva čvorova u mreži. Komentirajte dobivenu distribuciju.

Prosječni stupanj čvora u mreži računa se u alatu Gephi pokretanjem funkcije *Average Degree* u odjeljku *Statistics -> Network Overview*. Uz izračunati prosječni stupanj, vidljiva je i distribucija stupnjeva čvorova u mreži.

Prosječni stupanj čvora: 15.417

Results:

Average Degree: 15,417



Iz navedene distribucije može se primijetiti da je stupanj 24 prisutan u najviše čvorova, njih 103, dok je najveći stupanj nekog čvora, 244, prisutan u samo 1 čvoru. Ove točne vrijednosti mogu se iščitati namještanjem *Degree Range* parametra koji je detaljnije opisan u 5. zadatku.

4. Odredite:

1) gustoću mreže i koliko mreža mora imati grana da bi gustoća mreže bila 1 (izračunajte, potrebno je objasniti izračun, tj. postupak);

Gustoća mreže računa se u alatu Gephi pokretanjem funkcije *Graph Density* u odjeljku *Statistics -> Network Overview*.

Gustoća mreže: 0.031

Formula za izračunavanje gustoće usmjerene mreže je:

$$D = \frac{|E|}{2\binom{|V|}{2}} = \frac{|E|}{|V|(|V|-1)}$$

iz čega dobivamo broj grana (E) koji je potreban da bi gustoća (D) bila 1, uz isti broj čvorova (V):

E = 248502

2) promjer mreže (engl. *diameter*);
Promjer mreže računa se u alatu Gephi pokretanjem funkcije *Network Diameter* u odjeljku *Statistics -> Network Overview*.

Promjer mreže: 6

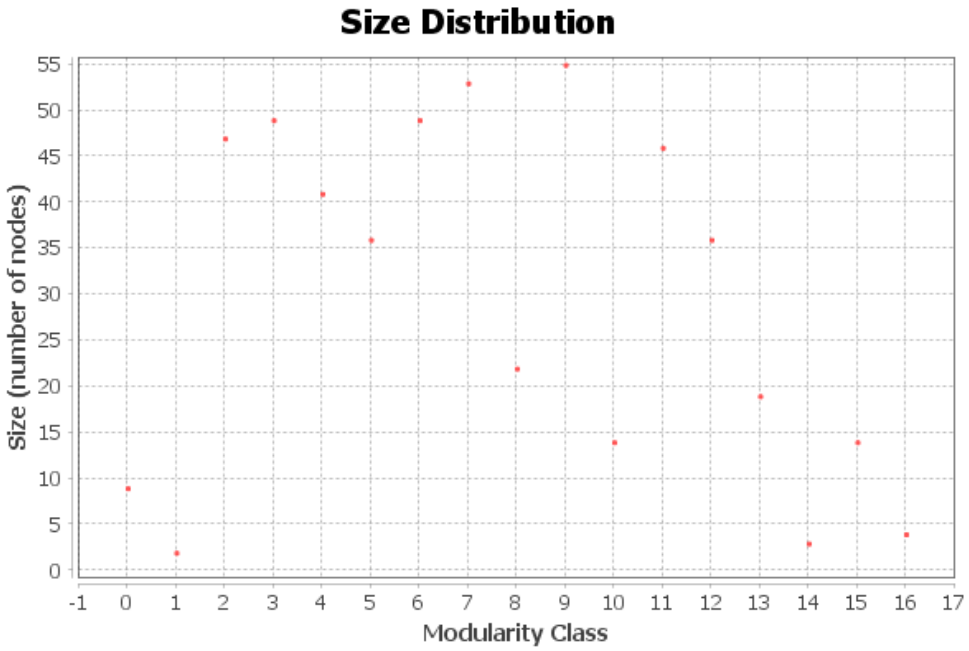
3) koeficijent modularnosti mreže (engl. *modularity*) i navedite koliko je detektirano zajednica (engl. communities) u mreži, priložite sliku te komentirajte.
Koeficijent modularnosti mreže računa se u alatu Gephi pokretanjem funkcije *Modularity* u odjeljku *Statistics -> Network Overview*.

Koeficijent modularnosti mreže: 0.673

Broj detektiranih zajednica: 17

Results:

Modularity: 0,673
Modularity with resolution: 0,673
Number of Communities: 17



Iz navedene distribucije može se primijetiti da zajednica s indeksom 1 ima najmanje čvorova, njih 2, dok zajednica s indeksom 9 ima najviše čvorova, njih 55.

