1. Introduction:-

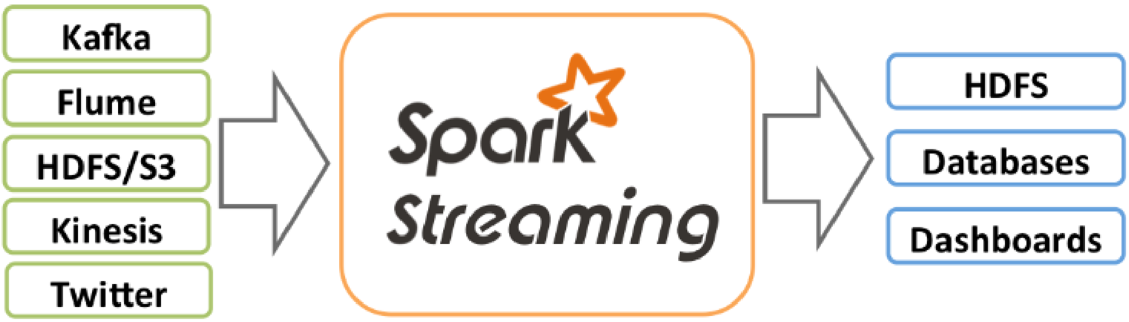
Sentiment analysis is the interpretation and classification of emotions (positive, negative and neutral) within text data using text analysis techniques. Sentiment analysis allows businesses to identify customer sentiment toward products, brands or services in online conversations and feedback.

Sentiment analysis models detect polarity within a text (e.g. a *positive* or *negative* opinion), whether it's a whole document, paragraph, sentence, or clause.
Understanding people's emotions is essential for businesses since customers are able to express their thoughts and feelings more openly than ever before. By automatically analyzing customer feedback, from survey responses to social media conversations, brands are able to listen attentively to their customers, and tailor products and services to meet their needs.

2. Technology Used:-

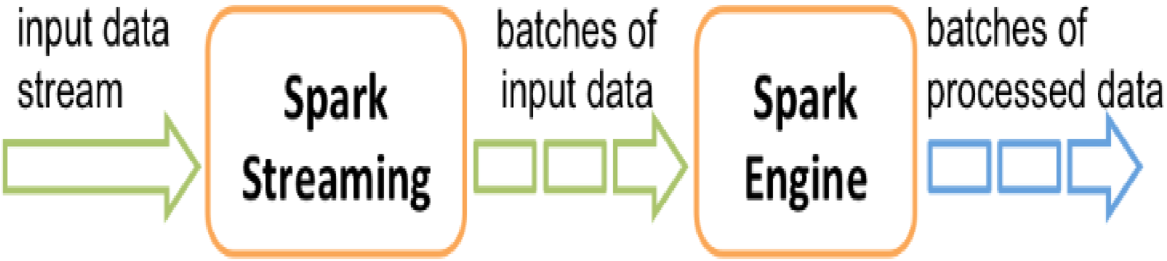
2.1 Apache Spark:-

Spark Streaming is an extension of the core Spark API that enables scalable, high-throughput, fault-tolerant stream processing of live data streams. Data can be ingested from many sources like Kafka, Flume, Kinesis, or TCP sockets, and can be processed using complex algorithms expressed with high-level functions like map, reduce, join and window. Finally, processed data can be pushed out to filesystems, databases, and live dashboards. In fact, you can apply Spark's machine learning and graph processing algorithms on data streams.



In this project we are using **Twitter API** python module for getting data into the system.

Internally, it works as follows. Spark Streaming receives live input data streams and divides the data into batches, which are then processed by the Spark engine to generate the final stream of results in batches.



3. Libraries Used:-

3.1 PySpark:- PySpark is the Python API written in python to support Apache Spark. Apache Spark is a distributed framework that can handle Big Data analysis. Apache Spark is written in Scala and can be integrated with Python, Scala, Java, R, SQL languages. Spark is basically a computational engine, that works with huge sets of data by processing them in parallel and batch systems.

