**Sentiment Analysis using Twitter API to predict election results based on tweets**

**Donald Trump vs. Joe Biden**

This project includes the basics of social media mining by collecting Twitter data, pre-processing the data, and conducting exploratory analysis.

**Task 1: Collecting Twitter Data**

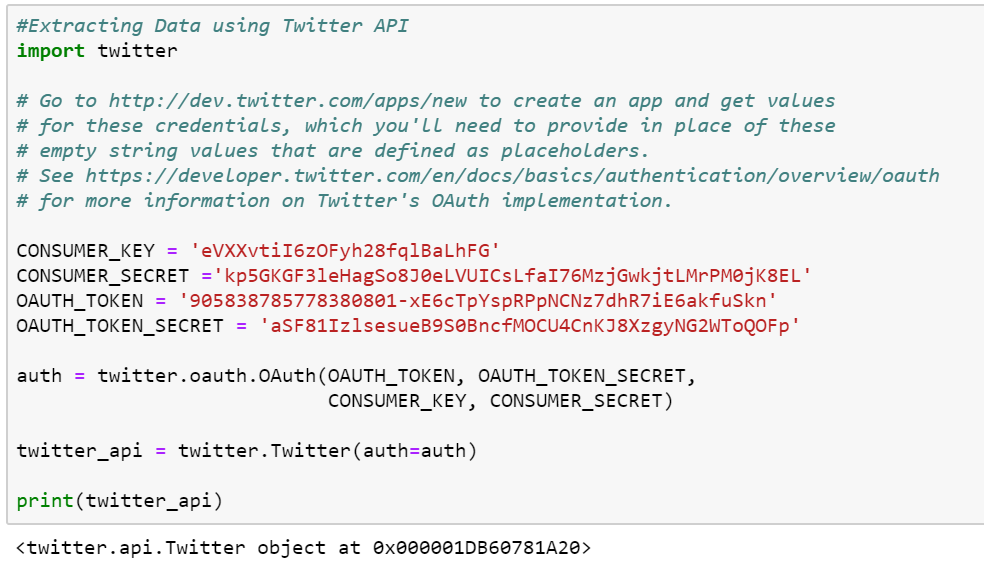
In order to collect tweets from Twitter API followed the below steps:

1. **Create a Twitter Developer Account**: Give application details/descriptions and click on create application.
2. **Key Generation**: Obtain Consumer API ‘Keys and Access Tokens’ used for authentication of twitter API.

Install the twitter package on the system using the command: **‘pip install twitter’**

For fetching tweets using the twitter API we first need to complete authentication using the keys and access tokens obtained after account creation.

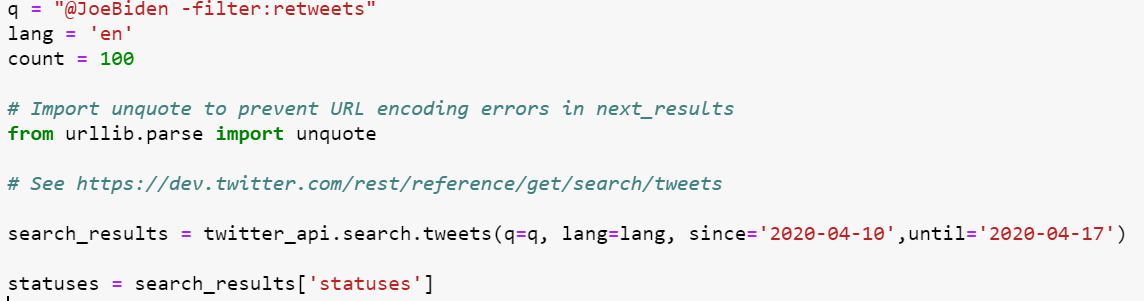
**Twitter Authentication:**



Next, we need to find tweets related to Joe Biden and Donald trump using the Twitter API. Hence, we can mention their twitter user ids in our search query and save the tweets in a JSON file.