5. Početnom detekcijom zajednica postoji određeni broj zajednica koje imaju premali broj članova te je potrebno „filtrirati šum“. Filtrirajte mrežu tako da eliminirate oko 10% (procijenite postotak proizvoljno) čvorova po parametru „*Degree Range*“. Opišite dobivene vrijednosti i postupak.
Filtriranje mreže provodi se u alatu Gephi namještanjem *Degree Range* parametra u odjeljku *Filters -> Topology.*

	Prije	Poslije	Postotak (Visible)
Broj čvorova:	499	460	92,18%
Broj grana:	7693	7394	96,11%

Namještanjem minimuma u parametru *Degree Range* na 18, eliminirano je 39 čvorova iz mreže, odnosno 7.82% od ukupnog broja čvorova. Uz tih eliminiranih 39 čvorova, broj grana smanjio se za 299.

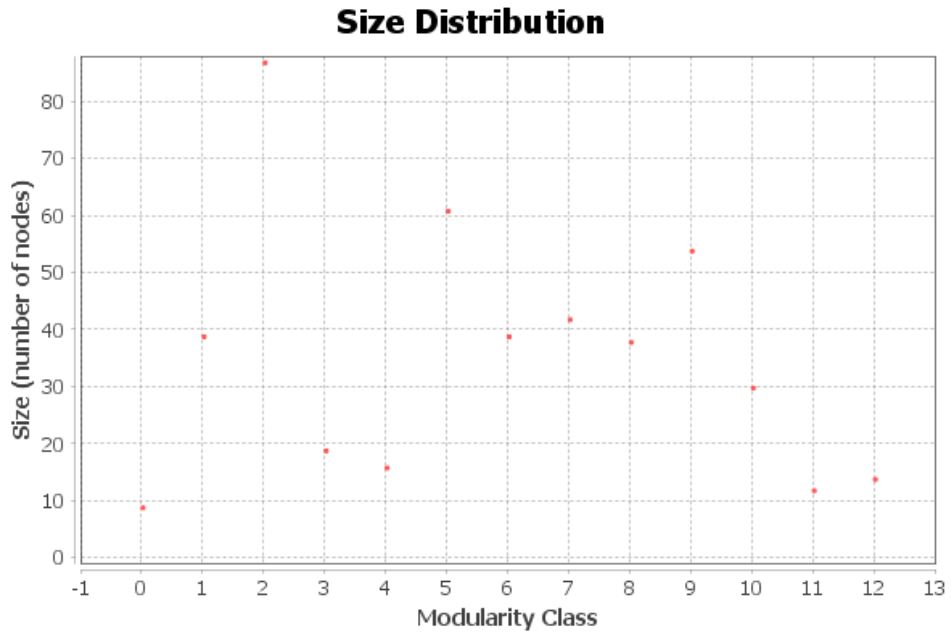
6. Ponovno odredite koeficijent modularnosti mreže (engl. *modularity*) i navedite koliko je detektirano zajednica (engl. *communities*) u mreži, priložite sliku te komentirajte.

Koeficijent modularnosti mreže: 0.271

Broj detektiranih zajednica: 13

Results:

Modularity: 0,271
Modularity with resolution: 0,271
Number of Communities: 13



Iz navedene distribucije može se primijetiti da zajednica s indeksom 0 ima najmanje čvorova, njih 9, dok zajednica s indeksom 2 ima najviše čvorova, njih 89. U oba slučaja, radi se značajnom povećanju broja čvorova u sada najmanjoj i najvećoj zajednici, u usporedbi s rezultatima prije filtriranja mreže.

7. Također ponovno odredite i komentirajte:

1) gustoću mreže;

Gustoća mreže: 0.035

Gustoća mreže povećala se za 0.004 u usporedbi s gustoćom prije filtriranja mreže.