3.2 Tweepy:- Tweepy is a Python library for accessing the Twitter API. It is great for simple automation and creating twitter bots.

3.3 TextBlob:- TextBlob is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more

3.4 **NumPy:-** NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.

3.5 **Pandas:-** Pandas is the most popular python library that is used for data analysis.

3.6 **Re:-** A regular expression (or RE) specifies a set of strings that matches it; the functions in this module let you check if a particular string matches a given regular expression (or if a given regular expression matches a particular string, which comes down to the same thing).

4. Code and Explanation of working with Outputs:-

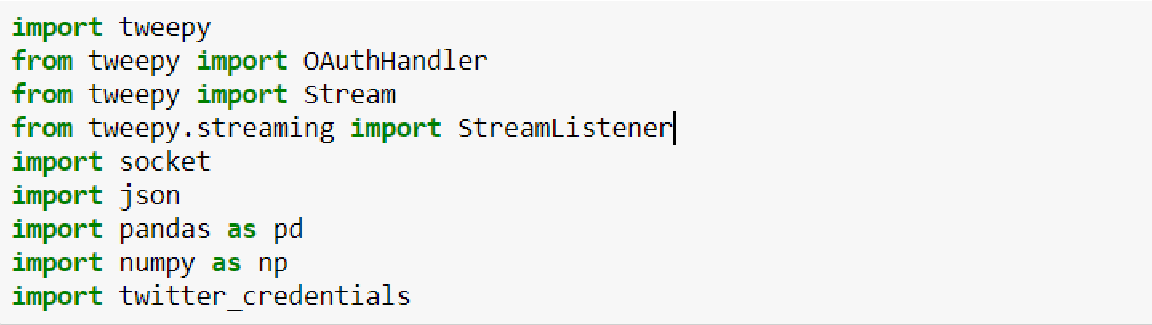
This whole project consists of 4 jupyter notebooks and one twitter credential file which consists of keys and tocken of twitter apps.

1. **TwitterStreaming.ipynb**
2. **SparkStreaming.ipynb**
3. **SentimentAnalysis.ipynb**
4. **Analytics.ipynb**

twitter_credentials.py

These files holds all the functioning of the project

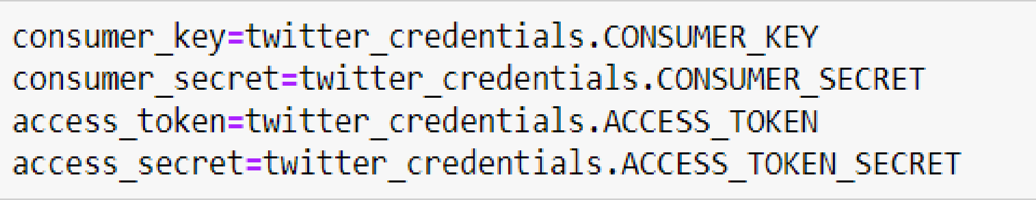
4.1 TwitterStreaming.ipynb



```
import tweepy
from tweepy import OAuthHandler
from tweepy import Stream
from tweepy.streaming import StreamListener
import socket
import json
import pandas as pd
import numpy as np
import twitter_credentials
```

