HomeWork #4

Donald Trump vs. Joe Biden

Task 1: Collecting Twitter Data :

Tweets are collected for Donald Trump and Joe Biden using the twitter API. For this Keys and Tokens are generated from the Twitter Developer Account. A generic function to search Tweets using a given Twitter API and search key is constructed to search for a maximum of 10000 tweets.

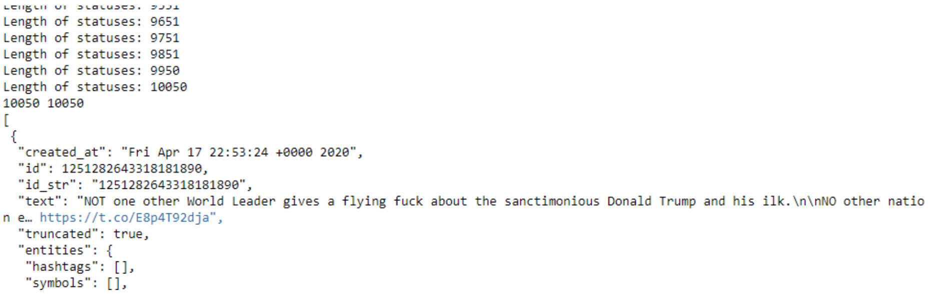
There are 2 methods of searching. We could search for a period of 7 days, by not using the "unquote" command while creating the dictionary of tweets. But the disadvantage is that we would receive a maximum of 100 tweets only, for this given period. As our initial requirement is to acquire atleast 1000 tweets, I have used the "unquote" command for the dictionary for 10000 tweets, which gives me tweets for a single day[here, April 17th 2020]. This generic search is then used to find the respective tweets for Donald Trump and Joe Biden. I have used the following filters for the tweets:

Filtering tweets that are only "English" and also filtering out the tweets that are retweets, replies and that contain the string "RT @". I have also included tweets that come from the respective candidates.

Steps:
1. Generate the Twitter API for the login
2. Create a Search String for Donald Trump /Joe Biden
3. Get 10000 Tweets about the topic and log to json file
4. Print the generated json file

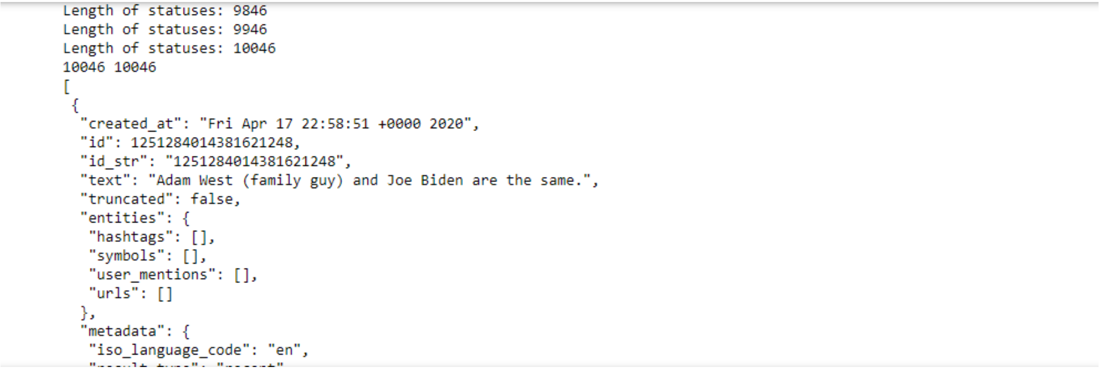
Please find a sample of the tweets obtained:

Donald Trump Tweet:



```
Length of statuses: 9551
Length of statuses: 9651
Length of statuses: 9751
Length of statuses: 9851
Length of statuses: 9950
Length of statuses: 10050
10050 10050
[
 {
  "created_at": "Fri Apr 17 22:53:24 +0000 2020",
  "id": 1251282643318181890,
  "id_str": "1251282643318181890",
  "text": "NOT one other World Leader gives a flying fuck about the sanctimonious Donald Trump and his ilk.\n\nNO other natio
n e… https://t.co/E8p4T92dja",
  "truncated": true,
  "entities": {
   "hashtags": [],
   "symbols": [],
```