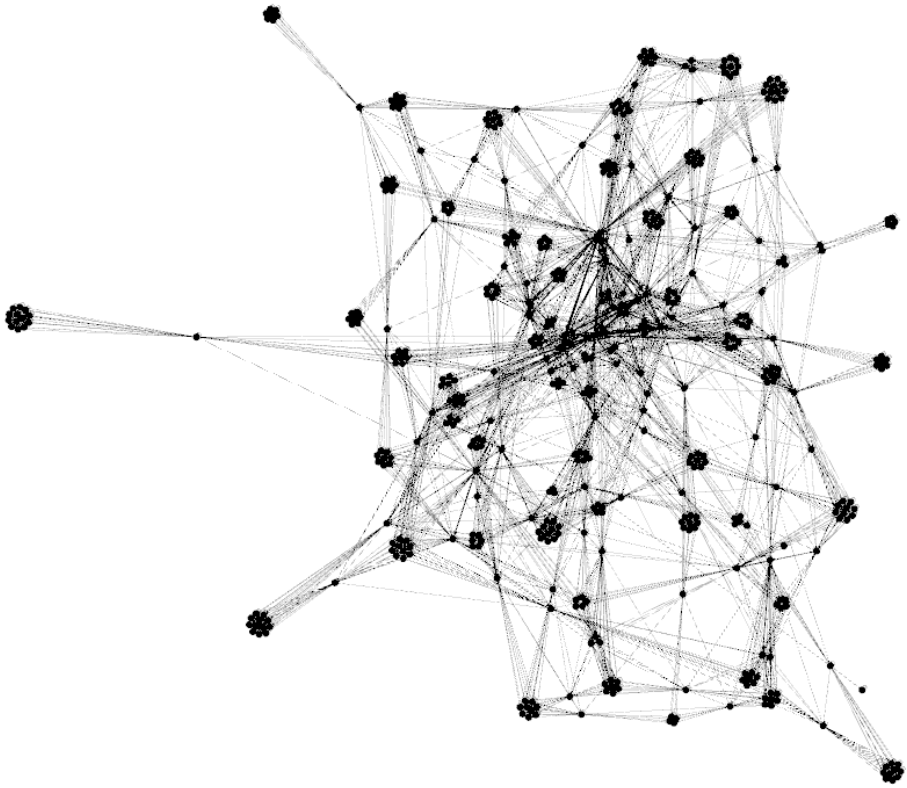
2) promjer mreže (engl. *diameter*);

Promjer mreže: 5
Promjer mreže smanjio se za 1 u usporedbi s promjerom prije filtriranja mreže.

<u>Komponente:</u>

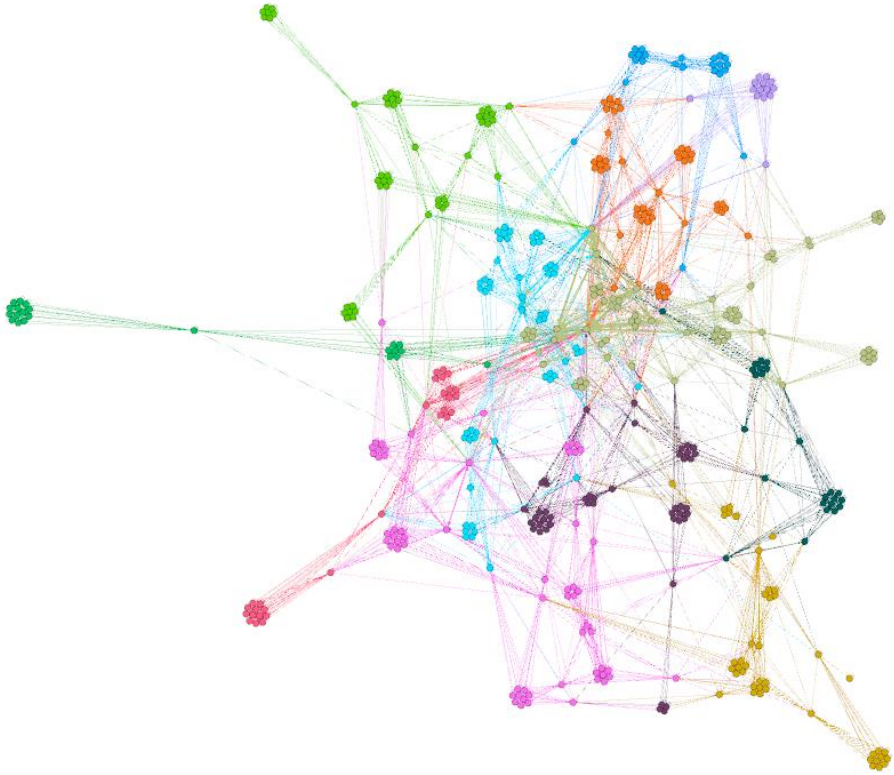
8. Za prikaz mreže definirajte jedan od željenih *layout-a Fruchterman Reingold* ili *Force Atlas*, te priložite dobivenu sliku.

Namještanje *layouta* u alatu Gephi nalazi se u odjeljku *Layout*, gdje je u ovom slučaju odabran *Force Atlas*.