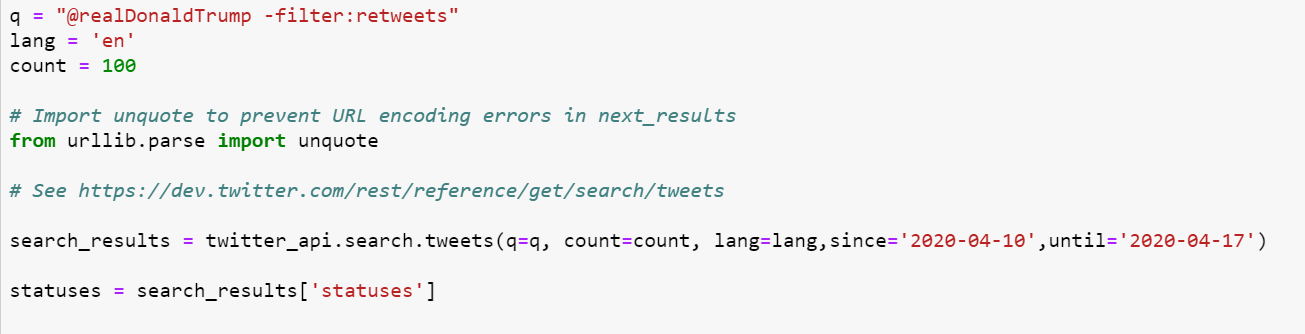
**Tweets should be written in English:**

In order to get all the tweets in English we need to set the “**lang = en”** in the search query as shown below in the code snippet.

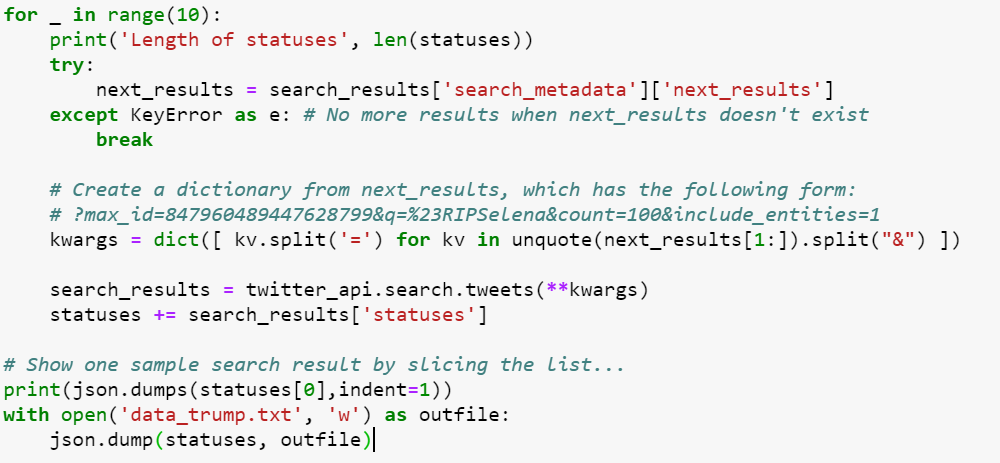


**Remove/filter retweets to get more meaningful information if you can get at least 1,000 tweets.**

In order to remove/filter retweets we need to set a filter in the search query **“-filter: retweets”**



To get at least 1000 tweets, we need to access the ‘**next\_results’** field from the ‘**search\_metadata’**.The API returns maximum of **100 tweets per page** and to access the remaining tweets we need to use the next\_results field.



**Sample tweets for Joe Biden in JSON:**

{

"created\_at": "Thu Apr 16 23:59:59 +0000 2020",

"id": 1250937009528549378,

"id\_str": "1250937009528549378",

"text": "@KamalaHarris Or mail in ? I sure hope that @JoeBiden endorses you for VP @KamalaHarris. I would be the happiest pe\u2026 [https://t.co/j4LV8T5ZcF",](https://t.co/j4LV8T5ZcF%22,)

"truncated": true,

"entities": {

"hashtags": [],

"symbols": [],

"user\_mentions": [

{

"screen\_name": "KamalaHarris",

"name": "Kamala Harris",

"id": 30354991,

"id\_str": "30354991",

"indices": [

0,

13

]

},

{

"screen\_name": "JoeBiden",

"name": "Joe Biden",

"id": 939091,

"id\_str": "939091",

"indices": [

44,

53

]

},

{

"screen\_name": "KamalaHarris",

"name": "Kamala Harris",

"id": 30354991,

"id\_str": "30354991",

"indices": [

74,

87

]

}

],

"urls": [

{

"url": "https://t.co/j4LV8T5ZcF",

"expanded\_url": "https://twitter.com/i/web/status/1250937009528549378",

"display\_url": "twitter.com/i/web/status/1\u2026",

"indices": [

117,

140

]

}

]

},

"metadata": {

"iso\_language\_code": "en",

"result\_type": "recent"

},

"source": "<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for Android</a>",

"in\_reply\_to\_status\_id": 1250935507242176512,

"in\_reply\_to\_status\_id\_str": "1250935507242176512",

"in\_reply\_to\_user\_id": 30354991,

"in\_reply\_to\_user\_id\_str": "30354991",

"in\_reply\_to\_screen\_name": "KamalaHarris",

"user": {

"id": 1249054457427566593,

"id\_str": "1249054457427566593",

"name": "Lori",

"screen\_name": "Lori90920955",

"location": "",

"description": "live your life, love everyone. Love my fur babies. grammy of 5 ! We got to take back our Country. Pray that WE VOTE HIM OUT #Bluewave2020",