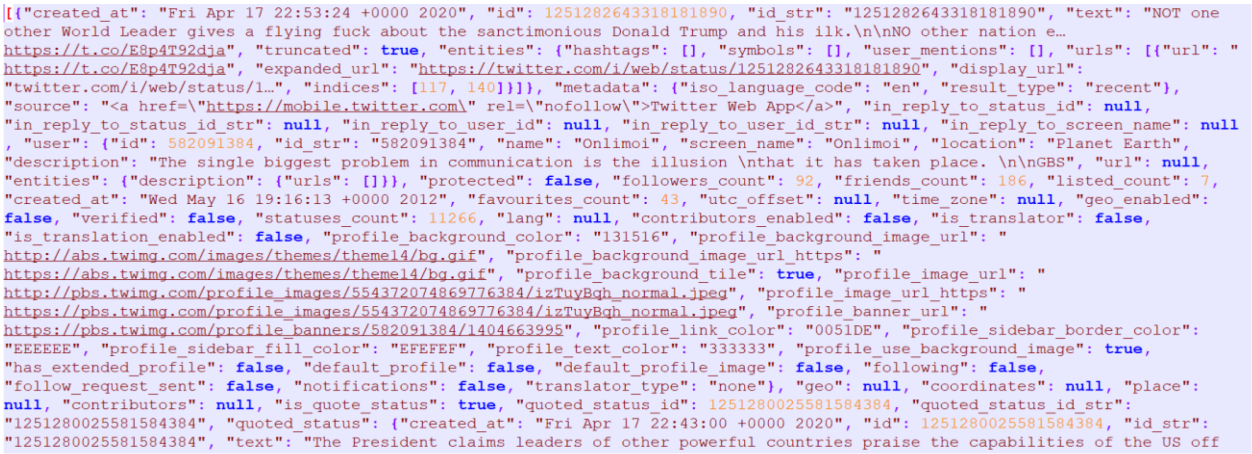
Joe Biden Tweet:



```
Length of statuses: 9846
Length of statuses: 9946
Length of statuses: 10046
10046 10046
[
 {
  "created_at": "Fri Apr 17 22:58:51 +0000 2020",
  "id": 1251284014381621248,
  "id_str": "1251284014381621248",
  "text": "Adam West (family guy) and Joe Biden are the same.",
  "truncated": false,
  "entities": {
   "hashtags": [],
   "symbols": [],
   "user_mentions": [],
   "urls": []
  },
  "metadata": {
   "iso_language_code": "en",
```

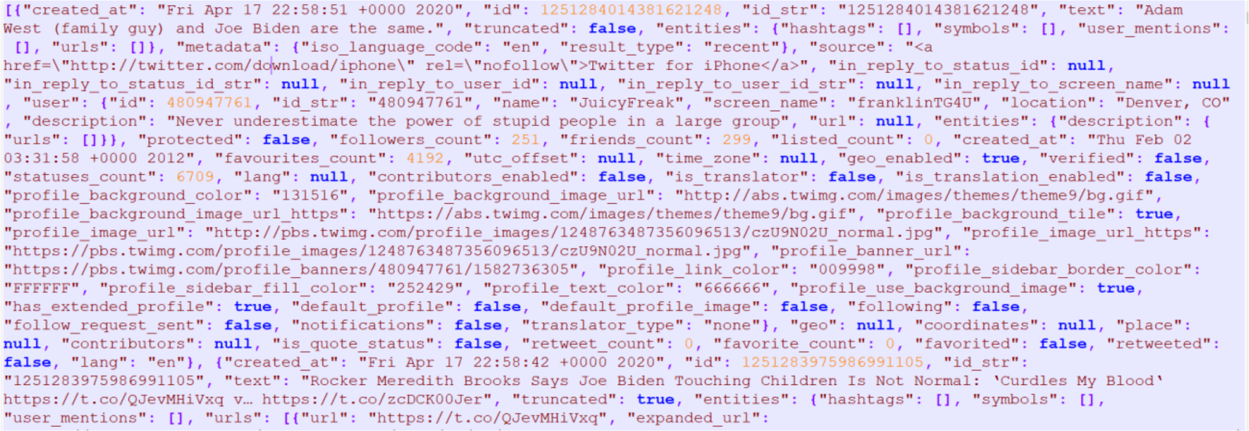
These tweets are then saved to a file in JSON format.

