ΤΗΜΜΥ

ΕΑΡΙΝΟ ΕΞΑΜΗΝΟ 2020

ΕΡΓΑΣΙΑ ΕΞΑΜΗΝΟΥ ΣΤΗΝ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

«ΑΝΑΛΥΣΗ ΑΜΕΡΙΚΑΝΙΚΩΝ ΕΚΛΟΓΩΝ ΑΠΟ ΤΗΝ ΣΚΟΠΙΑ ΤΟΥ DATA MINING»

Νίκος Πλέσσας – ΑΕΜ: 615

ΟΙ ΥΠΟΨΗΦΙΟΙ ΠΟΥ ΘΑ ΕΞΕΤΑΣΟΥΜΕ

Donald Trump

* POTUS
* Republican

Bernie Sanders

* Senator of Vermont
* Democrat

DATASETS

* 10,000 tweets από κάθε υποψήφιο
* 2.000 tweets από το @CNNpolitics
* 1.500 τίτλοι και περιλήψεις άρθρων από διάφορα μέσα

Όλα ανακτήθηκαν μέσω **crawlers** που γράφτηκαν σε python με χρήση των βιβλιοθηκών **GetOldTweets** & **Feedparser**

PYTHON.

- Οι περισσότεροι κώδικες γράφτηκαν στο Google Colab.

- Έγινε χρήση των βιβλιοθηκών της Python: Pandas, Sklearn,NLTK,TextBlob,Matplotlib

ΠΕΡΙΓΡΑΦΗ ΕΡΓΑΣΙΑΣ (1)

ΜΕΡΟΣ ΠΡΩΤΟ: ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ & ΕΞΑΓΩΓΗ ΣΥΜΠΕΡΑΣΜΑΤΩΝ

- **Clustering**: Ομαδοποίηση των tweets και προσδιορισμός θεματολογίας – πολιτικών θέσεων (Kmeans)
- **Sentiment Analysis**: Κατηγοριοποιήση των tweets σε θετικά-αρνητικά-ουδέτερα βάσει της φρασεολογίας τους
- **Δημοφιλία** κάθε υποψήφιου

ΠΕΡΙΓΡΑΦΗ ΕΡΓΑΣΙΑΣ (2)

ΜΕΡΟΣ ΔΕΥΤΕΡΟ: ΑΝΑΠΤΥΞΗ ΕΦΑΡΜΟΓΩΝ

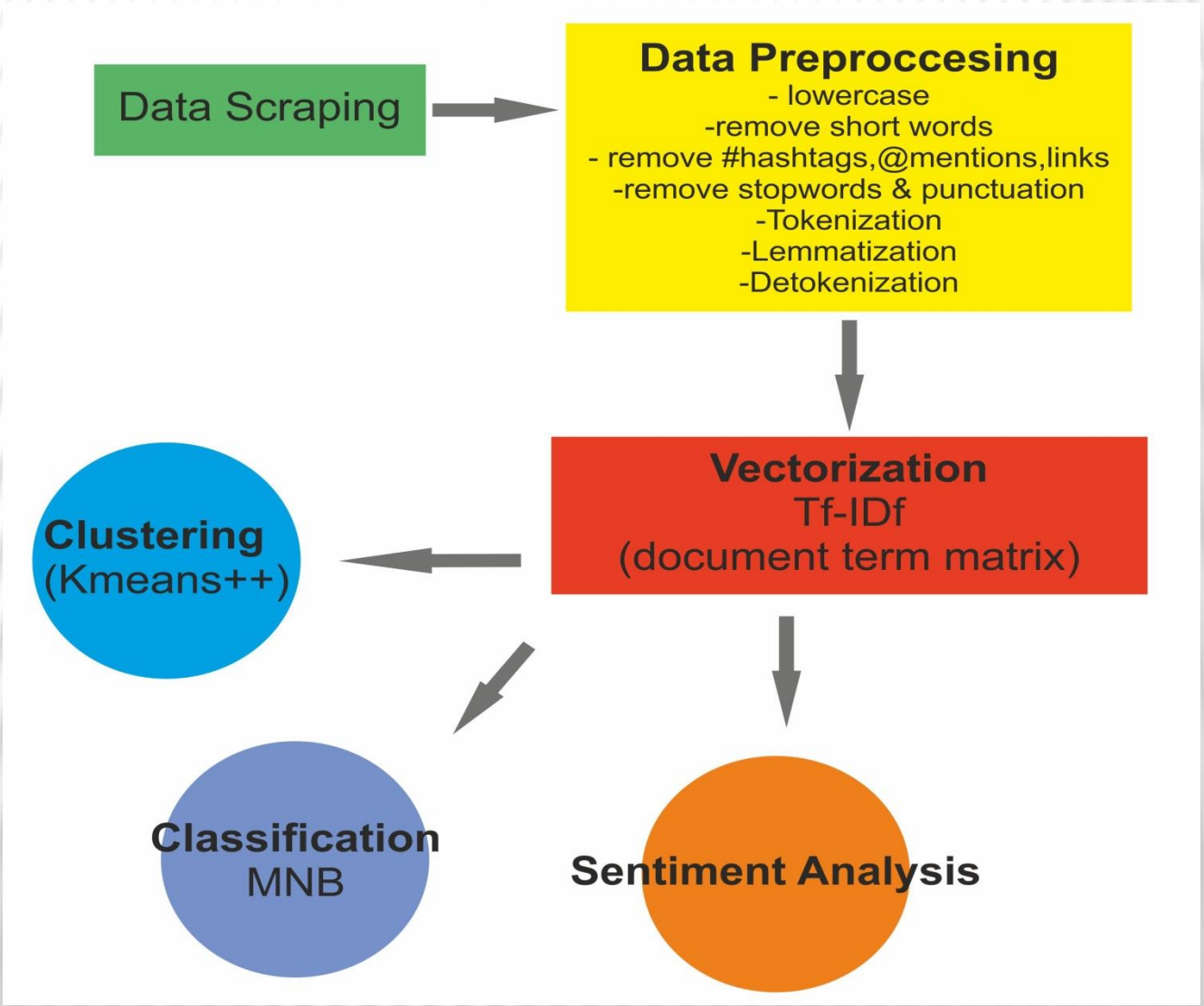
- **Classifier**: Δημιουργία ταξινομητή που θα βρίσκει τον «ιδιοκτήτη» κάθε tweet
- **Real Time Sentiment Analysis**: Ανάκτηση tweets και ανάλυση δημοφιλίας σε πραγματικό χρόνο από tweets χρηστών που αφορούν κάθε υποψήφιο

NATURAL LANGUAGE PROCCESSING (NLP)

NLP **pipeline** (data preproccesing):

- Μετατροπή σε **πεζά**
- Αφαίρεση μικρών λέξεων (<2 χαρακτήρες)
- Αφαίρεση hashtags,links,usernames κτλ
- **Tokenization**
- Αφαίρεση **stopwords**
- Lemmatization

NLP PIPELINE



ΜΕΡΟΣ 1ο

- **Ανάλυση** των tweets των υποψηφίων
- Τι **συμπεράσματα** μπορούμε να βγάλουμε για τις πολιτικές/προγραμματικές τους θέσεις μόνο από το twitter feed τους;
- Τι **γνώμη** έχουν τα media & οι χρήστες για αυτούς;

@REALDONALDTRUMP TWEETS ANALYSIS

- KMEANS Clustering (k = 40 clusters)

Cluster 3:
democrat
nothing
nothing democrat
impeachment
left
radical
radical left
want
party
hoax

Cluster 19:
thank
kag2020
maga
love
thank maga
great
approval rating
approval
working hard
rating republican

Cluster 18:
hillary
crooked
crooked hillary
clinton
comey
fbi
hillary clinton
james comey
james
mccabe

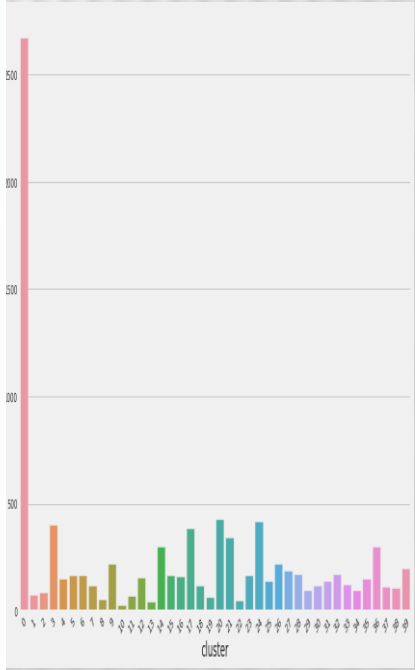
Cluster 17:
god bless
bless
god
bless usa
bless people
bless america
bless maga
usa
people
california

Cluster 0:
american
country
people
never
good
preside0:
nt
happy
day
time

Cluster 13:
endorsement
vet
total endorsement
military
total
strong
complete
crime
military vet
border