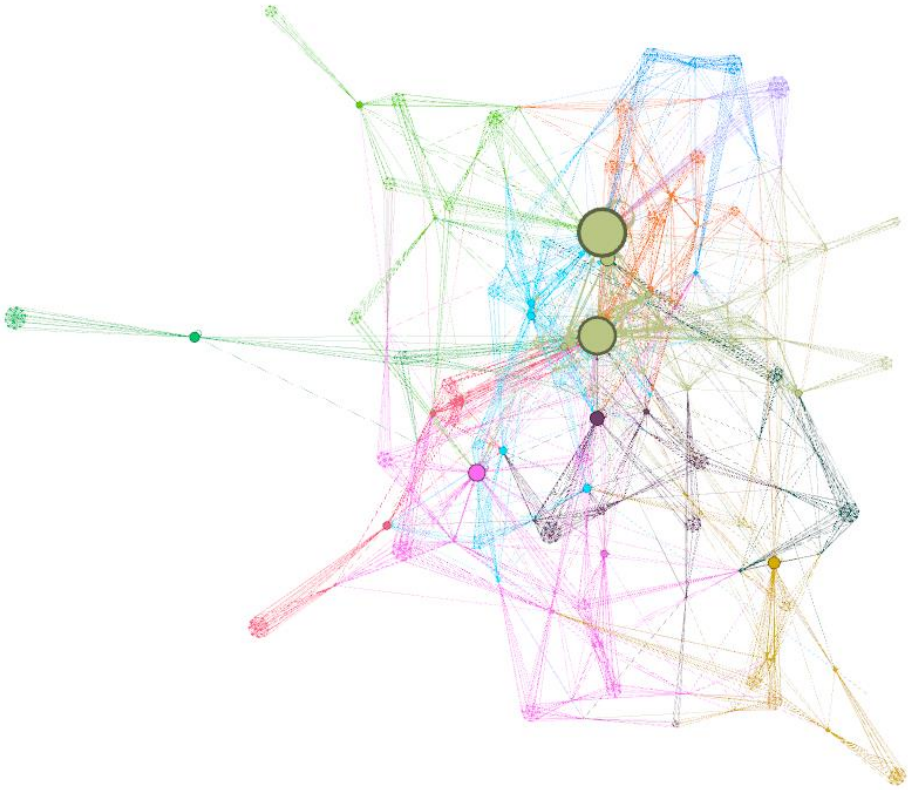
9. Odaberite karticu Appearance->Nodes->Attribute odabir boja te postavite parametar na Modularity Class za bojanje zajednica u mreži. Priložite dobivenu sliku.

Bojanje zajednica u mreži u alatu Gephi nalazi se u odjeljku *Appearance -> Nodes -> Partition*, gdje je potrebno odabrati *Modularity Class* i definirati paletu boja za broj čvorova u mreži.



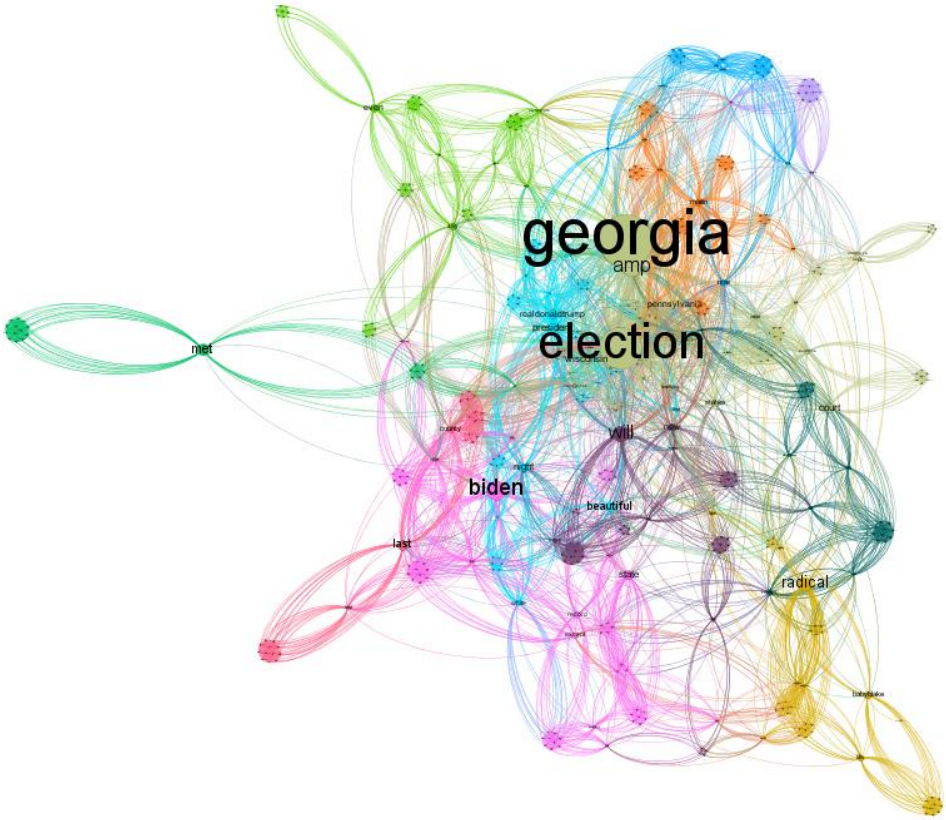
10. Nadalje, kliknite na Size ikonicu, odaberite parametar Betweenness Centrality te postavite parametre min-max u rasponu od 1-70 (možete i samostalno odrediti raspon po želji), pa priložite dobivenu sliku.