Donald_Trump.json:



```
[{"created_at": "Fri Apr 17 22:53:24 +0000 2020", "id": 1251282643318181890, "id_str": "1251282643318181890", "text": "NOT one
other World Leader gives a flying fuck about the sanctimonious Donald Trump and his ilk.\n\nNO other nation e…
https://t.co/E8p4T92dja", "truncated": true, "entities": {"hashtags": [], "symbols": [], "user_mentions": [], "urls": [{"url": "
https://t.co/E8p4T92dja", "expanded_url": "https://twitter.com/i/web/status/1251282643318181890", "display_url":
"twitter.com/i/web/status/1…", "indices": [117, 140]}]}, "metadata": {"iso_language_code": "en", "result_type": "recent"},
"source": "<a href=\"https://mobile.twitter.com\" rel=\"nofollow\">Twitter Web App</a>", "in_reply_to_status_id": null,
"in_reply_to_status_id_str": null, "in_reply_to_user_id": null, "in_reply_to_user_id_str": null, "in_reply_to_screen_name": null
, "user": {"id": 582091384, "id_str": "582091384", "name": "Onlimoi", "screen_name": "Onlimoi", "location": "Planet Earth",
"description": "The single biggest problem in communication is the illusion \nthat it has taken place. \n\nGBS", "url": null,
"entities": {"description": {"urls": []}}, "protected": false, "followers_count": 92, "friends_count": 186, "listed_count": 7,
"created_at": "Wed May 16 19:16:13 +0000 2012", "favourites_count": 43, "utc_offset": null, "time_zone": null, "geo_enabled":
false, "verified": false, "statuses_count": 11266, "lang": null, "contributors_enabled": false, "is_translator": false,
"is_translation_enabled": false, "profile_background_color": "131516", "profile_background_image_url": "
http://abs.twimg.com/images/themes/theme14/bg.gif", "profile_background_image_url_https": "
https://abs.twimg.com/images/themes/theme14/bg.gif", "profile_background_tile": true, "profile_image_url": "
http://pbs.twimg.com/profile_images/554372074869776384/izTuyBqh_normal.jpeg", "profile_image_url_https": "
https://pbs.twimg.com/profile_images/554372074869776384/izTuyBqh_normal.jpeg", "profile_banner_url": "
https://pbs.twimg.com/profile_banners/582091384/1404663995", "profile_link_color": "0051DE", "profile_sidebar_border_color":
"EEEEEE", "profile_sidebar_fill_color": "EFEFEF", "profile_text_color": "333333", "profile_use_background_image": true,
"has_extended_profile": false, "default_profile": false, "default_profile_image": false, "following": false,
"follow_request_sent": false, "notifications": false, "translator_type": "none"}, "geo": null, "coordinates": null, "place":
null, "contributors": null, "is_quote_status": true, "quoted_status_id": 1251280025581584384, "quoted_status_id_str":
"1251280025581584384", "quoted_status": {"created_at": "Fri Apr 17 22:43:00 +0000 2020", "id": 1251280025581584384, "id_str":
"1251280025581584384", "text": "The President claims leaders of other powerful countries praise the capabilities of the US off
```

