CS525
Information Retrieval & Social Web
Homework 4

Submitted by: Kratika Agrawal

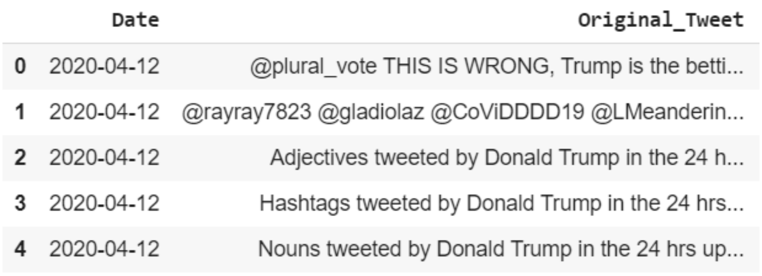
Task 1: Collecting Twitter Data

- Created Twitter Account
- Created App in developer.twitter.com
- Requested API access for tweet analysis to complete the assignment
- Got API keys in a day.
- Import twitter in notebook
- Set up a connection to the API passing Consumer API key, Consumer API key Secret, Access token, Access token secret.
- Using the authenticated API, hit Twitter Search API passing query.
    - Query for Donald Trump tweets: q=’donald trump’, lang=’en’, exclude=’retweets’, count=’1000’
    - Query for Joe Biden tweets: q=’joe biden’, lang=’en’, exclude=’retweets’, count=’1000’
- I hit these APIs 7 times for both candidates, as we are only allowed to fetch last 7 days tweets using the Standard API.
- I passed the parameter until with a specific date tweets and collected as many as 700 tweets for both candidates.

Task 2: Exploratory Analysis

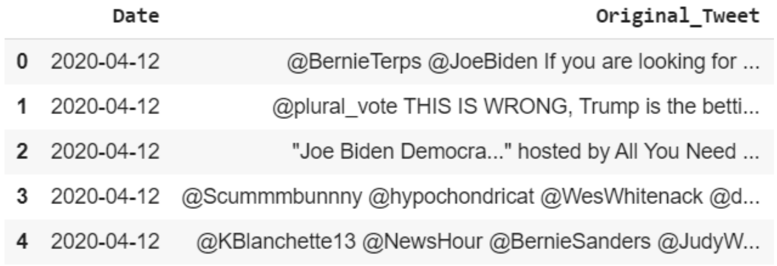
- Fetched tweets and created date(converted string formatted created datetime to absolute date value) for it and put it in a data frame for both candidates.

Donald Trump top 5 tweets:



	Date	Original_Tweet
0	2020-04-12	@plural_vote THIS IS WRONG, Trump is the betti...
1	2020-04-12	@rayray7823 @gladiolaz @CoViDDDD19 @LMeanderin...
2	2020-04-12	Adjectives tweeted by Donald Trump in the 24 h...
3	2020-04-12	Hashtags tweeted by Donald Trump in the 24 hrs...
4	2020-04-12	Nouns tweeted by Donald Trump in the 24 hrs up...

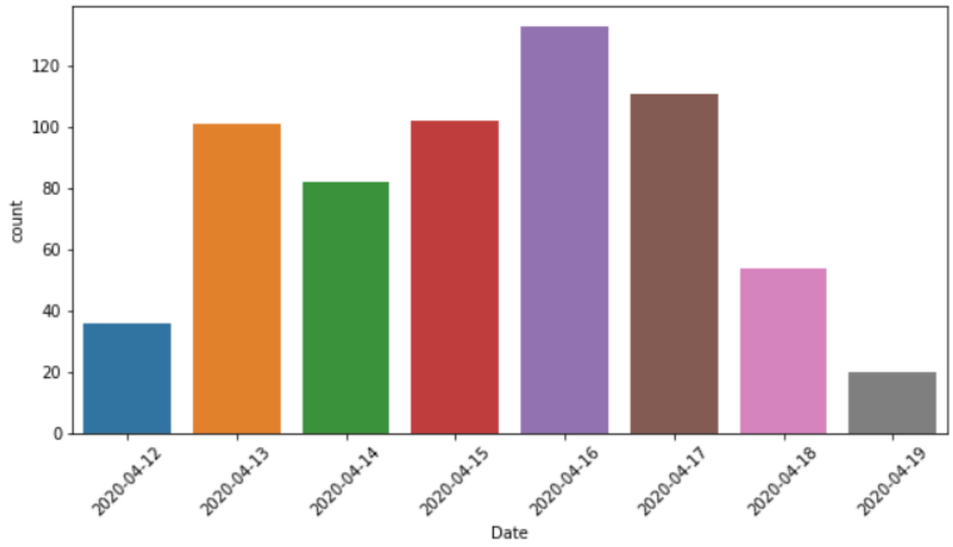
Joe Biden top 5 tweets:



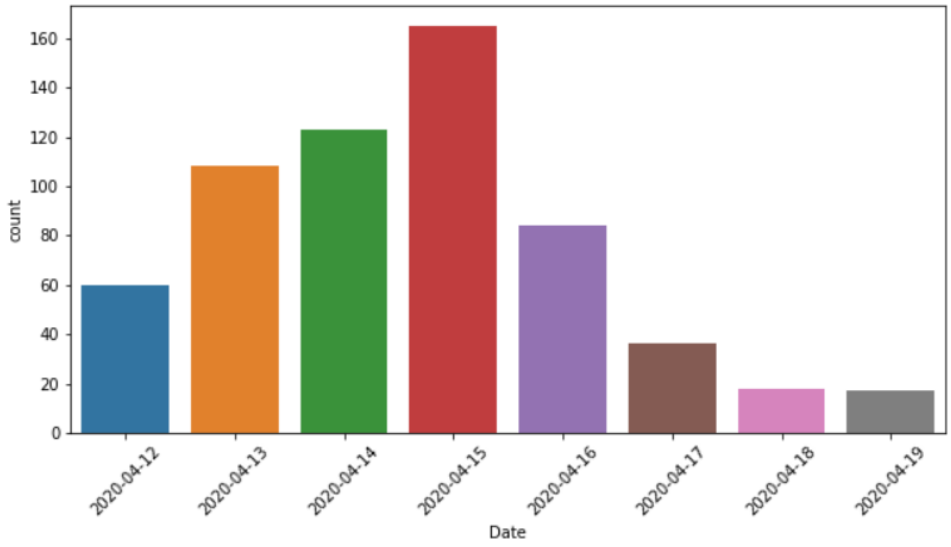
	Date	Original_Tweet
0	2020-04-12	@BernieTerps @JoeBiden If you are looking for ...
1	2020-04-12	@plural_vote THIS IS WRONG, Trump is the betti...
2	2020-04-12	"Joe Biden Democra..." hosted by All You Need ...
3	2020-04-12	@Scummmbunnny @hypochondricat @WesWhitenack @d...
4	2020-04-12	@KBlanchette13 @NewsHour @BernieSanders @JudyW...

- ➢ Number of tweets on each day for both candidates:

Donald Trump tweet frequency:



Joe Biden tweet frequency:



- ➢ We can see from these plots that number of tweets on each day in last week have been more for Donald Trump than Joe Biden except for 04/16/2020 where number of tweets for Joe Biden are more.

Text Pre-processing

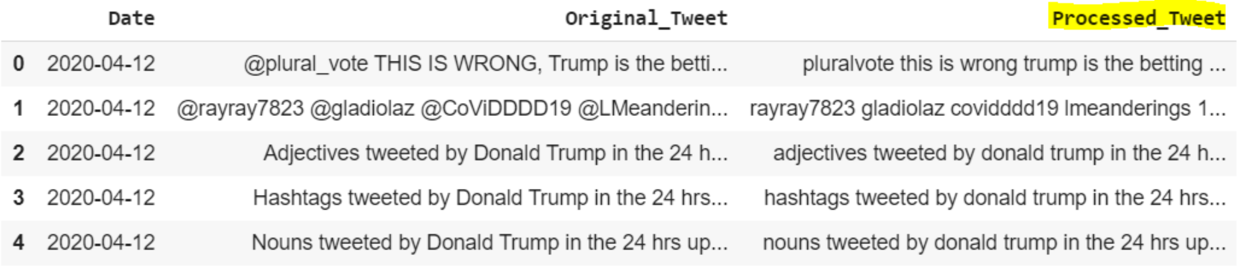
- ➢ I, then, performed topic modelling on the data.