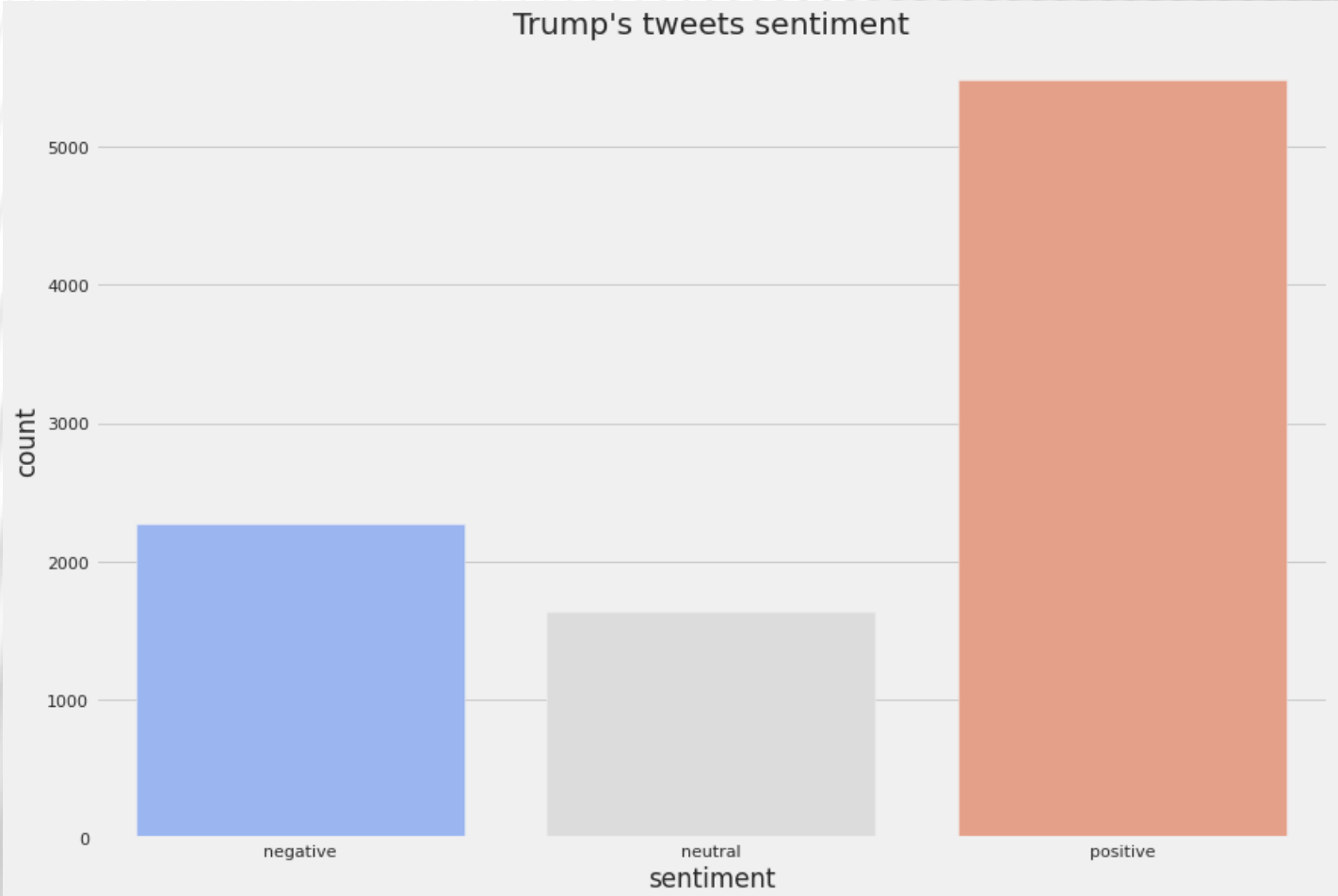
Cluster 15:
crazy bernie
crazy
joe
look
sleepy joe
sleepy
going
look like
great
like

Cluster 9:
border
wall
security
border security
southern border
southern
democrat
immigration
must
law



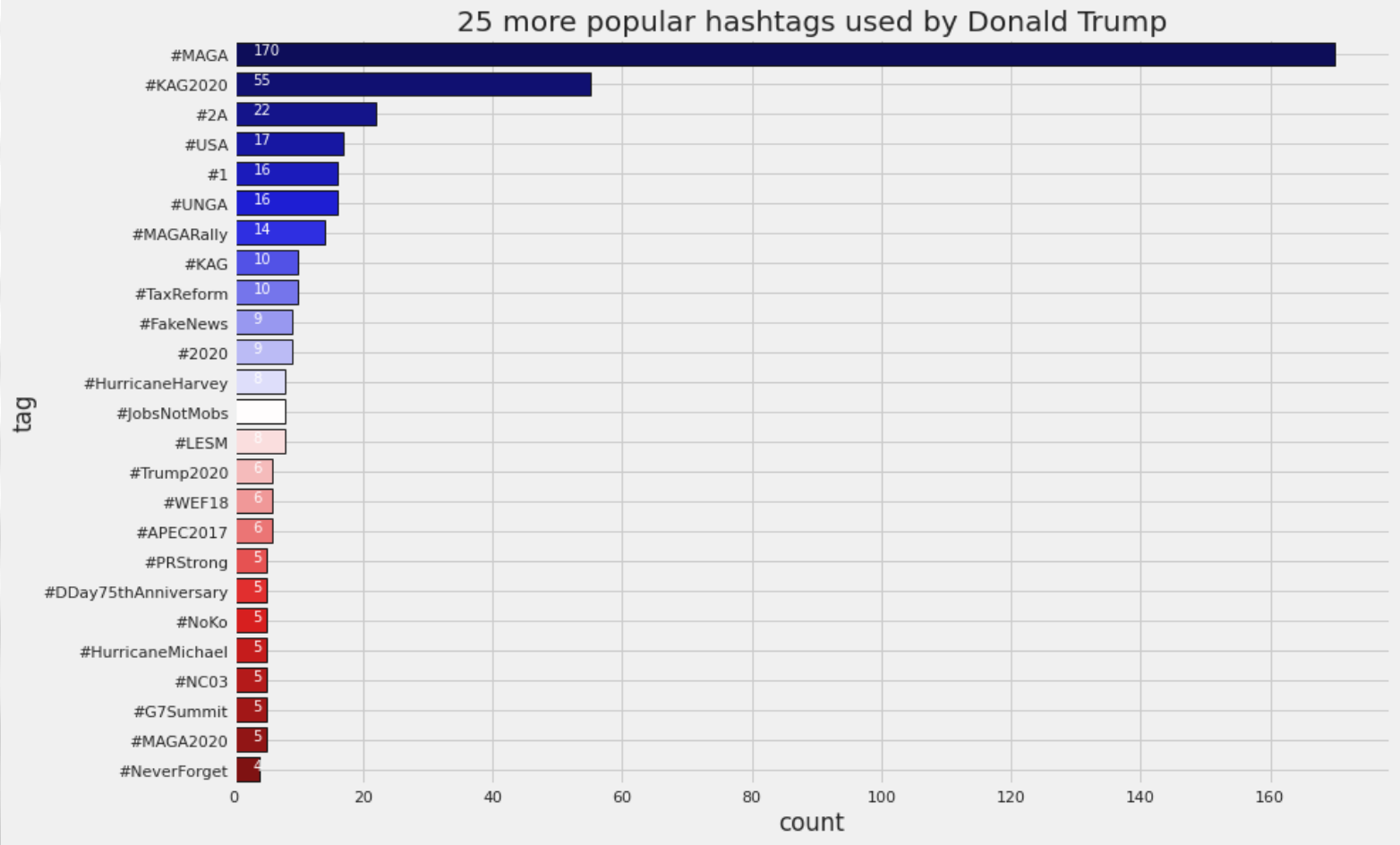
@REALDONALDTRUMP TWEETS ANALYSIS

- Sentiment analysis (TextBlob)