Definiranje raspona broja čvorova između koji se čvor nalazi u alatu Gephi nalazi se u odjeljku *Appearance -> Nodes -> Ranking* (uz prethodni klik na *Size* ikonicu), gdje je potrebno odabrati *Betweenness Centrality* i definirati *Min size:* 1 i *Max size:* 70.



11. Nakon toga odaberite karticu *Preview* te označite parametar *Show Labels* i kliknite na *Refresh*. Priložite dobivenu sliku. Također na slici označite zajednice koje možete detektirati i komentirajte dobivene rezultate.

Pregled mreže u alatu Gephi ostvaruje se u klikom na karticu *Preview* i označavanjem kvačice za parametar *Show Labels* u odjeljku *Node Labels* te potom klikom na *Refresh*.



U pregledu ove mreže najistaknutiji su čvorovi „georgia“ i „election“, zbog svojih najvećih stupnjeva naspram ostalih čvorova. Svih 17 zajednica i dalje je obojano različitim bojama te su vidljivi u ovom pregledu mreže.

12. Odaberite karticu *Data Laboratory* te odredite, usporedite, komentirajte i interpretirajte dobivene rezultate po parametrima:

Pregled rezultata iz tablice u alatu Gephi ostvaruje se u klikom na karticu *Data Laboratory* te odabirom parametra po kojem želimo poredati vrijednosti.

1) *Degree* top 5 s najvećim vrijednostima

1. election (244)
2. georgia (242)
3. amp (142)
4. will (132)
5. pennsylvania (130)

Čvorovi „election" i „georgia" imaju *Degree* s vrijednostima preko 240, što znači da su oba povezana sa skoro pola od ukupnog broja preostalih čvorova.

2) *Closeness Centrality top 5 s najvećim vrijednostima*

1. children (1.0)
2. insanity (1.0)
3. mention (1.0)
4. months (1.0)
5. saying (1.0)

Closeness Centrality top 5 s najmanjim vrijednostima

1. thanksgiving (0.2355)
2. tcoxtywknph (0.2412)
3. band (0.2636)
4. dancing (0.2636)
5. froz (0.2636)

S obzirom da su vrijednosti *Closeness Centrality*-a u rasponu između 0.263591 i 1, to pokazuje da se radi o dobro centraliziranoj mreži, jer je najveći prosječan broj skokova od nekog čvora ove mreže do ostatka mreže tek 1, dok su vrijednosti za ostale čvorove manje ili jednake.

3) *Betweenness Centrality top 5 s najvećim vrijednostima*

1. georgia (59214.5777)
2. election (46429.0876)
3. biden (20767.362)
4. amp (17916.2641)
5. will (17140.092)

Od ovih 5 čvorova, 4 ih se nalazi na *Degree* top 5 ljestvici, što pokazuje da se radi o dobro povezanoj mreži, jer se čvorovi s najviše veza najčešće pojavljuju na najkraćem putu u mreži. Ovdje se najviše ističu

čvorovi „georgia“ i „election“, s vrijednostima *Betweenness Centrality*-a višestruko većim od preostala 3 čvora na listi.

4) *PageRank top 5 s najvećim vrijednostima*

1. georgia (0.0115)
2. election (0.0108)
3. amp (0.0066)
4. will (0.0064)
5. biden (0.0061)

Čvorovi „georgia“ i „election“ imaju najveće vrijednosti *PageRank*-a, što znači da su dobro povezani s čvorovima koji su jako dobro povezani, no uzimajući u obzir samo ulazne veze (za razliku od svih veza koje se uzimaju u obzir kod *Eigenvector Centrality*-a). Kako se svih 5 čvorova iz ove liste nalazi na *Betweenness Centrality* top 5 ljestvici, a 4 čvora na *Degree* top 5 ljestvici, to ponovno pokazuje da se radi o dobro povezanoj mreži.