For perform LDA, I had to pre-process the data:

- Changing tweets to lower case
- Removing punctuations
- Removing numerical values
- Removing leading or trailing white spaces
- Tokenizing tweets using NLTK library
- Removing Stop words using NLTK library
- Lemmatizing tweets using NLTK library

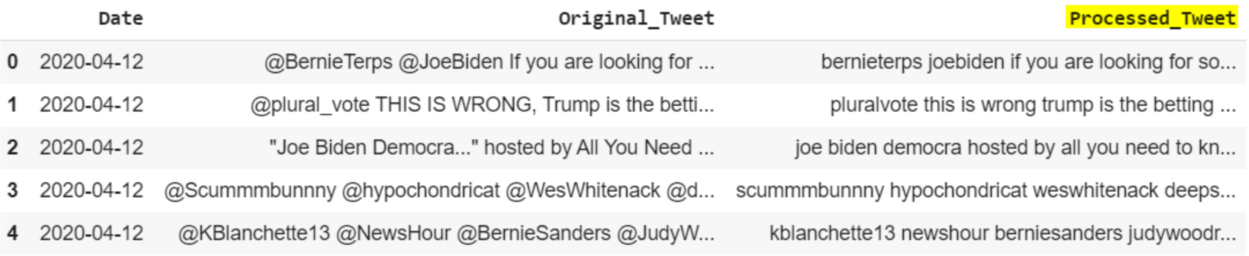
Updated processed tweets in a column on the Donald Trump dataframe and Joe Biden dataframe.

Donald Trump DataFrame



	Date	Original_Tweet	Processed_Tweet
0	2020-04-12	@plural_vote THIS IS WRONG, Trump is the betti...	pluralvote this is wrong trump is the betting ...
1	2020-04-12	@rayray7823 @gladiolaz @CoViDDDD19 @LMeanderin...	rayray7823 gladiolaz covidddd19 lmeanderings 1...
2	2020-04-12	Adjectives tweeted by Donald Trump in the 24 h...	adjectives tweeted by donald trump in the 24 h...
3	2020-04-12	Hashtags tweeted by Donald Trump in the 24 hrs...	hashtags tweeted by donald trump in the 24 hrs...
4	2020-04-12	Nouns tweeted by Donald Trump in the 24 hrs up...	nouns tweeted by donald trump in the 24 hrs up...

Joe Biden DataFrame

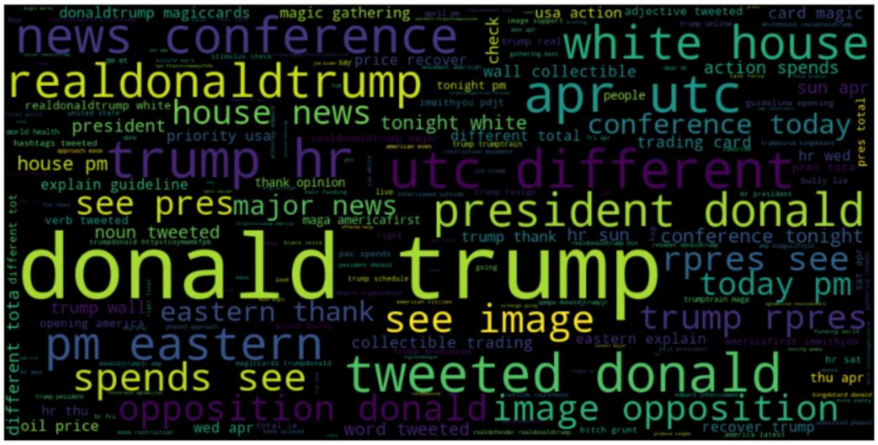


	Date	Original_Tweet	Processed_Tweet
0	2020-04-12	@BernieTerps @JoeBiden If you are looking for ...	bernieterps joebiden if you are looking for so...
1	2020-04-12	@plural_vote THIS IS WRONG, Trump is the betti...	pluralvote this is wrong trump is the betting ...
2	2020-04-12	"Joe Biden Democra..." hosted by All You Need ...	joe biden democra hosted by all you need to kn...
3	2020-04-12	@Scummmbunnny @hypochondricat @WesWhitenack @d...	scummmbunnny hypochondricat weswhitenack deeps...
4	2020-04-12	@KBlanchette13 @NewsHour @BernieSanders @JudyW...	kblanchette13 newshour berniesanders judywoodr...