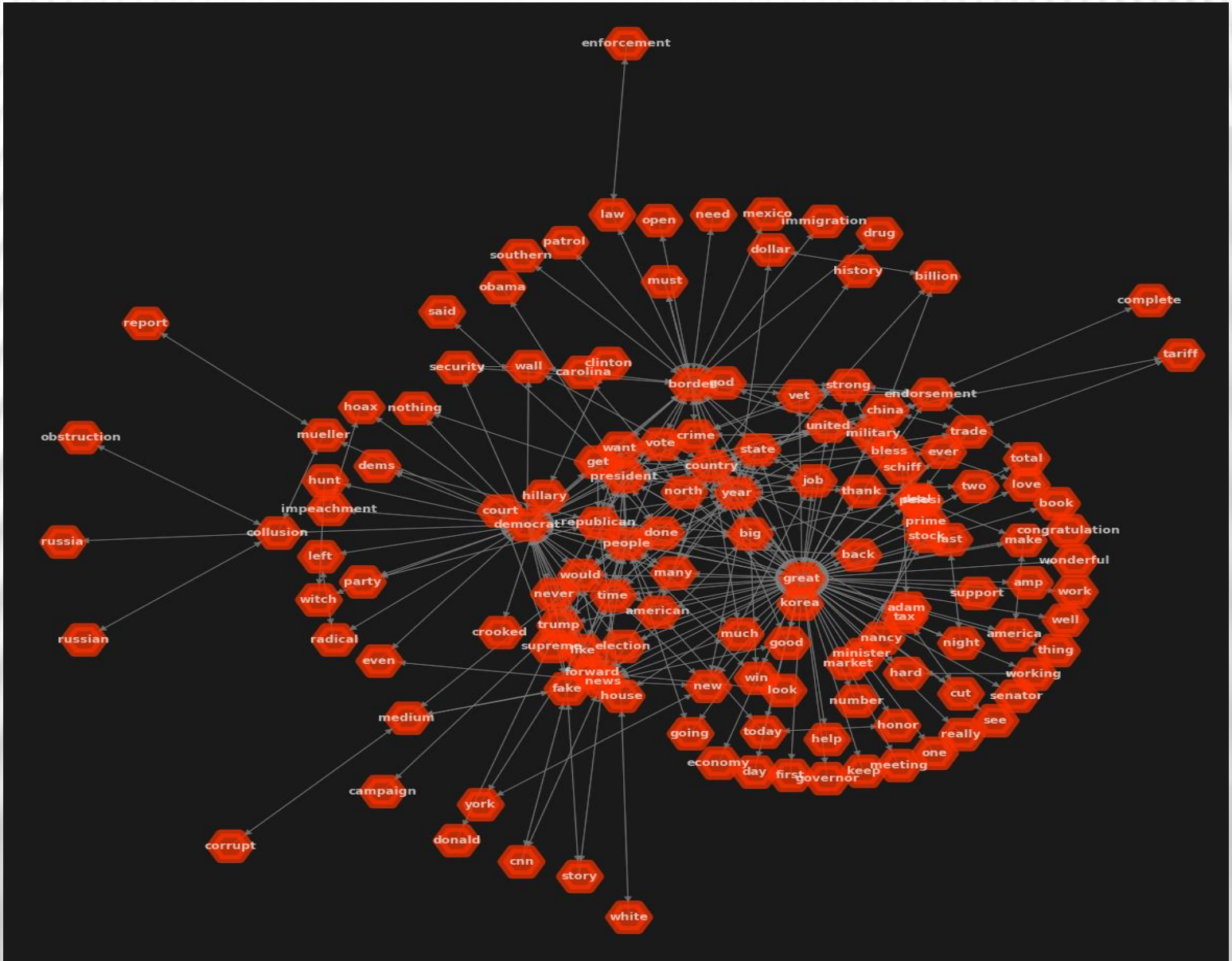


@REALDONALDTRUMP TWEETS ANALYSIS

- 25 most popular hashtags used by Trump

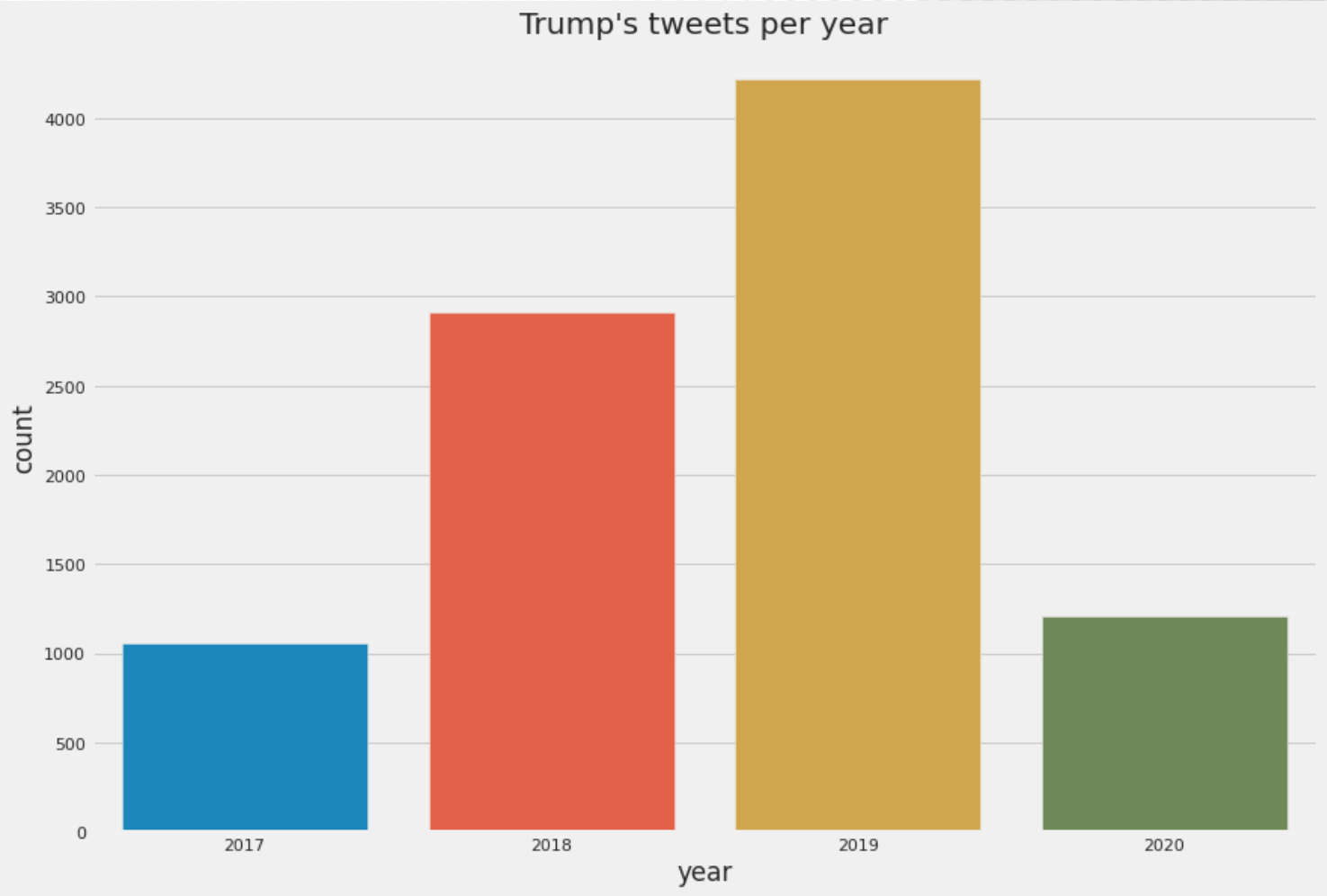


NETWORK CONNECTION GRAPH (FREQ>80 WORDS)