"url": null,

"entities": {

"description": {

"urls": []

}

},

"protected": false,

"followers\_count": 5,

"friends\_count": 244,

"listed\_count": 0,

"created\_at": "Sat Apr 11 19:19:46 +0000 2020",

"favourites\_count": 205,

"utc\_offset": null,

"time\_zone": null,

"geo\_enabled": false,

"verified": false,

"statuses\_count": 210,

"lang": null,

"contributors\_enabled": false,

"is\_translator": false,

"is\_translation\_enabled": false,

"profile\_background\_color": "F5F8FA",

"profile\_background\_image\_url": null,

"profile\_background\_image\_url\_https": null,

"profile\_background\_tile": false,

"profile\_image\_url": "http://pbs.twimg.com/profile\_images/1249054747212025856/MRMyohig\_normal.jpg",

"profile\_image\_url\_https": "https://pbs.twimg.com/profile\_images/1249054747212025856/MRMyohig\_normal.jpg",

"profile\_banner\_url": "https://pbs.twimg.com/profile\_banners/1249054457427566593/1586640713",

"profile\_link\_color": "1DA1F2",

"profile\_sidebar\_border\_color": "C0DEED",

"profile\_sidebar\_fill\_color": "DDEEF6",

"profile\_text\_color": "333333",

"profile\_use\_background\_image": true,

"has\_extended\_profile": true,

"default\_profile": true,

"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": 2,

"favorite\_count": 2,

"favorited": false,

"retweeted": false,

"lang": "en"

}

**Sample tweets for Donald Trump in JSON:**

{

"created\_at": "Thu Apr 16 23:59:59 +0000 2020",

"id": 1250937012816801792,

"id\_str": "1250937012816801792",

"text": "@Bunny\_Slick @carl\_cnp @JayMercer20 @itsJeffTiedrich @realDonaldTrump Jan 31 Trump Administration declared the coro\u2026 [https://t.co/6YwtfetL0n",](https://t.co/6YwtfetL0n%22,)

"truncated": true,

"entities": {

"hashtags": [],

"symbols": [],

"user\_mentions": [

{

"screen\_name": "Bunny\_Slick",

"name": "Bunny Slick",

"id": 793949726483693569,

"id\_str": "793949726483693569",

"indices": [

0,

12

]

},

{

"screen\_name": "carl\_cnp",

"name": "Carl Purseglove",

"id": 147672777,

"id\_str": "147672777",

"indices": [

13,

22

]

},

{

"screen\_name": "JayMercer20",

"name": "Jay Mercer",

"id": 1219007709757886464,

"id\_str": "1219007709757886464",

"indices": [

23,

35

]

},

{

"screen\_name": "itsJeffTiedrich",

"name": "Jeff Tiedrich",

"id": 1009577803304656896,

"id\_str": "1009577803304656896",

"indices": [

36,

52

]

},

{

"screen\_name": "realDonaldTrump",

"name": "Donald J. Trump",

"id": 25073877,

"id\_str": "25073877",

"indices": [

53,

69

]

}

],

"urls": [

{

"url": "https://t.co/6YwtfetL0n",

"expanded\_url": "https://twitter.com/i/web/status/1250937012816801792",

"display\_url": "twitter.com/i/web/status/1\u2026",

"indices": [

117,

140

]

}

]

},

"metadata": {

"iso\_language\_code": "en",

"result\_type": "recent"

},

"source": "<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for Android</a>",

"in\_reply\_to\_status\_id": 1250935646396362752,

"in\_reply\_to\_status\_id\_str": "1250935646396362752",

"in\_reply\_to\_user\_id": 1096464427,

"in\_reply\_to\_user\_id\_str": "1096464427",

"in\_reply\_to\_screen\_name": "Hammeredge1",

"user": {

"id": 1096464427,

"id\_str": "1096464427",

"name": "Hammeredge",

"screen\_name": "Hammeredge1",

"location": "Earth #Qanon #WeAreTheNewsNow",

"description": "Jer 23:29 Is not my word like as a fire? saith the LORD; and like a hammer that breaketh the rock in pieces? #Qanon #WeAreTheNewsNow #FactsMatter #WWG1WGA",

"url": null,

"entities": {

"description": {

"urls": []