Joe_Biden.json



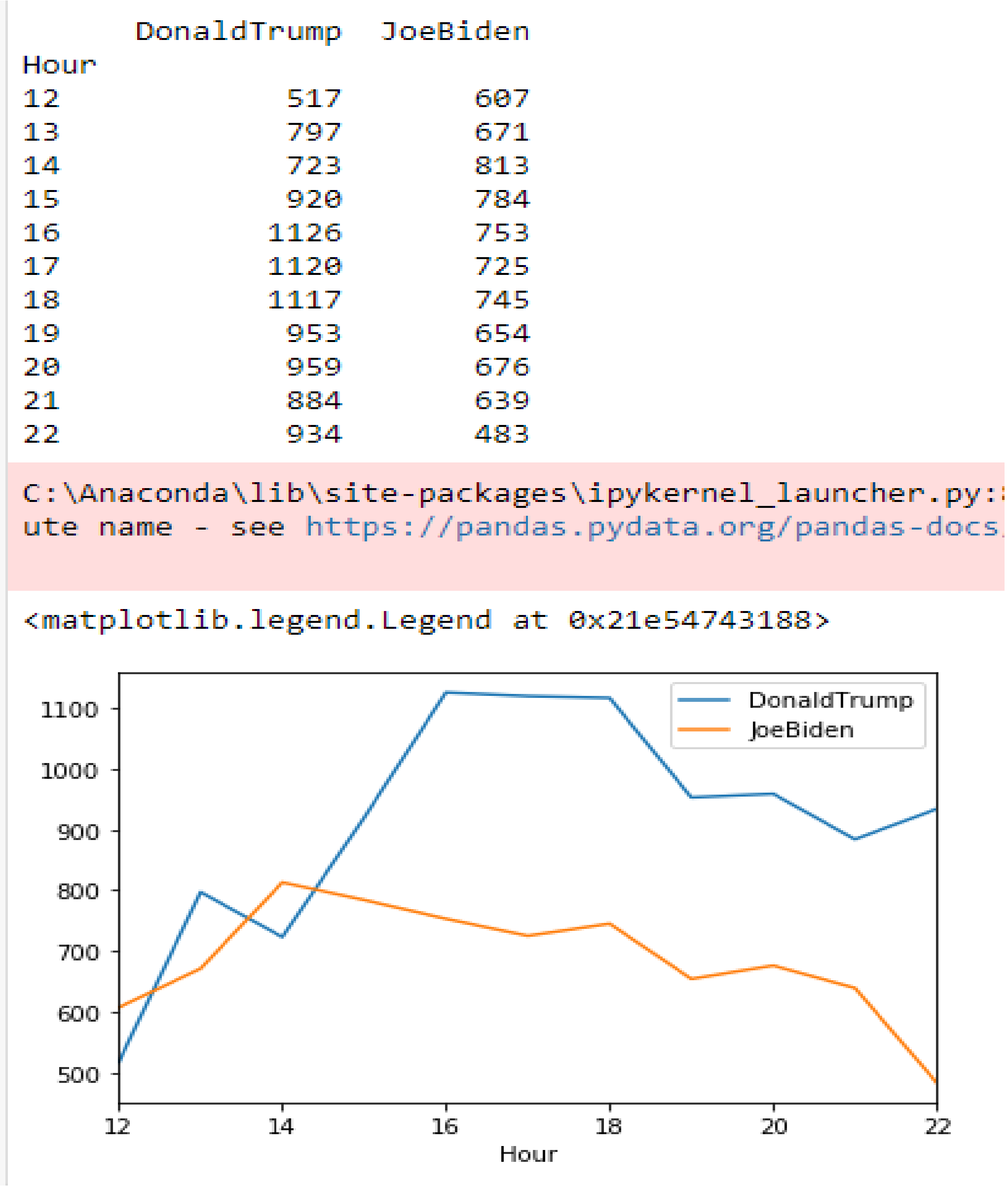
```
[{"created_at": "Fri Apr 17 22:58:51 +0000 2020", "id": 1251284014381621248, "id_str": "1251284014381621248", "text": "Adam
West (family guy) and Joe Biden are the same.", "truncated": false, "entities": {"hashtags": [], "symbols": [], "user_mentions":
 [], "urls": []}, "metadata": {"iso_language_code": "en", "result_type": "recent"}, "source": "<a
href=\"http://twitter.com/download/iphone\" rel=\"nofollow\">Twitter for iPhone</a>", "in_reply_to_status_id": null,
"in_reply_to_status_id_str": null, "in_reply_to_user_id": null, "in_reply_to_user_id_str": null, "in_reply_to_screen_name": null
, "user": {"id": 480947761, "id_str": "480947761", "name": "JuicyFreak", "screen_name": "franklinTG4U", "location": "Denver, CO"
, "description": "Never underestimate the power of stupid people in a large group", "url": null, "entities": {"description": {
"urls": []}}, "protected": false, "followers_count": 251, "friends_count": 299, "listed_count": 0, "created_at": "Thu Feb 02
03:31:58 +0000 2012", "favourites_count": 4192, "utc_offset": null, "time_zone": null, "geo_enabled": true, "verified": false,
"statuses_count": 6709, "lang": null, "contributors_enabled": false, "is_translator": false, "is_translation_enabled": false,
"profile_background_color": "131516", "profile_background_image_url": "http://abs.twimg.com/images/themes/theme9/bg.gif",
"profile_background_image_url_https": "https://abs.twimg.com/images/themes/theme9/bg.gif", "profile_background_tile": true,
"profile_image_url": "http://pbs.twimg.com/profile_images/1248763487356096513/czU9N02U_normal.jpg", "profile_image_url_https":
"https://pbs.twimg.com/profile_images/1248763487356096513/czU9N02U_normal.jpg", "profile_banner_url":
"https://pbs.twimg.com/profile_banners/480947761/1582736305", "profile_link_color": "009998", "profile_sidebar_border_color":
"FFFFFF", "profile_sidebar_fill_color": "252429", "profile_text_color": "666666", "profile_use_background_image": true,
"has_extended_profile": true, "default_profile": false, "default_profile_image": false, "following": false,
"follow_request_sent": false, "notifications": false, "translator_type": "none"}, "geo": null, "coordinates": null, "place":
null, "contributors": null, "is_quote_status": false, "retweet_count": 0, "favorite_count": 0, "favorited": false, "retweeted":
false, "lang": "en"}, {"created_at": "Fri Apr 17 22:58:42 +0000 2020", "id": 1251283975986991105, "id_str":
"1251283975986991105", "text": "Rocker Meredith Brooks Says Joe Biden Touching Children Is Not Normal: 'Curdles My Blood'
https://t.co/QJevMHiVxq v… https://t.co/zcDCK00Jer", "truncated": true, "entities": {"hashtags": [], "symbols": [],
"user_mentions": [], "urls": [{"url": "https://t.co/QJevMHiVxq", "expanded_url":
```