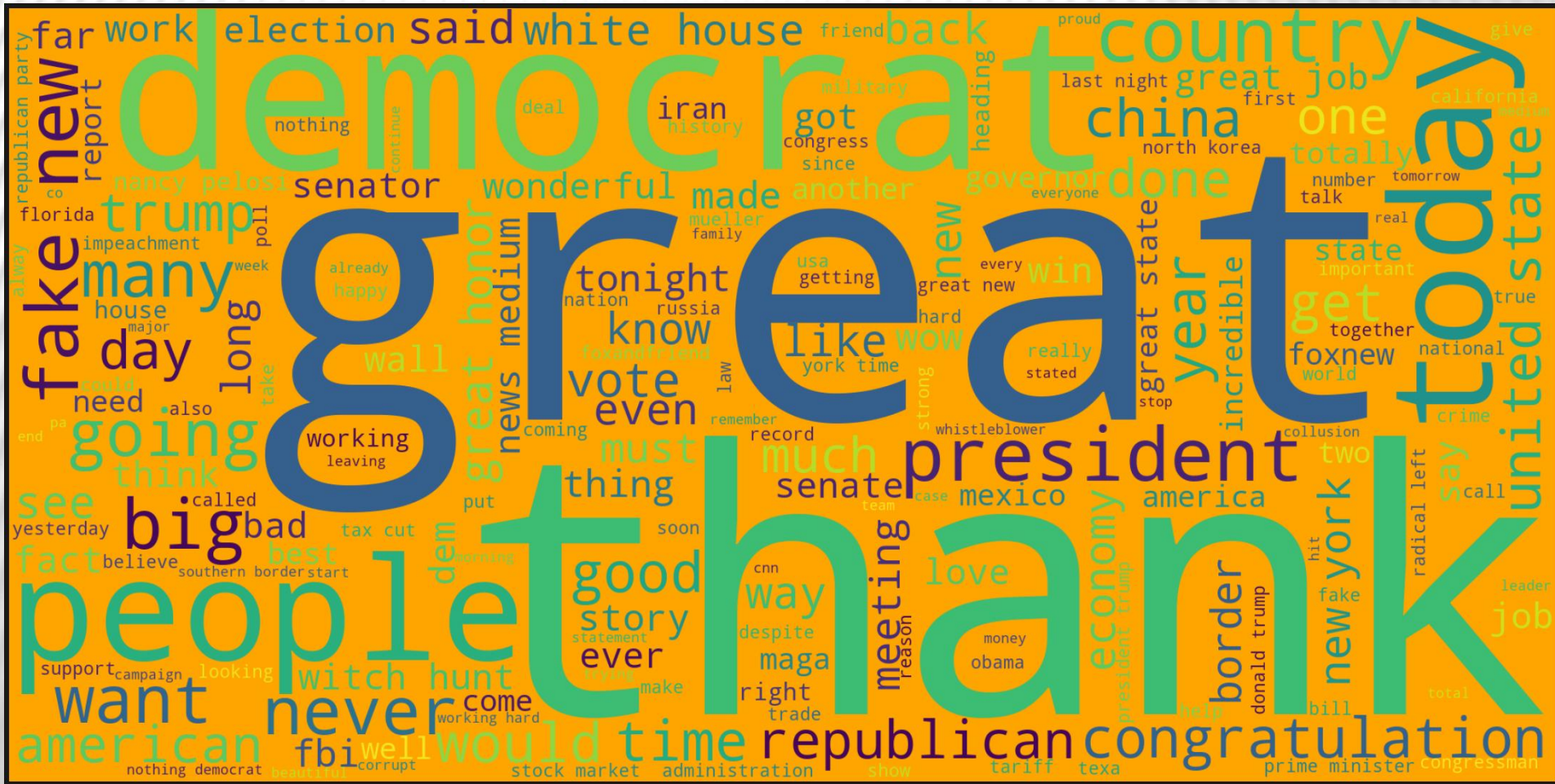
@REALDONALDTRUMP TWEETS ANALYSIS

- Trump’s tweets per year

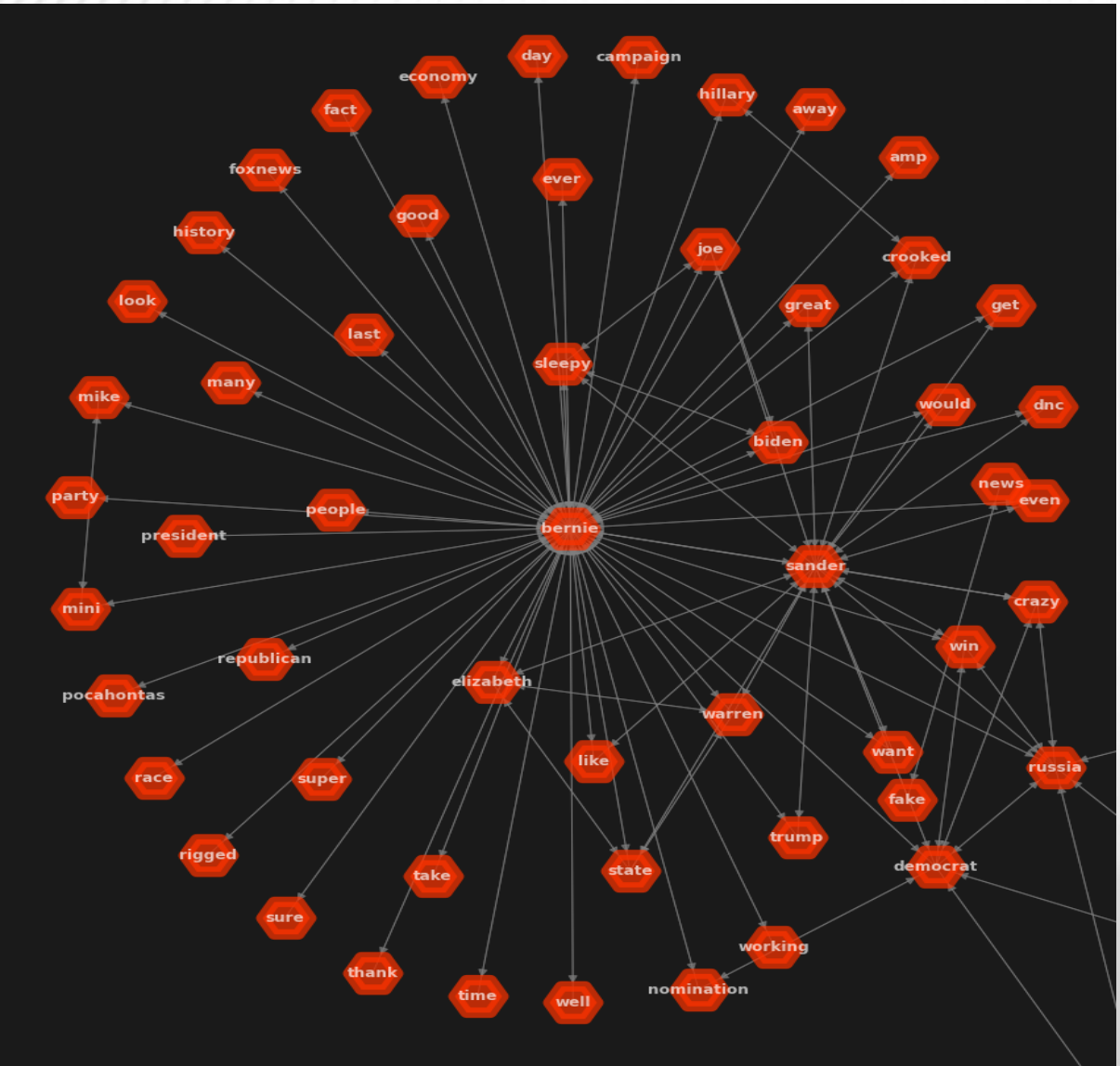


@REALDONALDTRUMP TWEETS ANALYSIS

- Trump’s wordcloud



TRUMP TWEETS ABOUT BERNIE...



Cluster 0:
crazy bernie
crazy
joe
look
sleepy joe
sleepy
going
look like
great
like

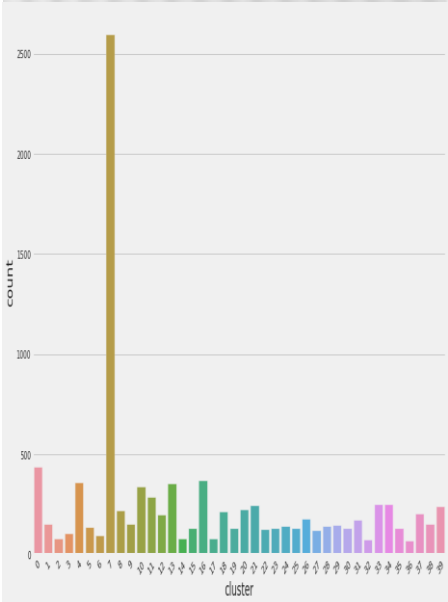
Cluster 1:
elizabeth
warren
elizabeth warren
race
even
state
would
news
fake
pocahontas

Cluster 2:
russia
crooked
democrat
russia russia
crooked hillary
hillary
right
nomination
rigged
dnc

@BERNIESANDERS TWEETS ANALYSIS

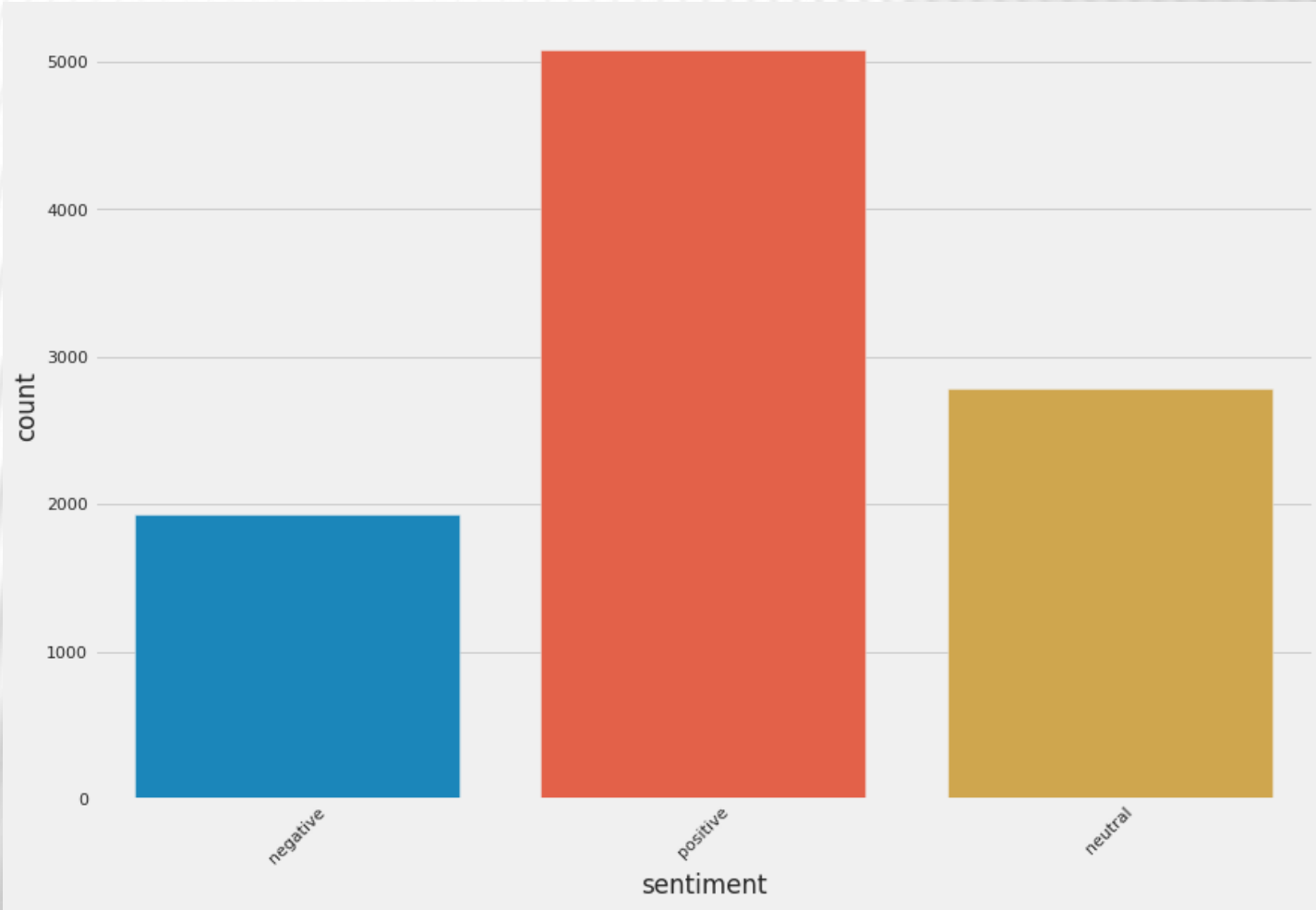
- KMEANS CLUSTERING (k=40 clusters)

Cluster 9: billionaire billionaire class class message billionaire message take country movement greed class cannot	**Cluster 13:** health care health care medicare system care system insurance health insurance need need medicare	**Cluster 17:** gun weapon assault weapon assault gun violence nra violence gun safety safety must	**Cluster 22:** wage minimum wage minimum 15 hour 15 hour wage 15 raise federal minimum living wage
Cluster 25: stand together together stand accomplish nothing cannot accomplish nothing cannot win together nothing cannot	**Cluster 33:** climate change climate change united state united state state america planet change real president	**Cluster 35:** woman right abortion woman right body control control body right control equal constitutional right	**Cluster 38**: young people young people political process process country political transform future people people