}

},

"protected": false,

"followers\_count": 1419,

"friends\_count": 1431,

"listed\_count": 2,

"created\_at": "Wed Jan 16 21:52:18 +0000 2013",

"favourites\_count": 14339,

"utc\_offset": null,

"time\_zone": null,

"geo\_enabled": false,

"verified": false,

"statuses\_count": 19239,

"lang": null,

"contributors\_enabled": false,

"is\_translator": false,

"is\_translation\_enabled": false,

"profile\_background\_color": "000000",

"profile\_background\_image\_url": "http://abs.twimg.com/images/themes/theme1/bg.png",

"profile\_background\_image\_url\_https": "https://abs.twimg.com/images/themes/theme1/bg.png",

"profile\_background\_tile": false,

"profile\_image\_url": "http://pbs.twimg.com/profile\_images/1248117452598161408/zDv0HR3l\_normal.jpg",

"profile\_image\_url\_https": "https://pbs.twimg.com/profile\_images/1248117452598161408/zDv0HR3l\_normal.jpg",

"profile\_banner\_url": "https://pbs.twimg.com/profile\_banners/1096464427/1551010712",

"profile\_link\_color": "1B95E0",

"profile\_sidebar\_border\_color": "000000",

"profile\_sidebar\_fill\_color": "000000",

"profile\_text\_color": "000000",

"profile\_use\_background\_image": false,

"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"

}

**Task 2: Exploratory Analysis**

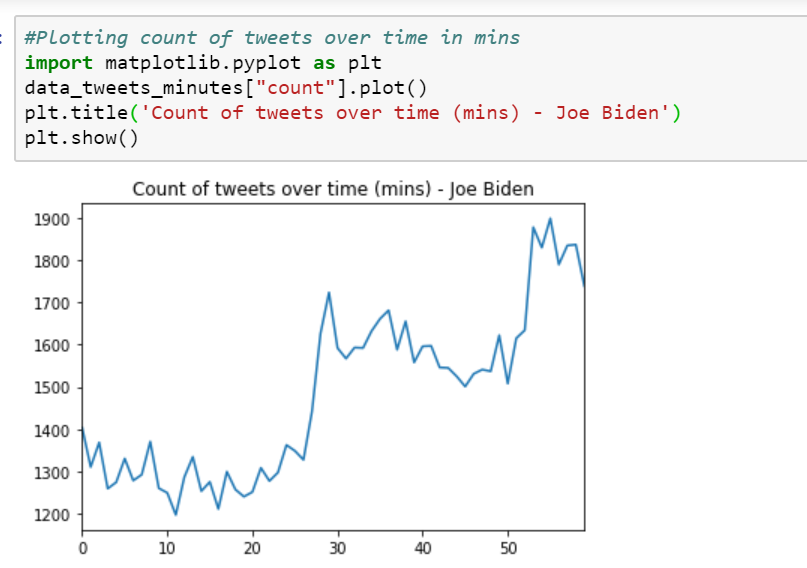
* **A time series figure with the number of tweets per day over time for both candidates**

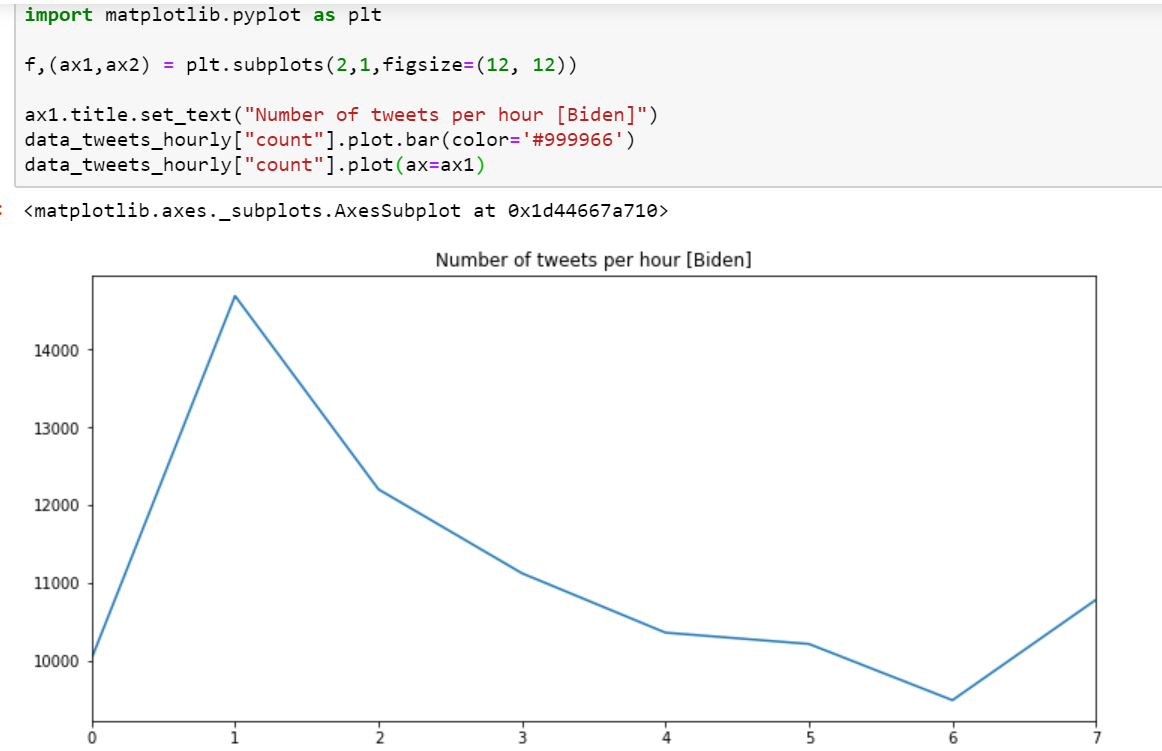
**NOTE: Collecting tweets for each day was taking a lot of time. Hence, I have used tweepy to collect the tweets for 14 April, 2020 from 4:00 pm to 23:59 pm for Joe Biden and 14 April, 2020 from 23:00 pm to 23:59 pm for Donald Trump.**