Task 2: Exploratory Analysis

To start with exploratory analysis, I converted the JSON file to a DataFrame by saving it as pickle file. This helped me with the ease of access to all fields of the tweet.

a. Time Series Plot:

As the tweets were obtained for a single day, I have created a time series plot on an hourly basis, from 12 Noon to 10 PM. This plot is a comparison of the number of tweets obtained by the 2 candidates. Please find the plot below:



```
      DonaldTrump  JoeBiden
Hour
12            517       607
13            797       671
14            723       813
15            920       784
16           1126       753
17           1120       725
18           1117       745
19            953       654
20            959       676
21            884       639
22            934       483
```

```
C:\Anaconda\lib\site-packages\ipykernel_launcher.py:
ute name - see https://pandas.pydata.org/pandas-docs
```

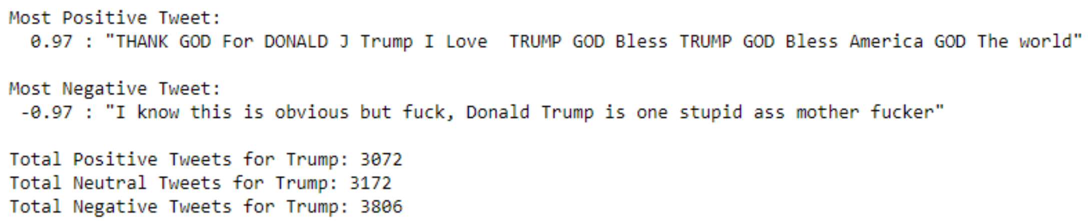
```
<matplotlib.legend.Legend at 0x21e54743188>
```