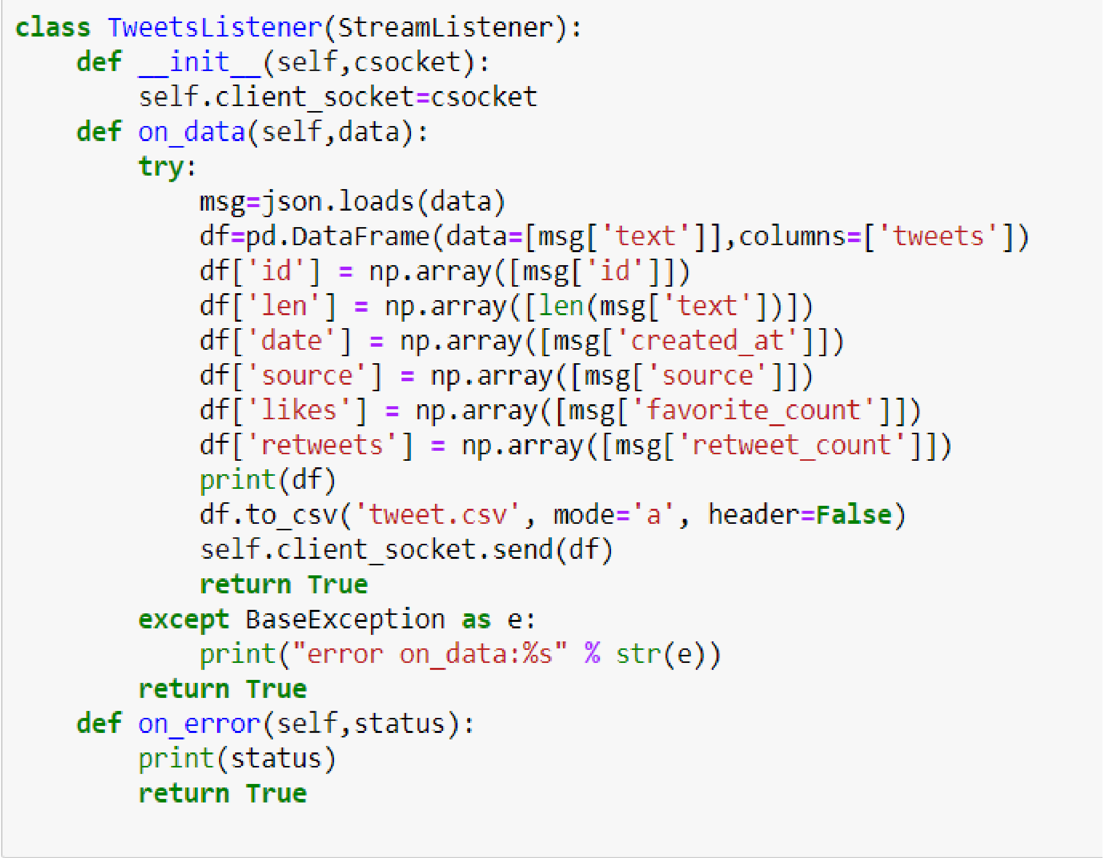
In this we are importing all the libraries which we need for getting data out

Note:- twiiter_credentials is my twitter credentials file



```
consumer_key=twitter_credentials.CONSUMER_KEY
consumer_secret=twitter_credentials.CONSUMER_SECRET
access_token=twitter_credentials.ACCESS_TOKEN
access_secret=twitter_credentials.ACCESS_TOKEN_SECRET
```

Getting all the twitter credentials for streaming twitter data



```
class TweetsListener(StreamListener):
    def __init__(self,csocket):
        self.client_socket=csocket
    def on_data(self,data):
        try:
            msg=json.loads(data)
            df=pd.DataFrame(data=[msg['text']],columns=['tweets'])
            df['id'] = np.array([msg['id']])
            df['len'] = np.array([len(msg['text'])])
            df['date'] = np.array([msg['created_at']])
            df['source'] = np.array([msg['source']])
            df['likes'] = np.array([msg['favorite_count']])
            df['retweets'] = np.array([msg['retweet_count']])
            print(df)
            df.to_csv('tweet.csv', mode='a', header=False)
            self.client_socket.send(df)
            return True
        except BaseException as e:
            print("error on_data:%s" % str(e))
        return True
    def on_error(self,status):
        print(status)
        return True
```

This class is most import in this notebook because it will get tweet and get only important parts of the data and then save it to the .csv file

Msg=json.loads(data) //loads all the tweet data to msg

df=pd.DataFrame(data=[msg[‘text’]],columns=[‘tweet’])

// in this line of code we are extracting only text of tweet it is most important part of this project because from this part only we will do sentiment analysis.