**Tweet Collection:**

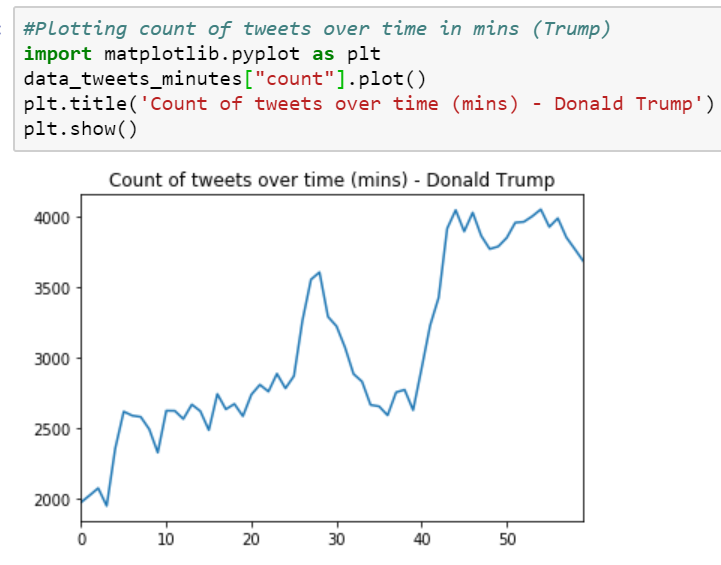


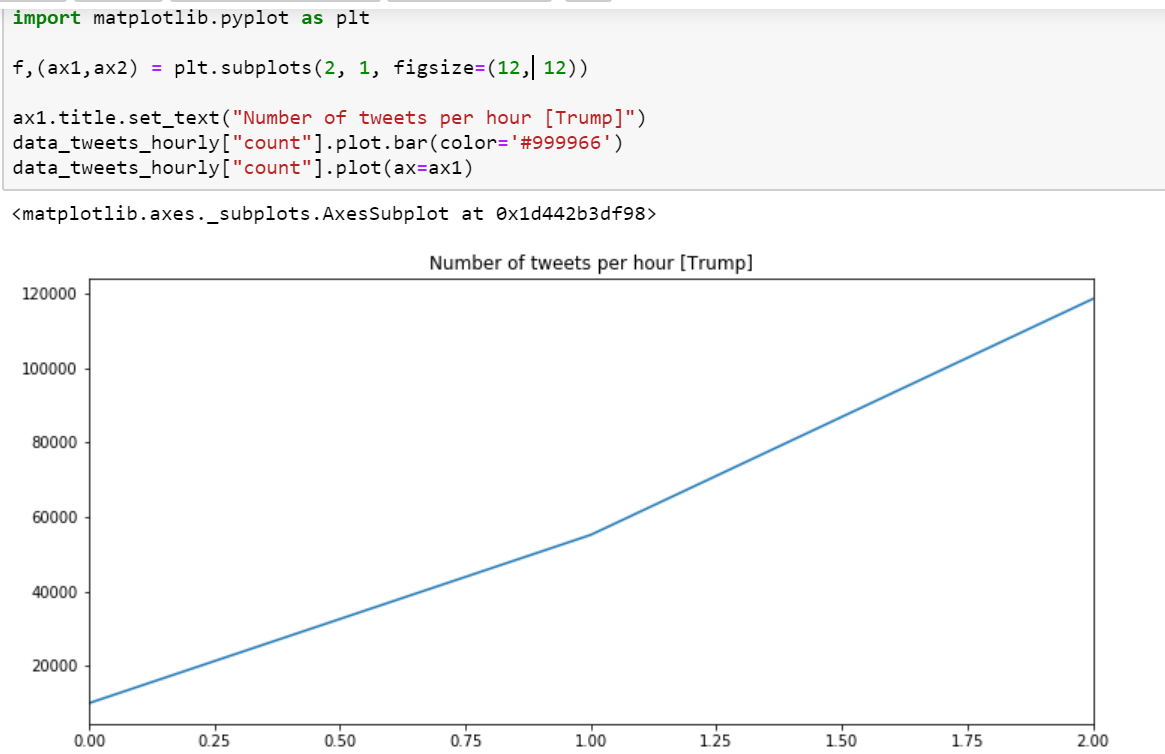
**Time Series [Joe Biden]:**





**Time Series [Donald Trump]**





**Perform two more interesting analysis of your choice (e.g., sentiment analysis, clustering and so on).**

**1] Sentiment Analysis using VADER:**

I have used python’s **VADER** package for doing sentiment analysis. VADER (**Valence Aware Dictionary and sentiment Reasoner**) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.

**VADER** provides us with how negative or positive a tweet is and it doesn’t require any training data but is