The table compares the counts of the tweets obtained by them.

b. Sentiment Intensity Analysis:

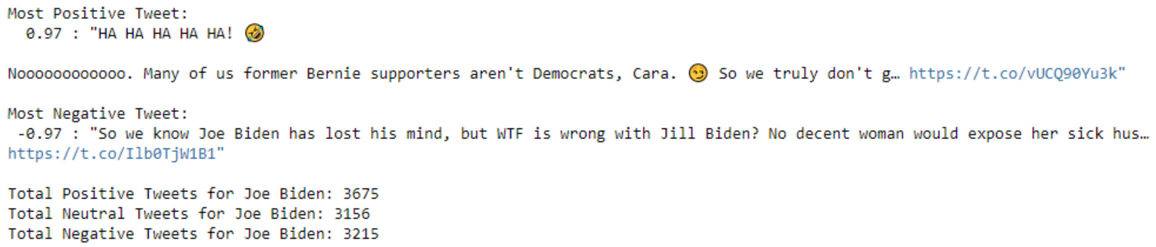
As part of the exploratory analysis, I tried analysis the polarity of the tweets received by each candidate. By classifying the tweets as positive , neutral and negative; we can get an idea of who gets the most positive comments which might say , who could likely win the election. The most positive and negative tweet is also shown:

Donald Trump:



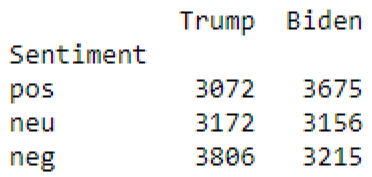
```
Most Positive Tweet:
  0.97 : "THANK GOD For DONALD J Trump I Love  TRUMP GOD Bless TRUMP GOD Bless America GOD The world"

Most Negative Tweet:
 -0.97 : "I know this is obvious but fuck, Donald Trump is one stupid ass mother fucker"

Total Positive Tweets for Trump: 3072
Total Neutral Tweets for Trump: 3172
Total Negative Tweets for Trump: 3806
```

Joe Biden:



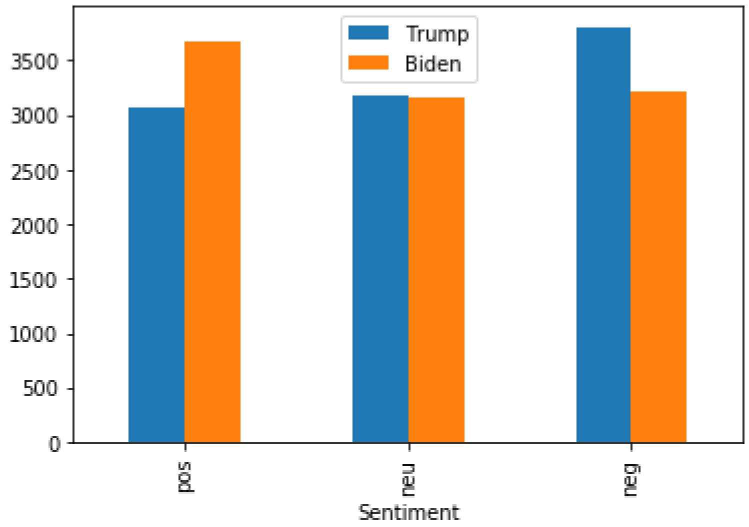
```
Most Positive Tweet:
  0.97 : "HA HA HA HA HA! 🤣

Noooooooooooo. Many of us former Bernie supporters aren't Democrats, Cara. 😏 So we truly don't g… https://t.co/vUCQ90Yu3k"

Most Negative Tweet:
 -0.97 : "So we know Joe Biden has lost his mind, but WTF is wrong with Jill Biden? No decent woman would expose her sick hus…
https://t.co/Ilb0TjW1B1"

Total Positive Tweets for Joe Biden: 3675
Total Neutral Tweets for Joe Biden: 3156
Total Negative Tweets for Joe Biden: 3215
```