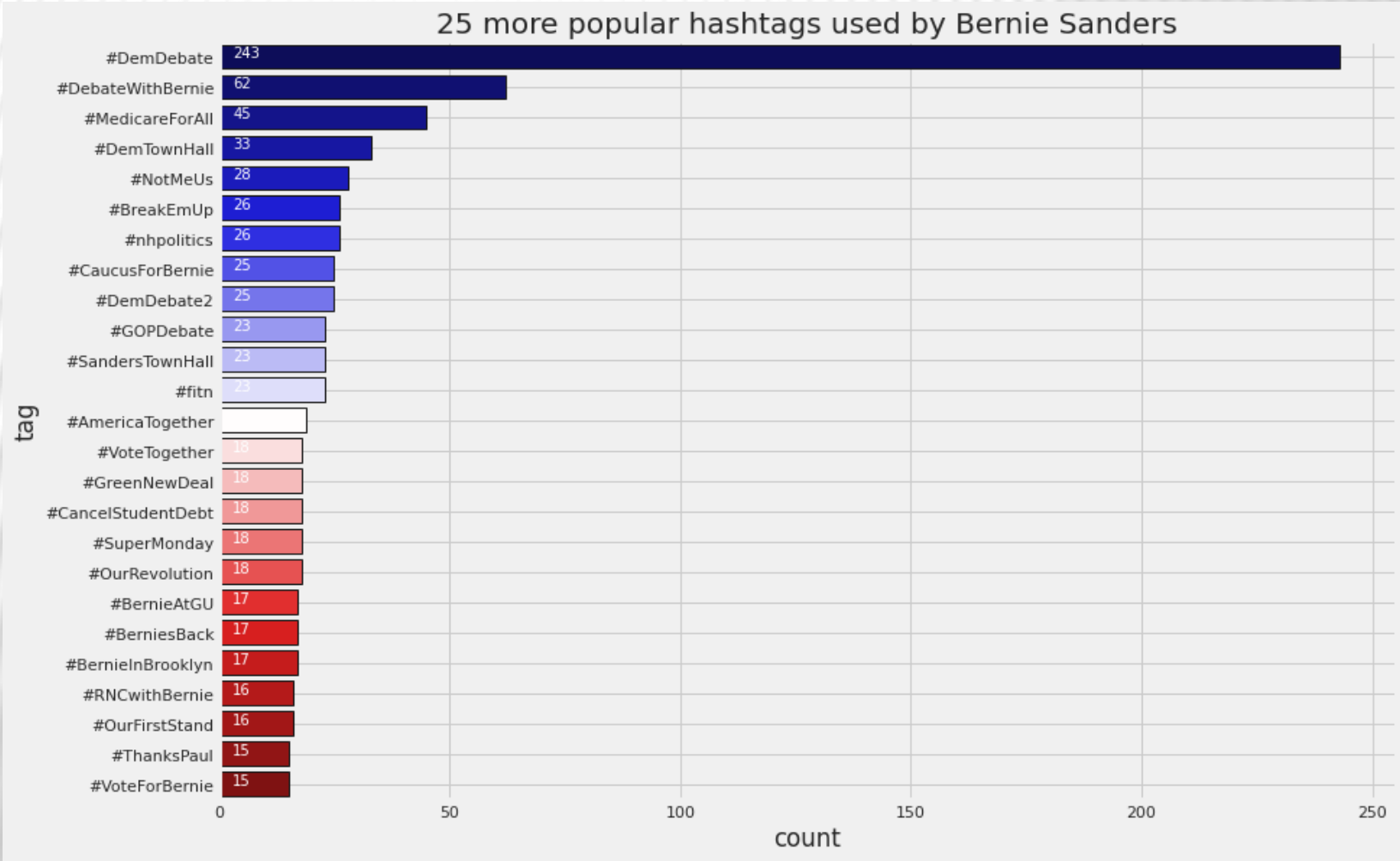
@BERNIESANDERS TWEETS ANALYSIS

- Sentiment analysis (TextBlob)



@BERNIESANDERS TWEETS ANALYSIS

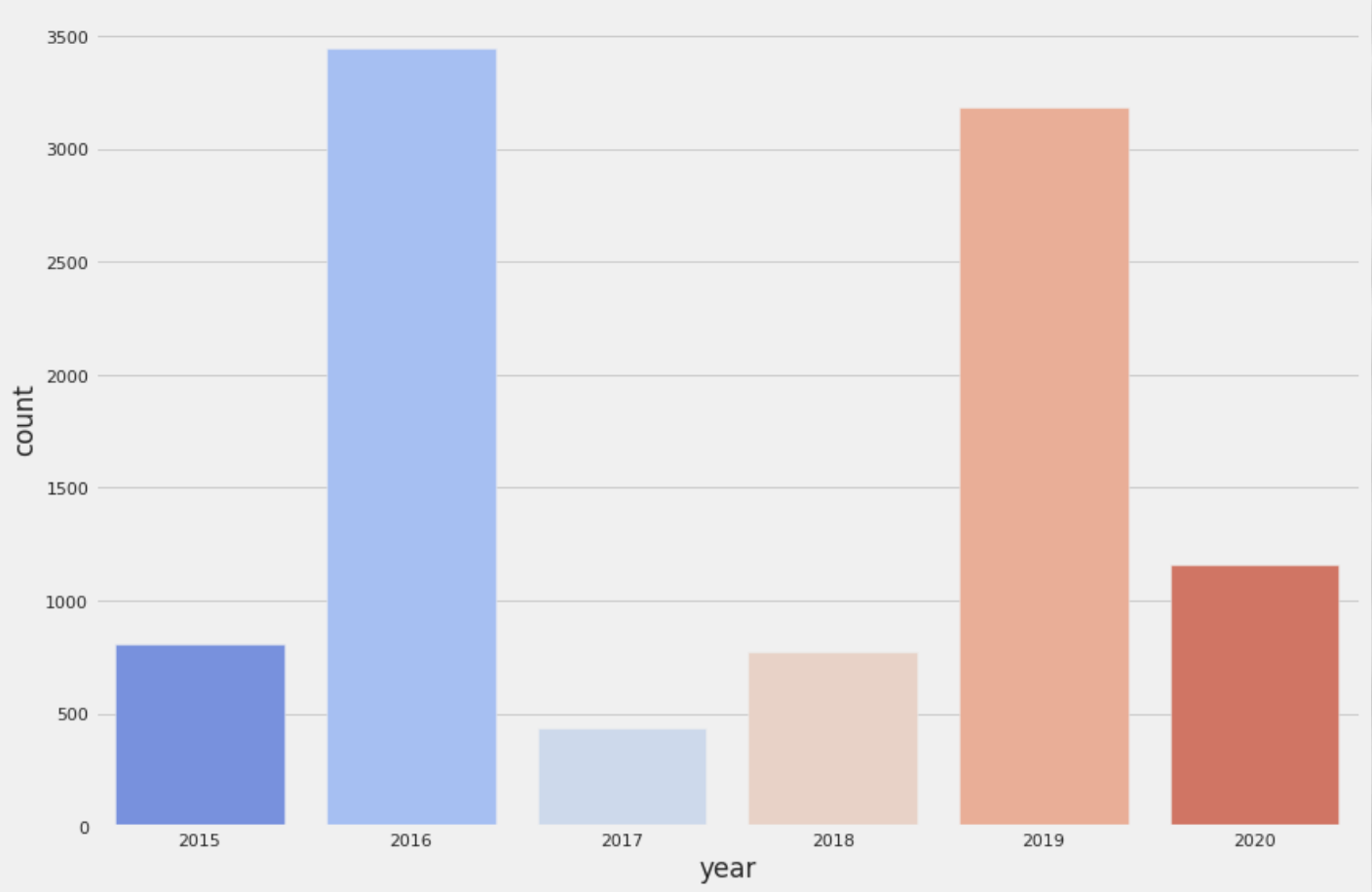
- 25 most popular hashtags used by Sanders



NETWORK CONNECTION GRAPH (FREQ>80 WORDS)