constructed from a generalizable, valence-based, human-curated gold standard sentiment lexicon.

* We first need to install VADER package on the system:

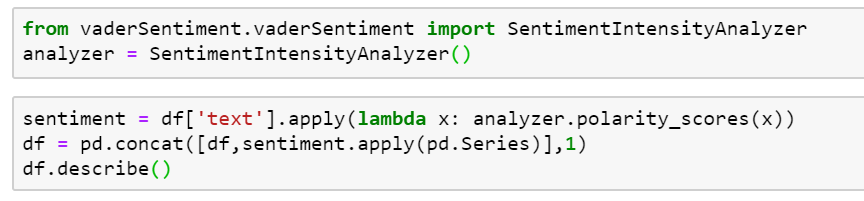
**pip install vaderSentiment**

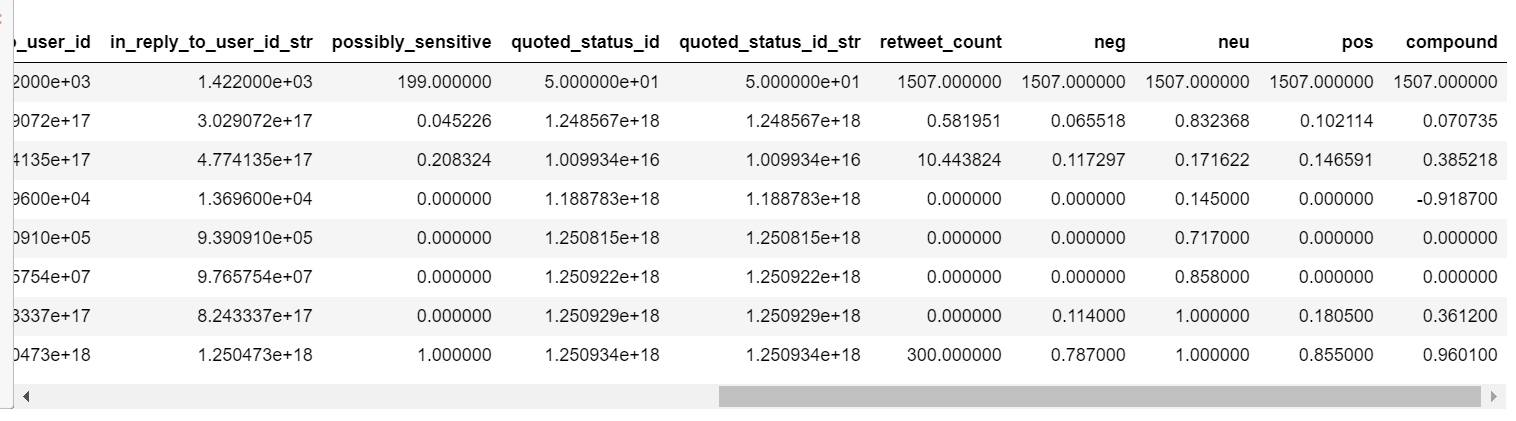
* It uses polarity\_scores() to get the polarity score of each sentence.
* This function returns four scores based on the input sentence i.e Positive Score, Negative Score, Compound Score and Neutral Score.
* The compound score is the normalized weighted composite score and is the single most useful metric to determine the sentiment of the sentence.