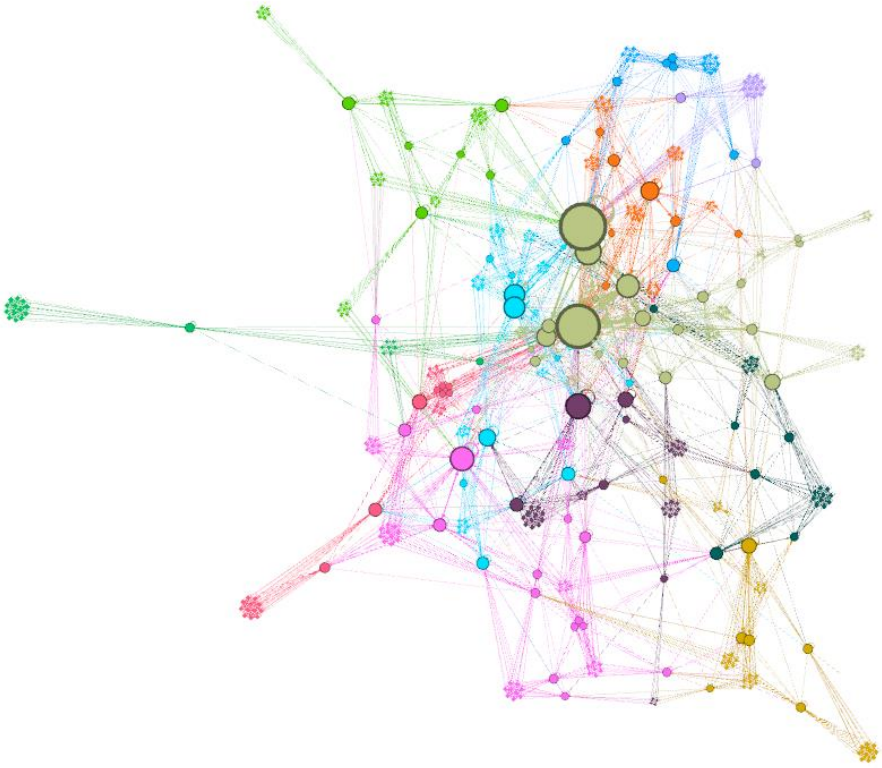
5) *Eigenvector Centrality* top 5 s najvećim vrijednostima

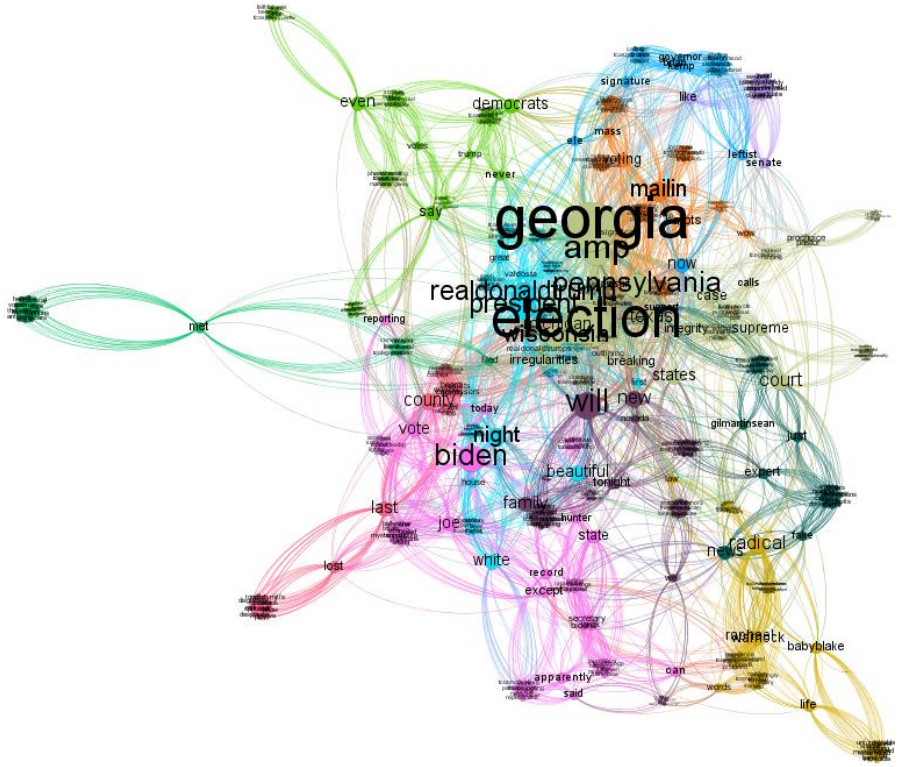
1. election (1.0)
2. georgia (0.8708)
3. pennsylvania (0.6123)
4. amp (0.581)
5. wisconsin (0.5235)

Čvorovi „election“ i „georgia“ imaju najveće vrijednosti *Eigenvector Centrality*-a, što znači da su dobro povezani s čvorovima koji su jako dobro povezani (sve veze se uzimaju u obzir). Kako ta ista dva čvora imaju najveće vrijednosti *Degree*-a i *Betweenness Centrality*-a, to ponovno pokazuje da se radi o dobro povezanoj mreži. Ostali čvorovi iz *Eigenvector Centrality* top 5 ljestvice, osim čvora „wisconsin“, također se nalaze u *Degree* i/ili *Betweenness Centrality* top 5 ljestvicama.