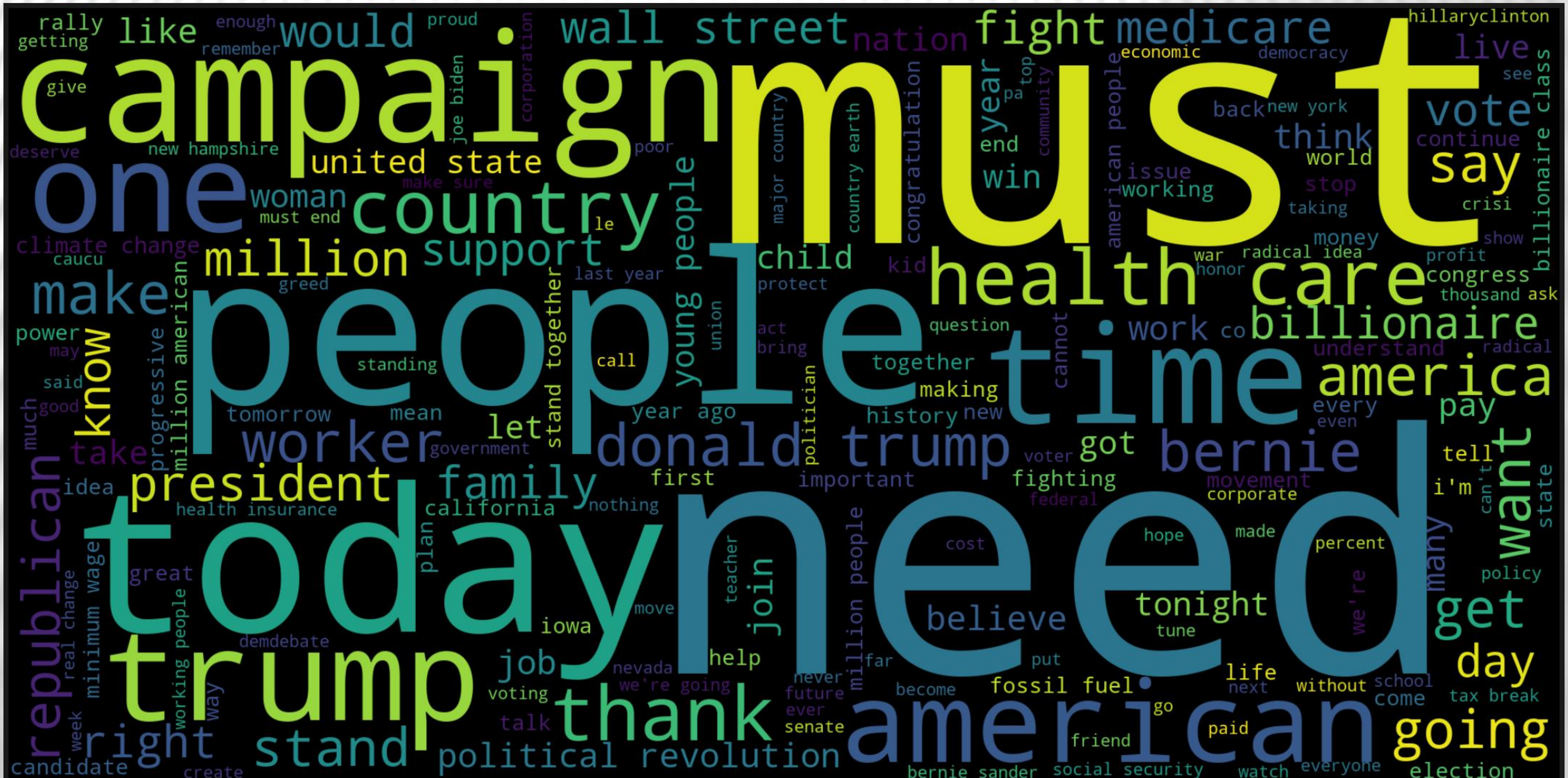
@BERNIESANDERS TWEETS ANALYSIS

- Sanders' tweets per year

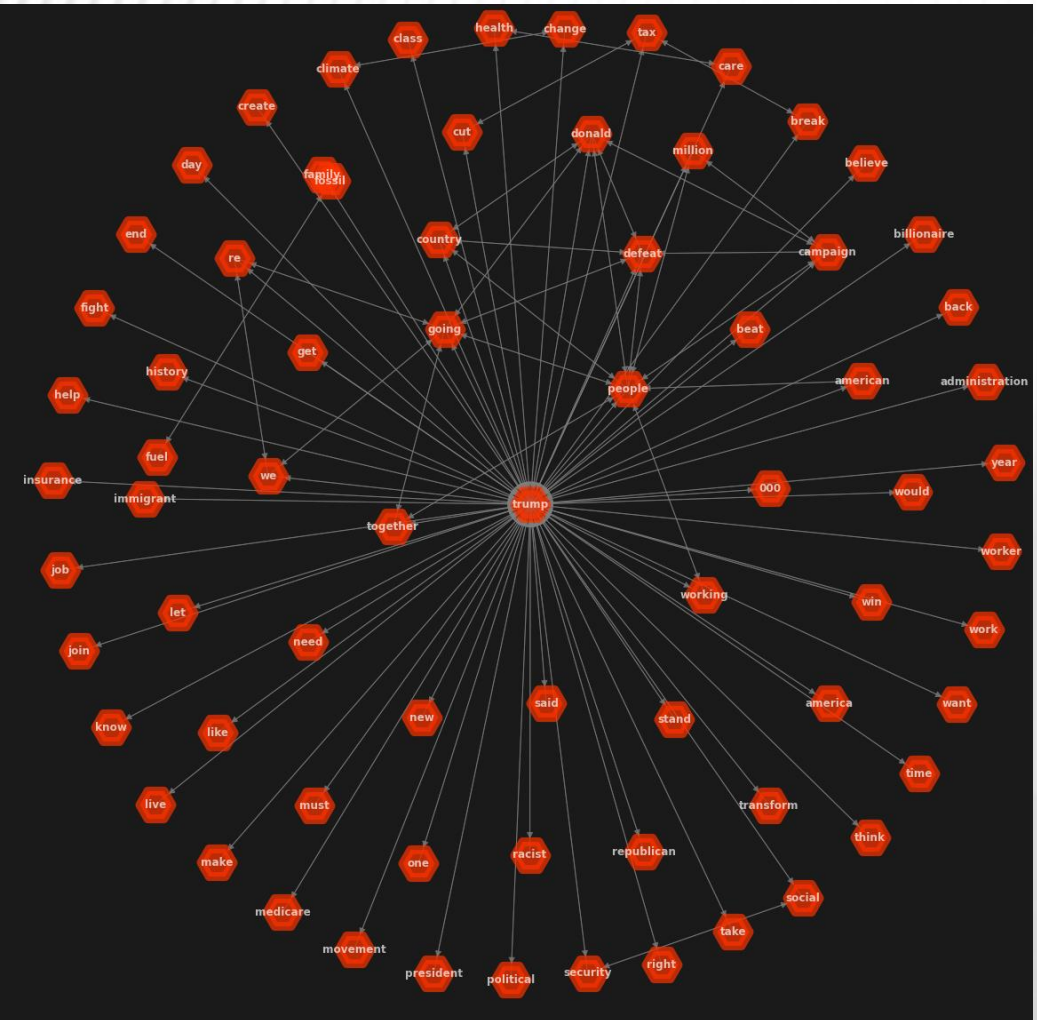


@BERNIESANDERS TWEETS ANALYSIS

- Sander's wordcloud



BERNIE TWEETS ABOUT TRUMP...



Cluster 4:
defeat trump
defeat
transform
people
transform country
trump transform
country
trump transform country
campaign
political process

Cluster 0:
friend
trump friend
fraud
make
liar
pathological
voter
pathological liar
trump pathological
donald

Cluster 9:
beat
beat trump
people
win
campaign
election
people together
donald
donald trump
think

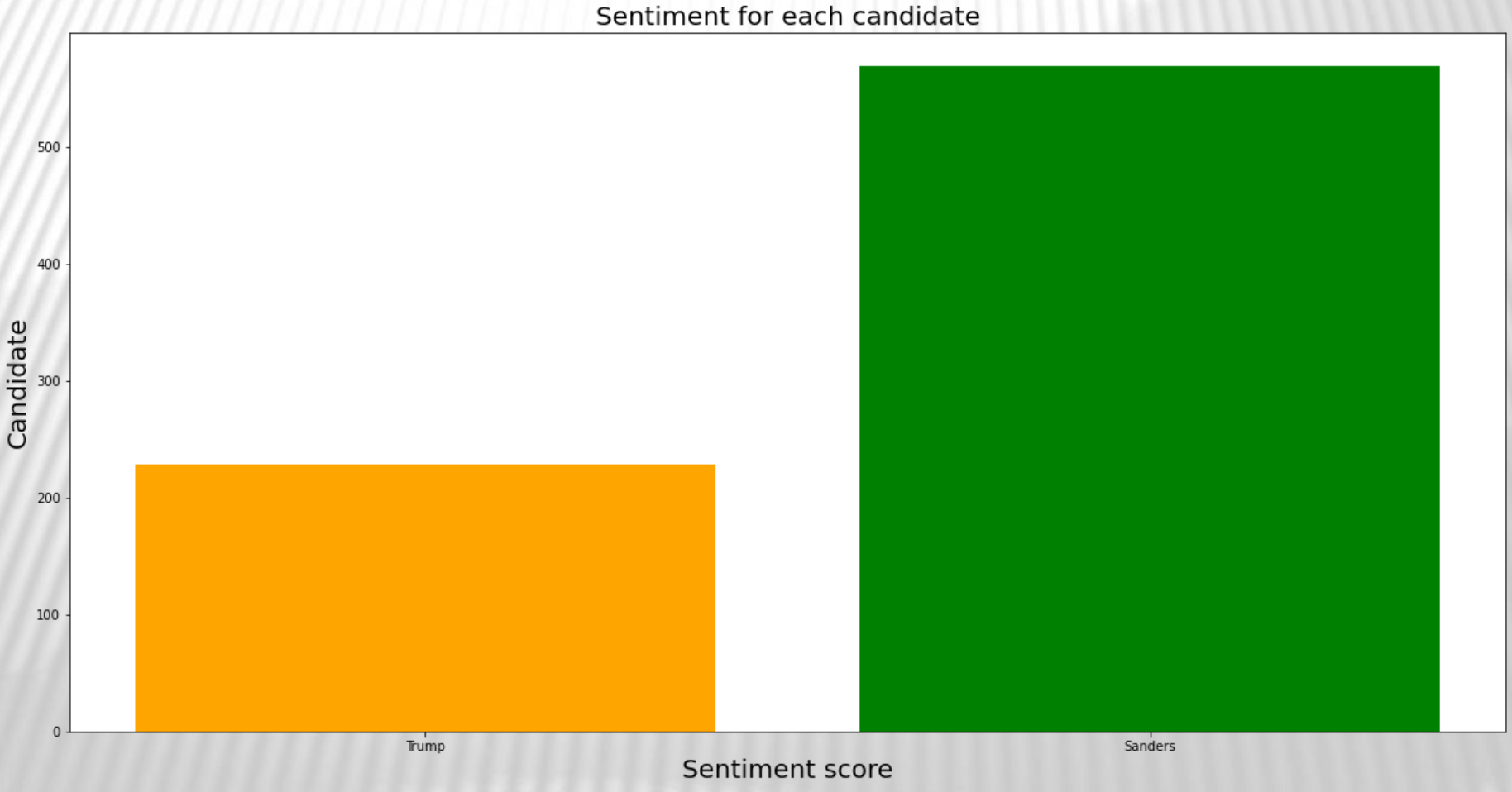
Cluster 11:
racism
xenophobia
sexism
let stop
racism sexism
homophobia
let
stop
racism sexism
sexism xenophobia