df['id'] = np.array([msg['id']]) //getting id of the tweet

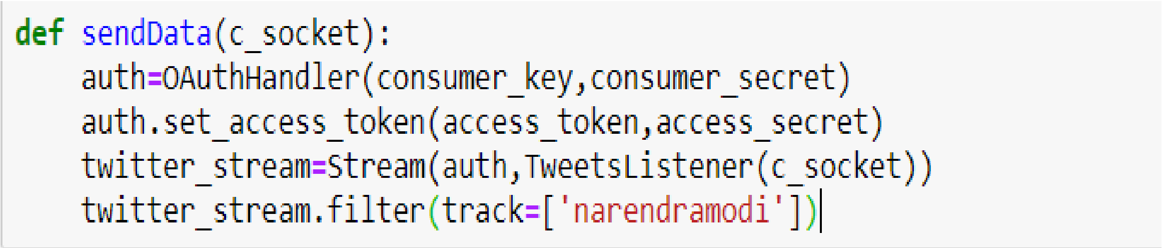
df['len'] = np.array([len(msg['text'])]) //getting length of the tweet text

df['date'] = np.array([msg['created_at']]) //getting time at which tweet is tweeted it is under created_at

df['source'] = np.array([msg['source']]) //this field tell from which device twitter is used

df['likes'] = np.array([msg['favorite_count']]) //likes on a particular tweet

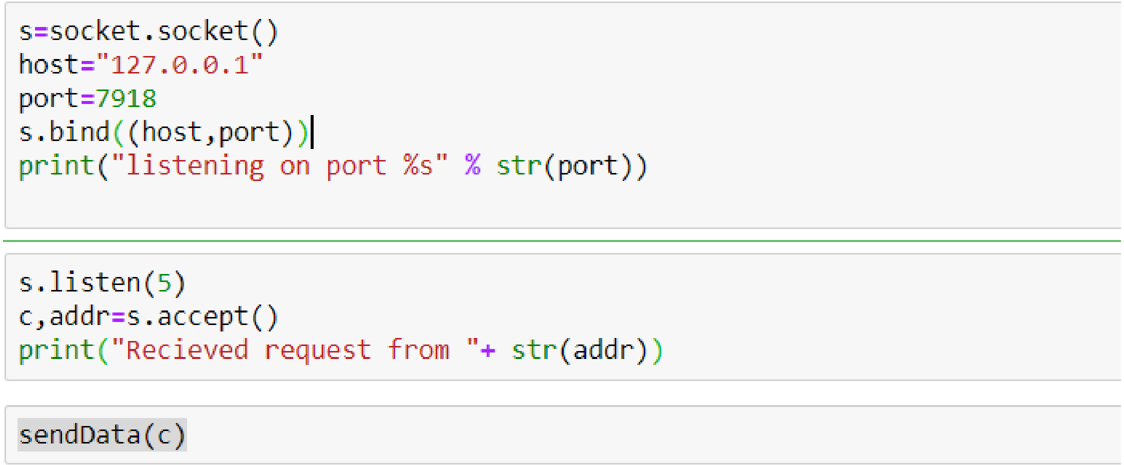
df['retweets'] = np.array([msg['retweet_count']]) //retweet count



```
def sendData(c_socket):
    auth=OAuthHandler(consumer_key,consumer_secret)
    auth.set_access_token(access_token,access_secret)
    twitter_stream=Stream(auth,TweetsListener(c_socket))
    twitter_stream.filter(track=['narendramodi'])
```

In this part of code we are going to pass twitter credentials to a method **OAuthHandler** which will pass the credentials to extract the tweet from the twitter

Twitter_stream.filter(track=[‘narendramodi’]) //this line basically extract the tweets of narendramodi only



```
s=socket.socket()
host="127.0.0.1"
port=7918
s.bind((host,port))
print("listening on port %s" % str(port))
```

```
s.listen(5)
c,addr=s.accept()
print("Recieved request from "+ str(addr))
```

```
sendData(c)
```