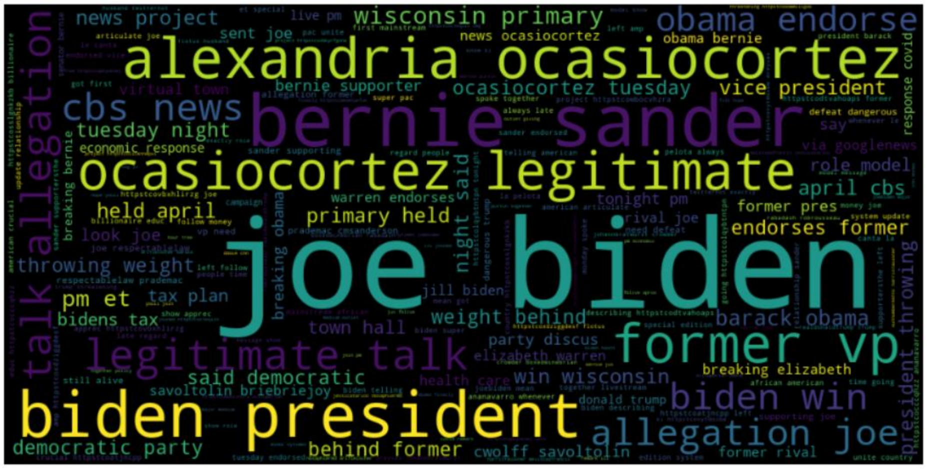
Topic Modelling (LDA):

- ➢ On performing LDA using 'gensim' library, extracted top 3 mostly talked about topics in case of Donald Trump and Joe Biden.
- ➢ I even displayed wordCloud for both of the candidates' tweets:

Donald Trump tweets Word Cloud



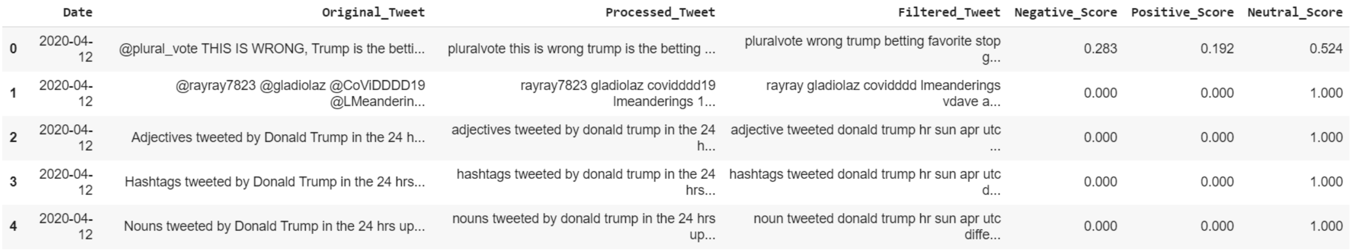
Joe Biden tweets Word Cloud



Sentiment Analysis (VaderSentiment)

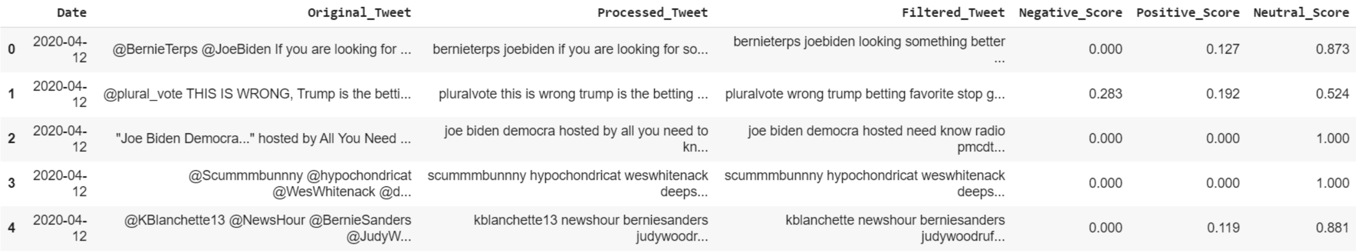
- Last step is to perform Rule based Sentiment Analysis (vaderSentiment) to determine how negative, positive and neutral a tweet is and updated the data table for both candidates.