ΔΗΜΟΦΙΛΙΑ ΥΠΟΨΗΦΙΩΝ

- Sentiment Analysis σε tweets χρηστών απο τις ΗΠΑ (5.000 tweets σε καθημερινή βάση)
- Sentiment Analysis στα tweets του CNN politics που αφορούν τους υποψήφιους
- Sentiment Analysis σε 1.500 άρθρα από διάφορες πηγές (Nytimes, Aljazzera, BBC, RT)

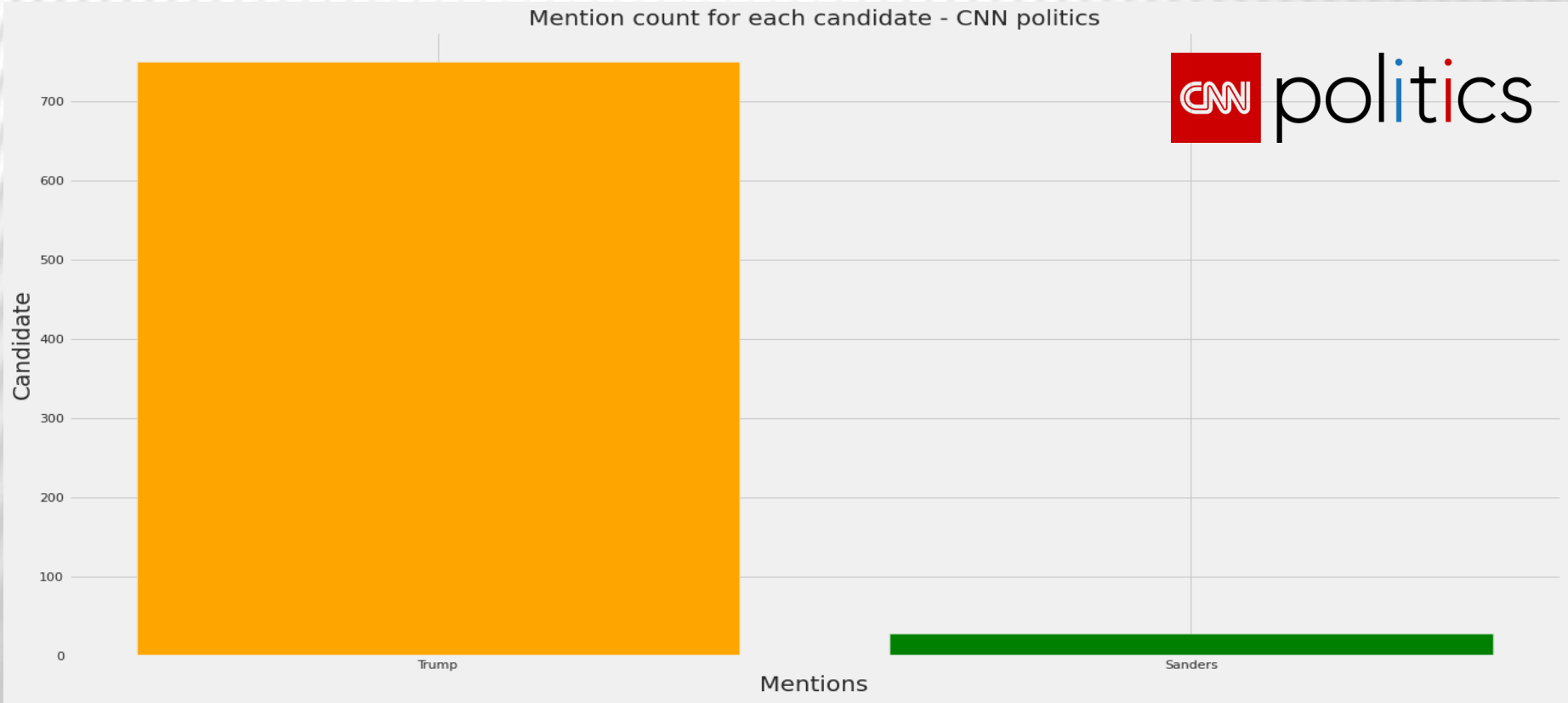
TWITTER USERS VOTE FOR…

- Δείγμα: 25.000 tweets χρηστών



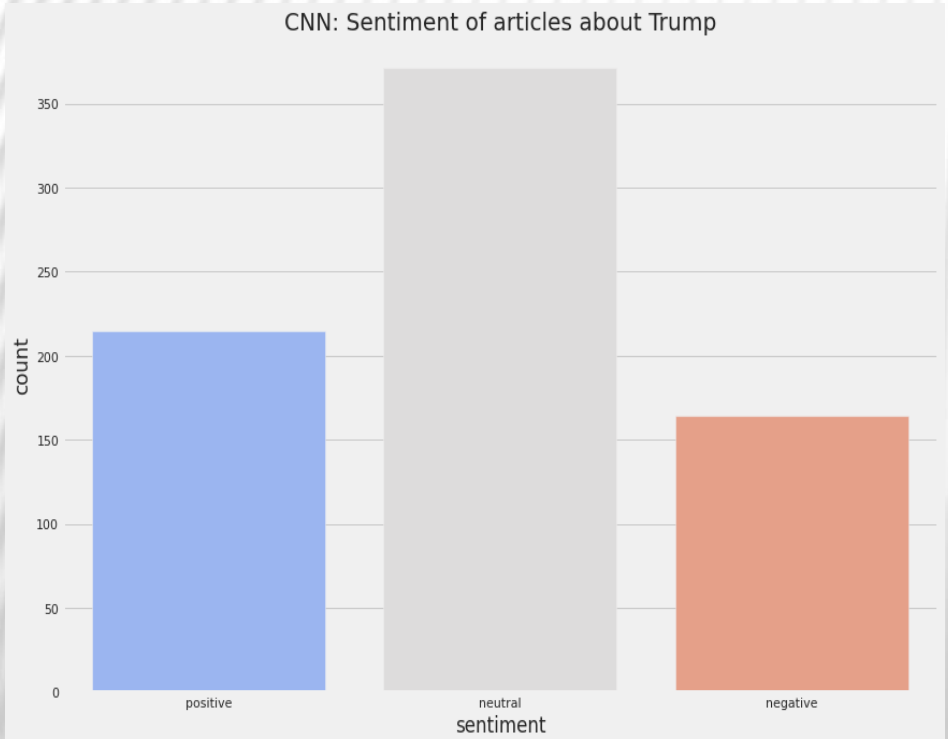
WHAT ABOUT THE MEDIA?

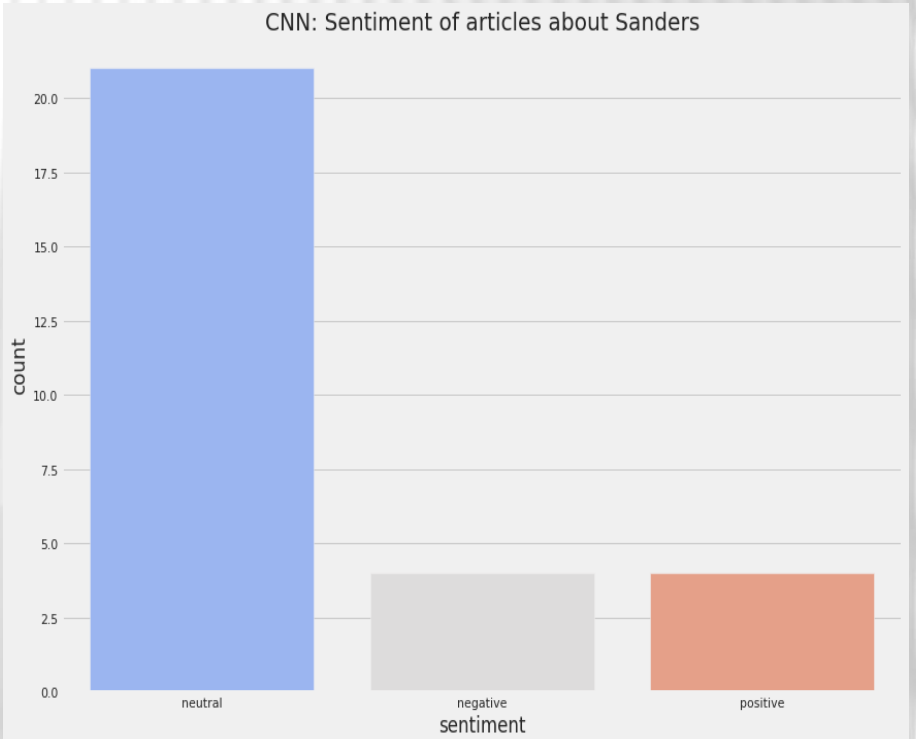
- Ο Trump αναφέρεται μακράν περισσότερο.
- Διάστημα: (20/3 – 16/4)



WHAT ABOUT THE MEDIA?