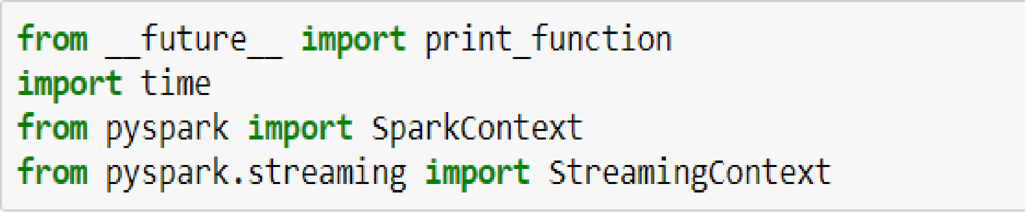
This part helps to activate the port and host to stream incoming tweets

s.listen(5) //will listen the incoming tweets for 5 seconds

c,addr=s.accept() //will accept the tweet data and pass it to c

4.2 SparkStreaming.ipynb

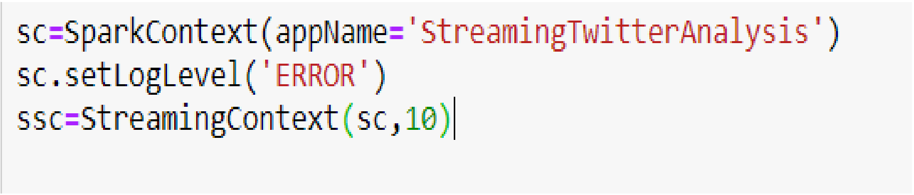
In this jupyter notebook we will make a request to twitter api to fetch the tweets



```
from __future__ import print_function
import time
from pyspark import SparkContext
from pyspark.streaming import StreamingContext
```

Importing all the basic liabraries

Print_function:-use to print the outputs

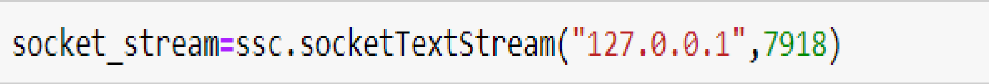


```
sc=SparkContext(appName='StreamingTwitterAnalysis')
sc.setLogLevel('ERROR')
ssc=StreamingContext(sc,10)
```

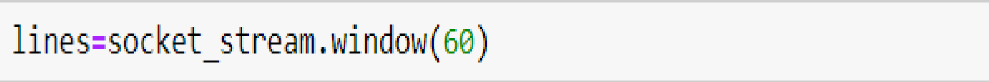
Initalizing the spark context and our app name=StreamingTwitterAnalysis

In this every 10 seconds we are receiving data from twitter API

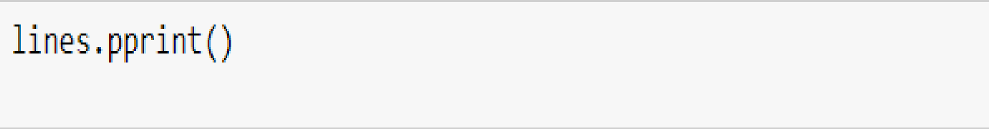
So 10 seconds is our batch size



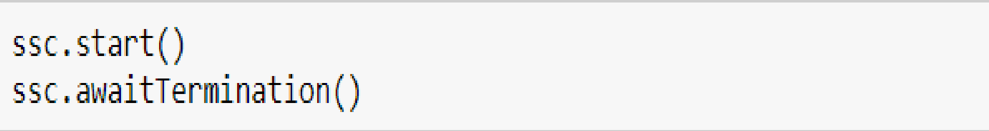
```
socket_stream=ssc.socketTextStream("127.0.0.1",7918)
```



```
lines=socket_stream.window(60)
```



```
lines.pprint()
```



```
ssc.start()
ssc.awaitTermination()
```

First we are streaming on port and host we are using same host and port as we use in Twitter API so to fetch incoming data