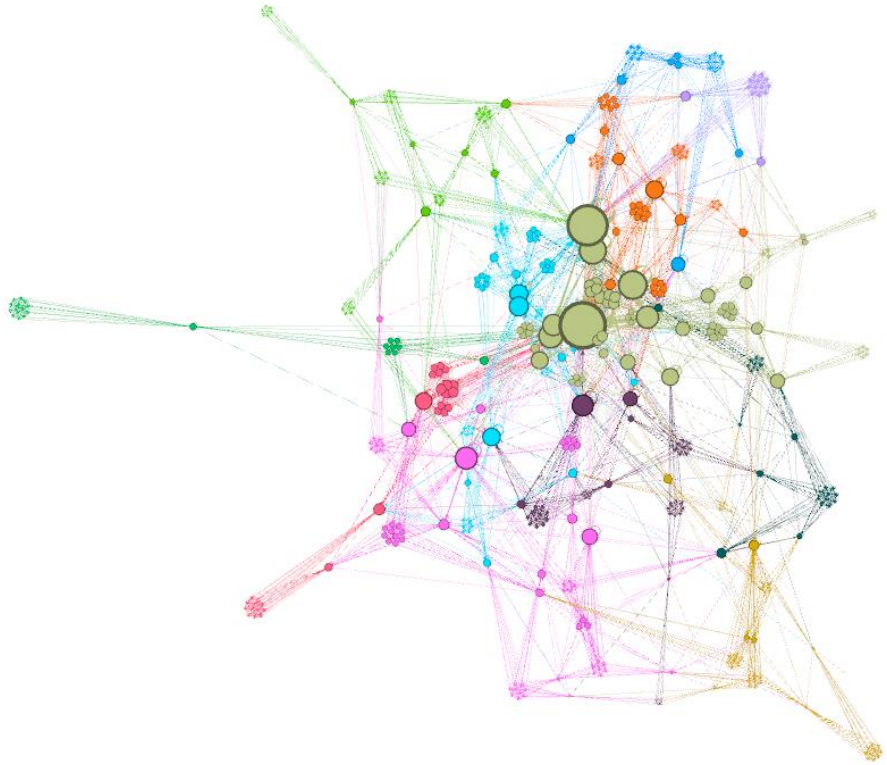
13. Kreirajte i usporedite slike mreža kada veličinu čvorova definirate korištenjem *PageRank* i *Eigenvector Centrality*, te priložite slike mreža. Odredite i komentirajte koja od analiziranih mjera najbolje određuje najvažnije čvorove u mreži i zašto.

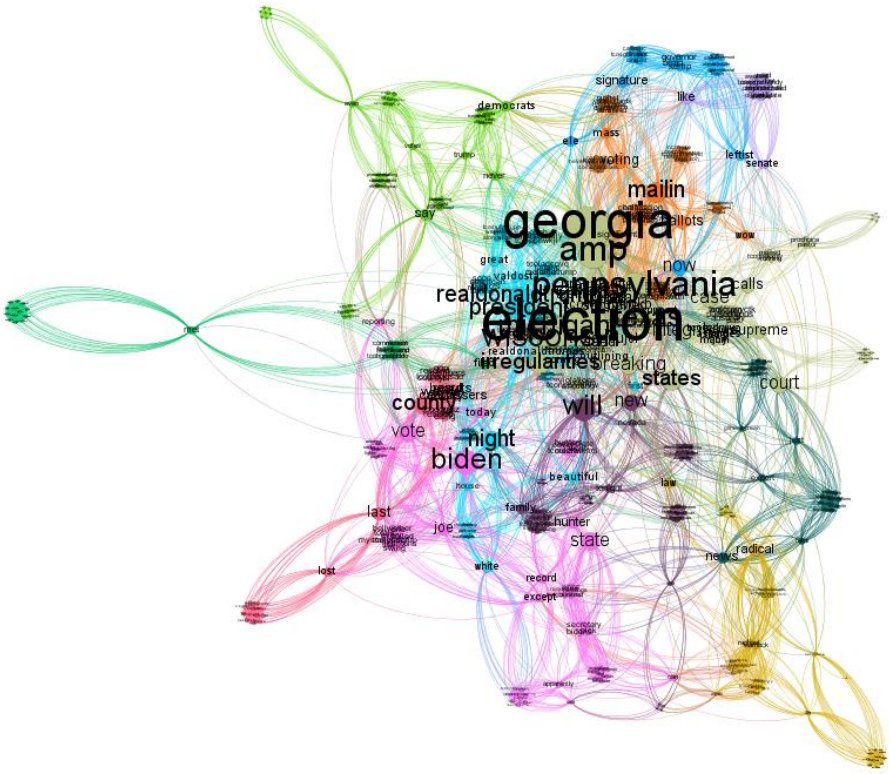
1) *PageRank*




even
democrats
signature
like
mass
voting
never
say
georgia
amp
now
pennsylvania
case
calls
election
wisconsin
supreme
reporting
irregularities
breaking
states
court
county
today
will
new
vote
night
biden
beautiful
tonight
expert
last
family
joe
hunter
state
radical
white
record
except
lost
babyblake
apparently
said
can
life