The results show that Trump receives more negative tweets when compared to Biden.



```
           Trump  Biden
Sentiment
pos         3072   3675
neu         3172   3156
neg         3806   3215
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x21e0e758a88>
```



As per the sentiment intensity analysis, Biden has more probability of winning the elections.

c. Vader Sentiment Analysis:
This is a sentiment analysis technique that is similar to Sentiment Intensity Analyser but we use the Machine Learning models to train and test the given data.
Steps:

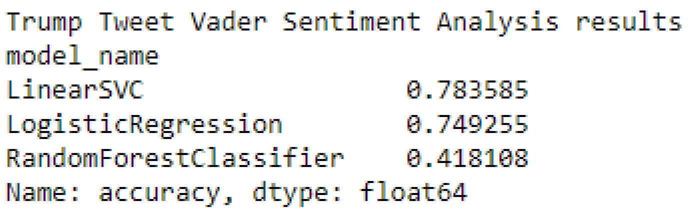
1. Merge both dataframes to a single one using the required columns like “Created at” “User” “Tweet Text” for both the candidates.

2. Create a list of stopwords
3. Create a tokenizer function and perform lemmatization
4. Remove the stop words and clean text
5. Custom Transform using SpaCY
6. Vectorize the text using bow ,TF-iDF
7. Use bag of words vector (parameter ngram_range)

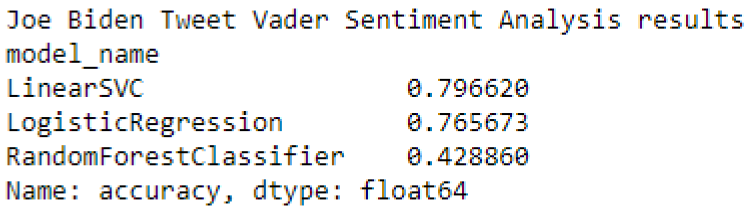
We then perform the

1. Vader sentiment analyzer for trump tweet
2. Define the positive, neutral and negative tweet
3. Label pos/neg/neu based on VaderSentiment result

Similar analysis is done for Biden tweets as well. After this, the data was split into train and test data and 3 Machine Learning models RandomForestClassifier, LinearSVC & LogisticRegression were used for classification. The results are as follows:



```
Trump Tweet Vader Sentiment Analysis results
model_name
LinearSVC                 0.783585
LogisticRegression        0.749255
RandomForestClassifier    0.418108
Name: accuracy, dtype: float64
```



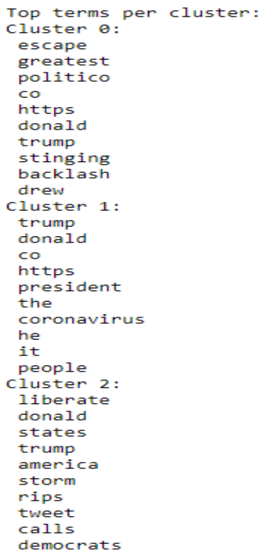
```
Joe Biden Tweet Vader Sentiment Analysis results
model_name
LinearSVC                 0.796620
LogisticRegression        0.765673
RandomForestClassifier    0.428860
Name: accuracy, dtype: float64
```

As we can see, accuracy for Joe's Tweets are slightly higher than Trump tweets. Thus we can conclude that Biden has slightly more chances of winning the election when compared to Trump.

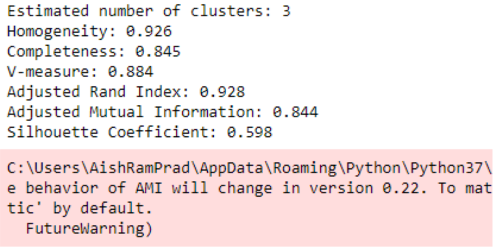
d. KMeans Clustering

I tried clustering of tweets using KMeans. Here "K" was taken as 3 as it gave the most clear clusters and sample size considered was 500.