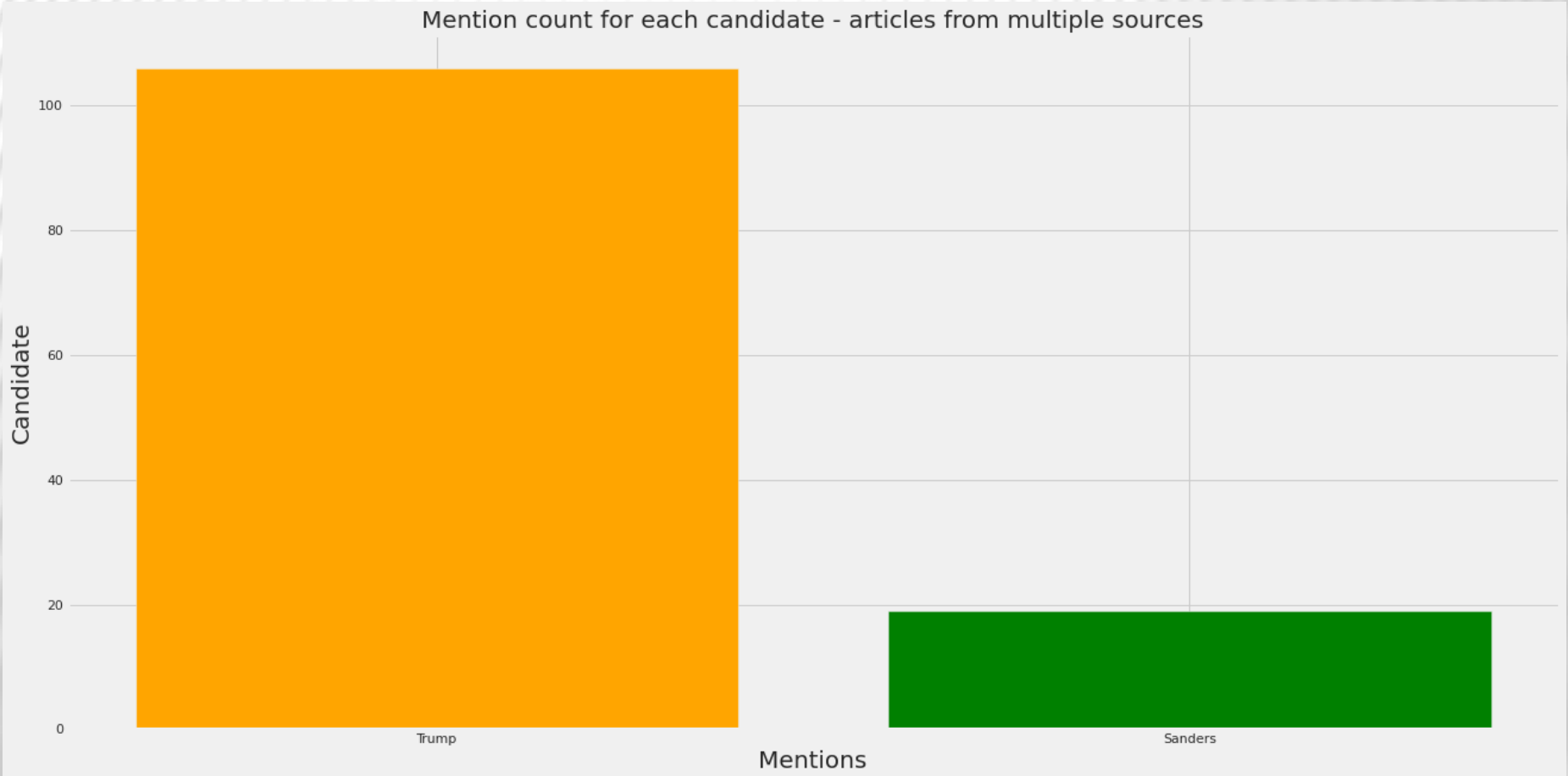
- Sentiment Analysis





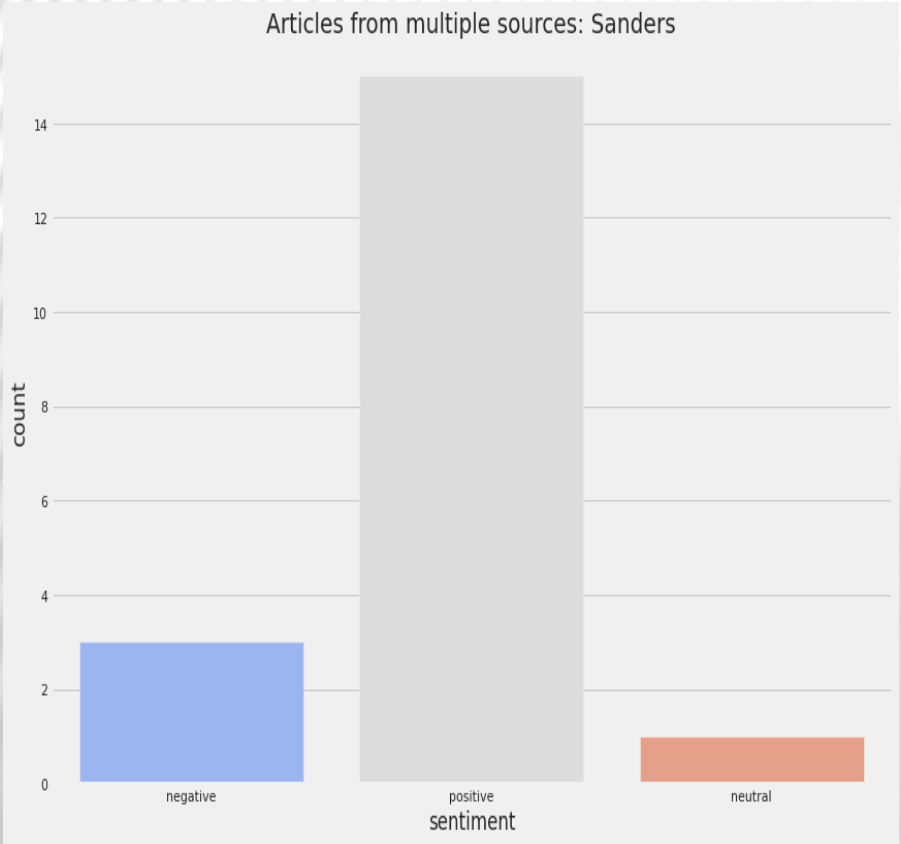
ARTICLES FROM VARIOUS SOURCES

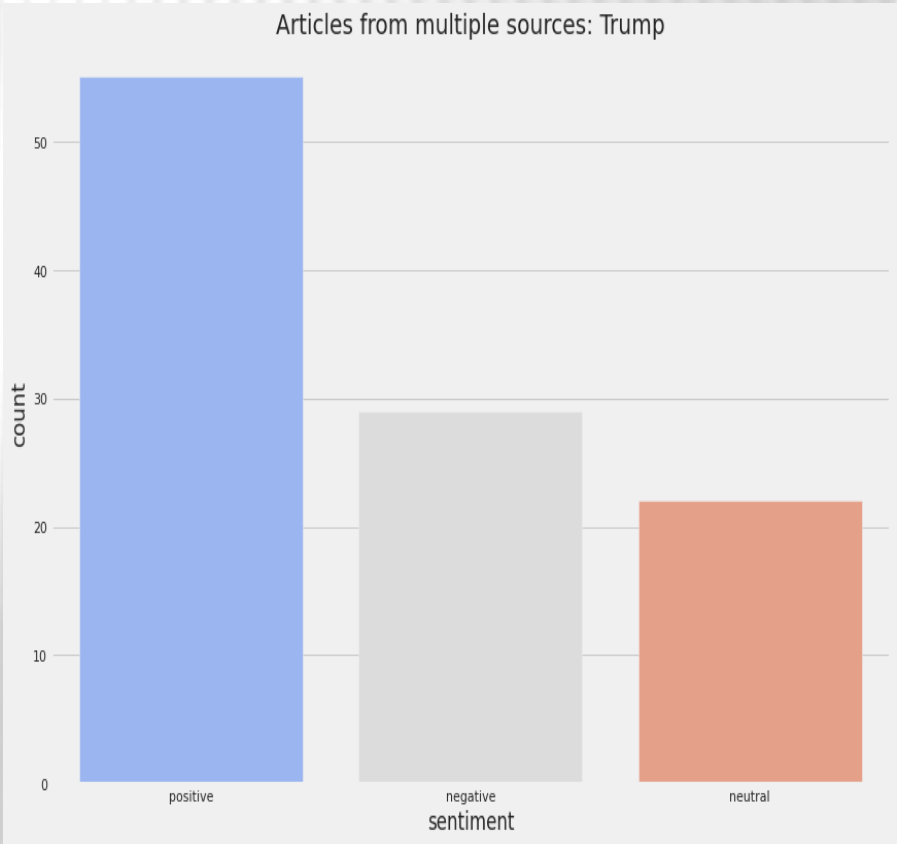
- 11/3 – 16/4 (CNN, AlJazeera, RT, BBC etc)



ARTICLES FROM VARIOUS SOURCES

- Sentiment Analysis





ARTICLES FROM VARIOUS SOURCES

- Clustering

Cluster 0:
sanders
vote
super
super tuesday
youth
tuesday
sanders campaign
campaign
election
youth vote

Cluster 1:
coronavirus
rally
campaign
sanders
biden bernie
biden
sander
election
outbreak
coronavirus outbreak

ΣΥΜΠΕΡΑΣΜΑΤΑ

- Ο Trump αναφέρεται πολύ περισσότερο στα μέσα από τον Sanders (any publicity is good publicity?)
- Ο Sanders twit-άρει πολύ πιο πολιτικά από τον Trump. Από το clustering των tweets μπορούμε να εξάγουμε κάποια συμπεράσματα για το πολιτικό του πρόγραμμα.

TRUMP OR BERNIE?

- Δημιουργία classifier
- Ένωση των 2 dataset σε ένα
- Training set: Tfidf Vectorization -> Data term Matrix
- Classifer: Multinomial Naive Bayes

- Accuracy: 0.951