2) ***Eigenvector Centrality***





Iz navedenih slika može se primjetiti da mjera *PageRank* bolje određuje najvažnije čvorove u mreži. Na mreži definiranom mjerom *PageRank* prisutno je manje čvorova koji nemaju toliku važnost, što je potvrđeno i u teoriji, jer *PageRank* uzima u obzir samo ulazne veze čvorova, dok *Eigenvector Centrality* uzima u obzir sve veze (ulazne i izlazne).

Zadatak 2. Kreiranje sentiment analize prikupljenih poruka s Twittera

14. Kao rezultat ispisati:

1) Top 3 riječi kod pozitivnog segmenta

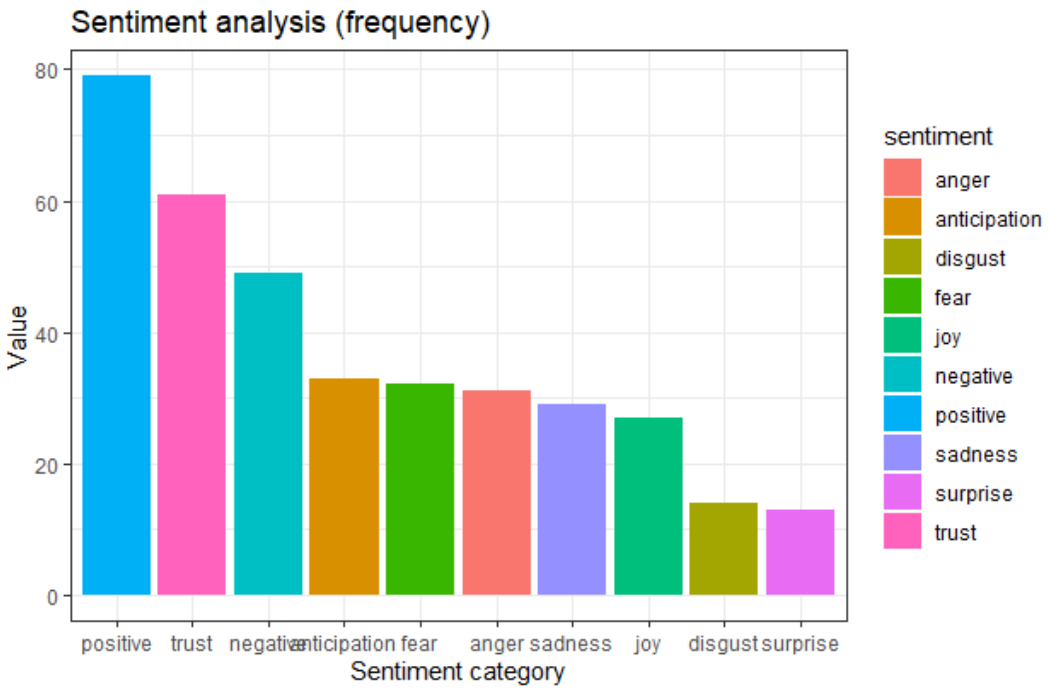
1. president
2. democracy
3. beautiful

2) Top 3 riječi kod negativnog segmenta

1. assault
2. words
3. case

Pokretanjem skripte *01_Twitter_core_addon - sent.R*, uz generiranu *termMatrix.cvs* datoteku, koja je služila kao izvor za Zadatak 1, dobivena je i sentiment analiza koja služi kao izvor za ovaj zadatak. U alatu RStudio, u odjeljku *Environment* potrebno je otvoriti varijablu *sentiment_val* u kojoj se nalaze riječi s pripadajućim sentimentom i učestalošću (*freq*). U gornjim top 3 listama nalaze se riječi iz pozitivnog i negativnog segmenta s najvećom učestalošću.

15. Ukupan ranking sentimenta koristeći *word frequency* (svih emocija) – graf slika (objasniti ukratko rezultate)



Iz provedene sentiment analize može se primijetiti da je sentiment pozitive (engl. *positive*) najviše zastupljen, s ukupnom učestalošću jednakom 79, dok je sentiment iznenađenja (engl. *surprise*) najmanje zastupljen, s ukupnom učestalošću jednakom 13.