Donald Trump Tweets



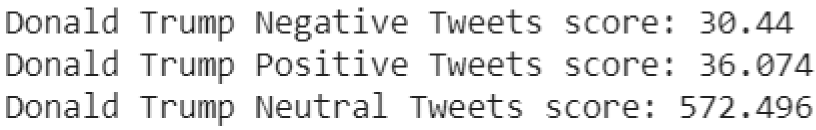
	Date	Original_Tweet	Processed_Tweet	Filtered_Tweet	Negative_Score	Positive_Score	Neutral_Score
0	2020-04-12	@plural_vote THIS IS WRONG, Trump is the betti...	pluralvote this is wrong trump is the betting ...	pluralvote wrong trump betting favorite stop g...	0.283	0.192	0.524
1	2020-04-12	@rayray7823 @gladiolaz @CoViDDDD19 @LMeanderin...	rayray7823 gladiolaz covidddd19 lmeanderings 1...	rayray gladiolaz covidddd lmeanderings vdave a...	0.000	0.000	1.000
2	2020-04-12	Adjectives tweeted by Donald Trump in the 24 h...	adjectives tweeted by donald trump in the 24 h...	adjective tweeted donald trump hr sun apr utc ...	0.000	0.000	1.000
3	2020-04-12	Hashtags tweeted by Donald Trump in the 24 hrs...	hashtags tweeted by donald trump in the 24 hrs...	hashtags tweeted donald trump hr sun apr utc d...	0.000	0.000	1.000
4	2020-04-12	Nouns tweeted by Donald Trump in the 24 hrs up...	nouns tweeted by donald trump in the 24 hrs up...	noun tweeted donald trump hr sun apr utc diffe...	0.000	0.000	1.000

Joe Biden Tweets

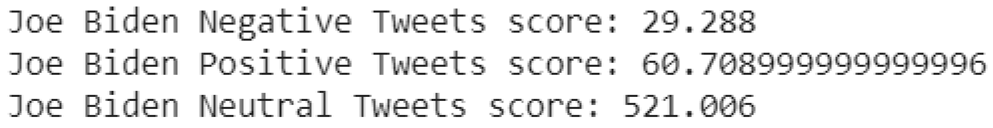


	Date	Original_Tweet	Processed_Tweet	Filtered_Tweet	Negative_Score	Positive_Score	Neutral_Score
0	2020-04-12	@BernieTerps @JoeBiden If you are looking for ...	bernieterps joebiden if you are looking for so...	bernieterps joebiden looking something better ...	0.000	0.127	0.873
1	2020-04-12	@plural_vote THIS IS WRONG, Trump is the betti...	pluralvote this is wrong trump is the betting ...	pluralvote wrong trump betting favorite stop g...	0.283	0.192	0.524
2	2020-04-12	"Joe Biden Democra..." hosted by All You Need ...	joe biden democra hosted by all you need to kn...	joe biden democra hosted need know radio pmcdt...	0.000	0.000	1.000
3	2020-04-12	@Scummmbunnny @hypochondricat @WesWhitenack @d...	scummmbunnny hypochondricat weswhitenack deeps...	scummmbunnny hypochondricat weswhitenack deeps...	0.000	0.000	1.000
4	2020-04-12	@KBlanchette13 @NewsHour @BernieSanders @JudyW...	kblanchette13 newshour berniesanders judywoodr...	kblanchette newshour berniesanders judywoodruf...	0.000	0.119	0.881

- Now, for each candidate, I evaluated total number of Negative Score, Positive Score and Neutral Score. Total comes out to be:



```
Donald Trump Negative Tweets score: 30.44
Donald Trump Positive Tweets score: 36.074
Donald Trump Neutral Tweets score: 572.496
```



```
Joe Biden Negative Tweets score: 29.288
Joe Biden Positive Tweets score: 60.708999999999996
Joe Biden Neutral Tweets score: 521.006
```

- Seeing the total score for positive sentiment for Joe Biden is more than that of Donald Trump.

  Thus, we can conclude that based on the tweets collected and the analysis done here, there are more chances of Joe Biden winning the election.