Trump KMeans:

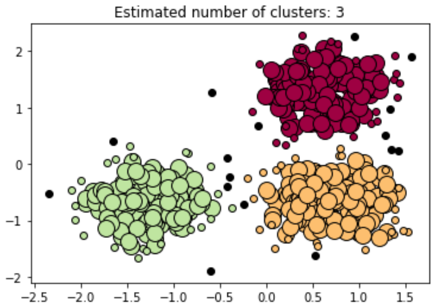


```
Top terms per cluster:
Cluster 0:
 escape
 greatest
 politico
 co
 https
 donald
 trump
 stinging
 backlash
 drew
Cluster 1:
 trump
 donald
 co
 https
 president
 the
 coronavirus
 he
 it
 people
Cluster 2:
 liberate
 donald
 states
 trump
 america
 storm
 rips
 tweet
 calls
 democrats
```

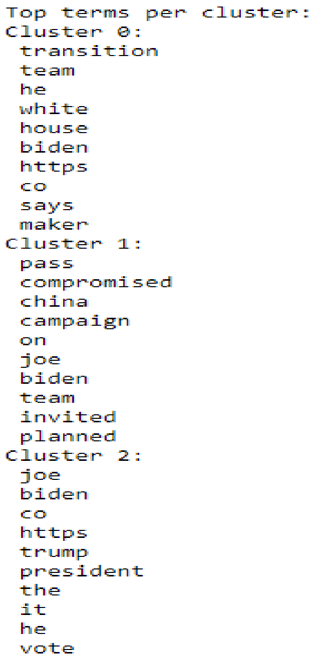


```
Estimated number of clusters: 3
Homogeneity: 0.926
Completeness: 0.845
V-measure: 0.884
Adjusted Rand Index: 0.928
Adjusted Mutual Information: 0.844
Silhouette Coefficient: 0.598
```

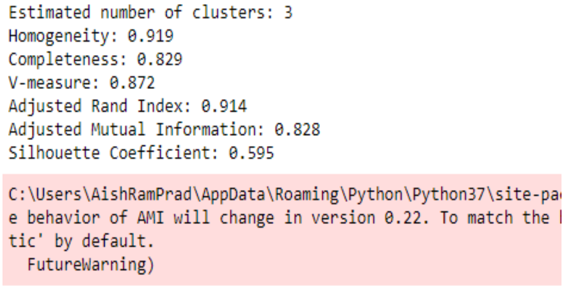
```
C:\Users\AishRamPrad\AppData\Roaming\Python\Python37\
e behavior of AMI will change in version 0.22. To mat
tic' by default.
  FutureWarning)
```



Biden KMeans:

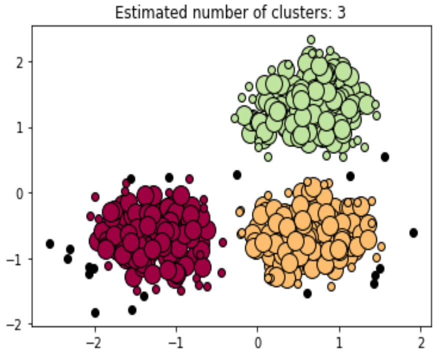


```
Top terms per cluster:
Cluster 0:
 transition
 team
 he
 white
 house
 biden
 https
 co
 says
 maker
Cluster 1:
 pass
 compromised
 china
 campaign
 on
 joe
 biden
 team
 invited
 planned
Cluster 2:
 joe
 biden
 co
 https
 trump
 president
 the
 it
 he
 vote
```



```
Estimated number of clusters: 3
Homogeneity: 0.919
Completeness: 0.829
V-measure: 0.872
Adjusted Rand Index: 0.914
Adjusted Mutual Information: 0.828
Silhouette Coefficient: 0.595
```

```
C:\Users\AishRamPrad\AppData\Roaming\Python\Python37\site-pa
e behavior of AMI will change in version 0.22. To match the 
tic' by default.
  FutureWarning)
```



Based on the above exploratory analysis and time series plot on single day tweets, **Joe Biden** has better chance of winning the 2020 US Presidential Elections when compared to Donald Trump.