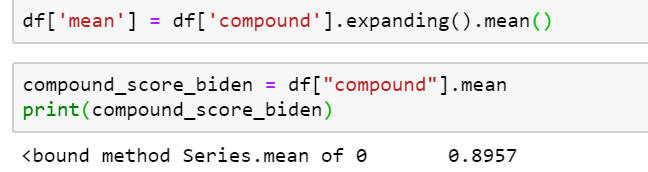
Below is the code snippet of VADER sentiment analyser.

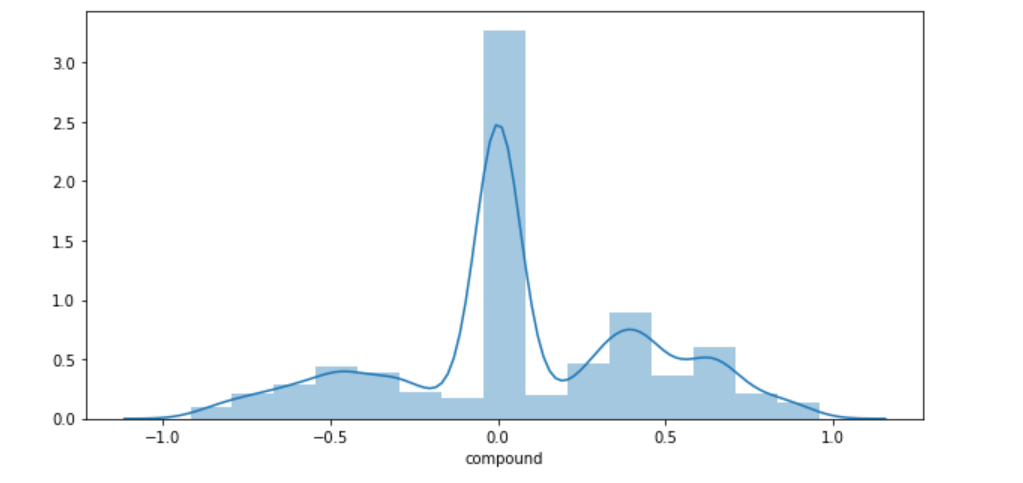
**VADER for Biden:**

**Compound Score: 0.8957**



****

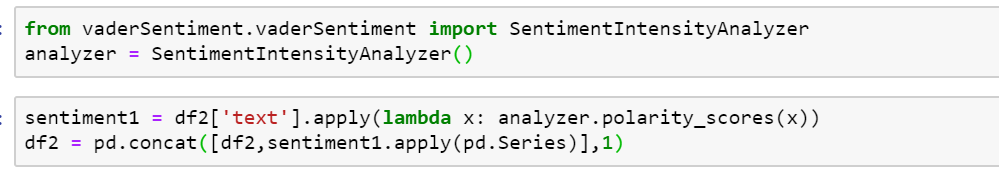


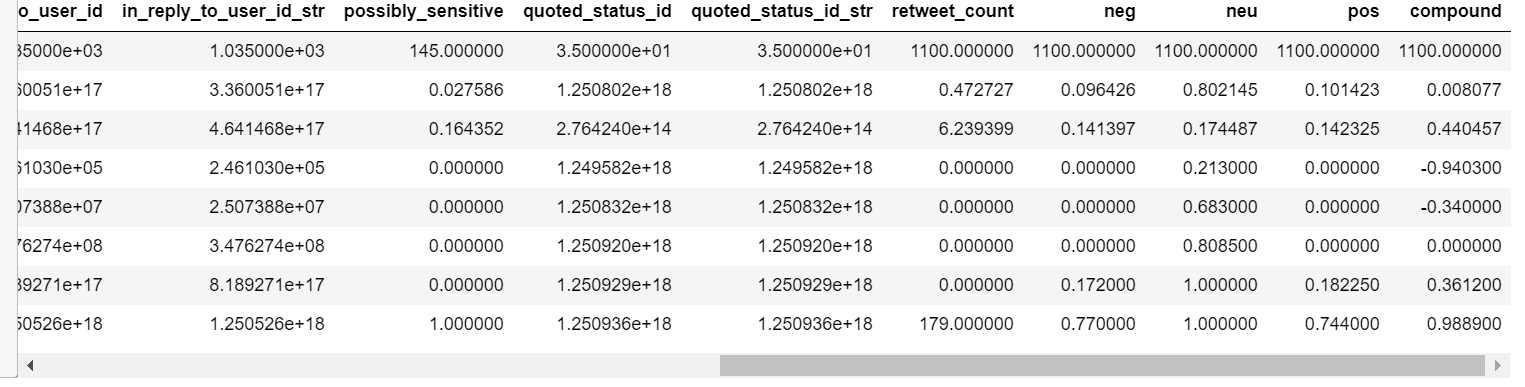


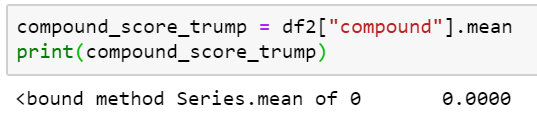
**The mean compound score for Biden is 0.8957 which means that the overall sentiment of the public towards Biden is positive.**

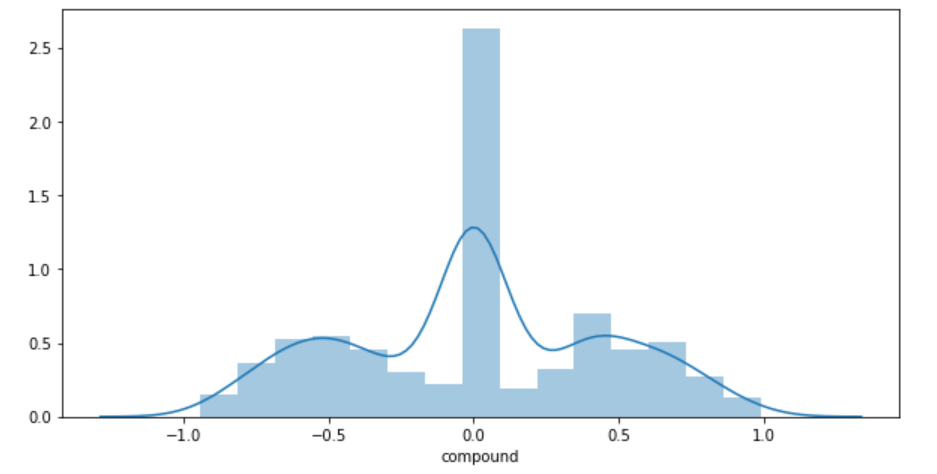
**VADER for Trump:**

**Compound Score: 0.00**

****

****

****

****

**The mean compound score for Trump is 0.00 which means that the overall sentiment of the public towards Trump is neutral.**

**Based on the compound scores of VADER analysis, Joe Biden would win the 2020 Elections as he has a more positive compound score as compared to Trump.**

**2]Sentiment Analysis using Textblob (NLTK):**

Textblob is the python library for processing textual data.

Install it using following command:

**pip install textblob**

Also, we need to install some NLTK corpora using following command:

**python -m textblob.download\_corpora**

The code includes 3 major steps in our program:

* Authorize twitter API client.
* Make a GET request to Twitter API to fetch tweets for a particular query.
* Parse the tweets. Classify each tweet as positive, negative or neutral.

First, the **clean\_tweet** method is used to remove links, special characters, etc. from the tweet using some simple regex. Then, as we pass **tweet** to create a **TextBlob** object, following processing is done over text by textblob library:

* Tokenize the tweet, i.e. split words from body of text.
* Remove stop words from the tokens.
* Do POS (part of speech) tagging of the tokens and select only significant features/tokens like adjectives, adverbs, etc.
* Pass the tokens to a **sentiment classifier**which classifies the tweet sentiment as positive, negative or neutral by assigning it a polarity between -1.0 to 1.0.
* Positive and negative features are extracted from each positive and negative review respectively.
* Training data now consists of labelled positive and negative features. This data is trained on a Naive Bayes Classifier.
* Then, use **sentiment.polarity** method of **TextBlob** class to get the polarity of tweet between -1 to 1.
* Finally, parsed tweets are returned. Then, we can do various type of statistical analysis on the tweets. For example, in the code below I tried to find the percentage of positive, negative and neutral tweets about a query.

****

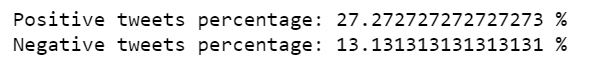
**Analysis for Biden:**

Positive tweets percentage: 29.0 %

Negative tweets percentage: 17.0 %

**Based on the above scores, it can be said that Biden has a more positive impact on the audience.**

**Analysis for Trump:**

****

**Based on the above scores, it can be said that Trump has a more positive impact on the audience. However, a lesser positive impact than Biden. Hence, it can be concluded that Joe Biden has higher chances of winning the 2020 elections as per the above analysis.**