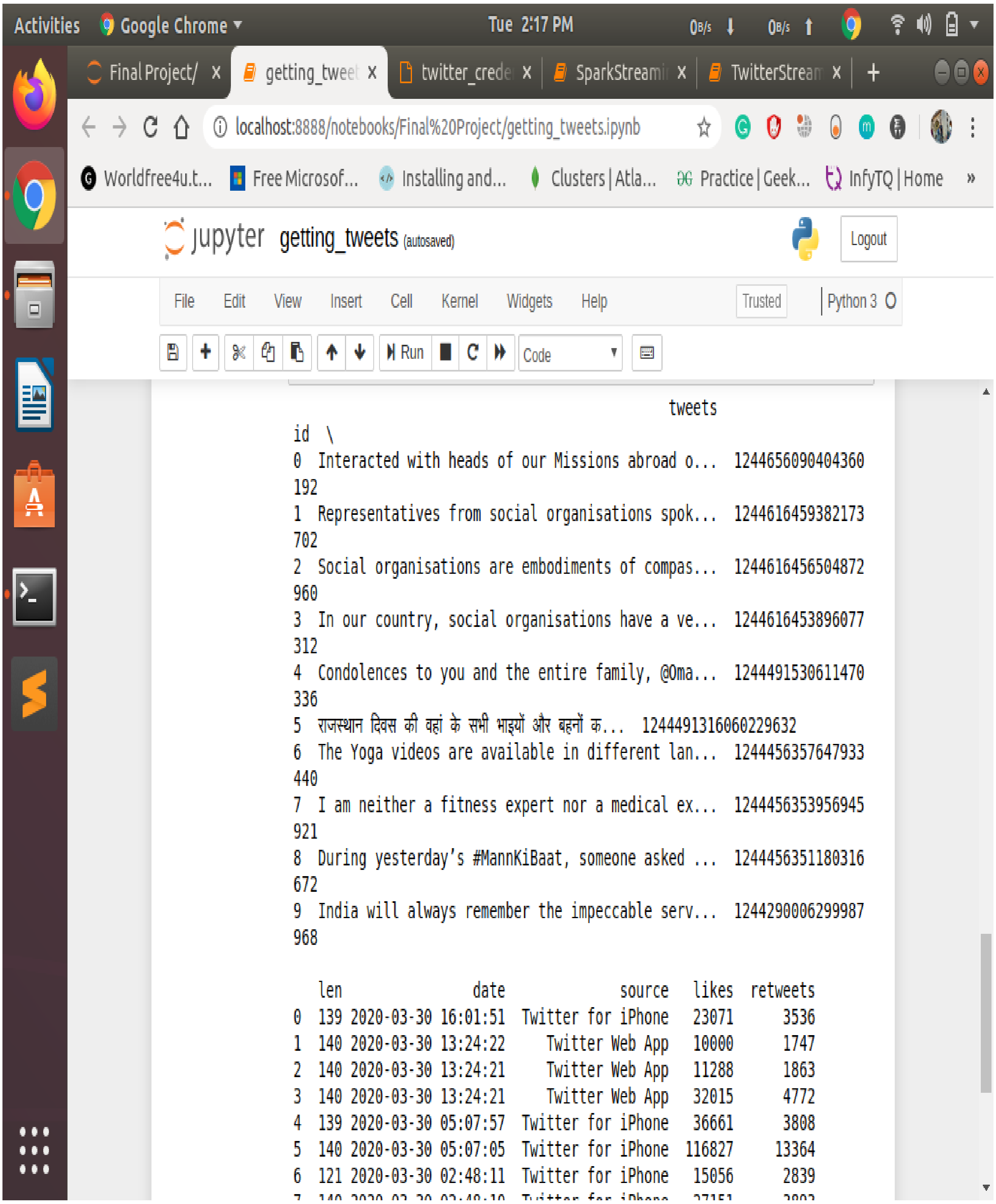
Window size =60 seconds

Lines =we are getting data into lines

Lines.pprint() //printing the data

ssc.start() //starting spark streaming

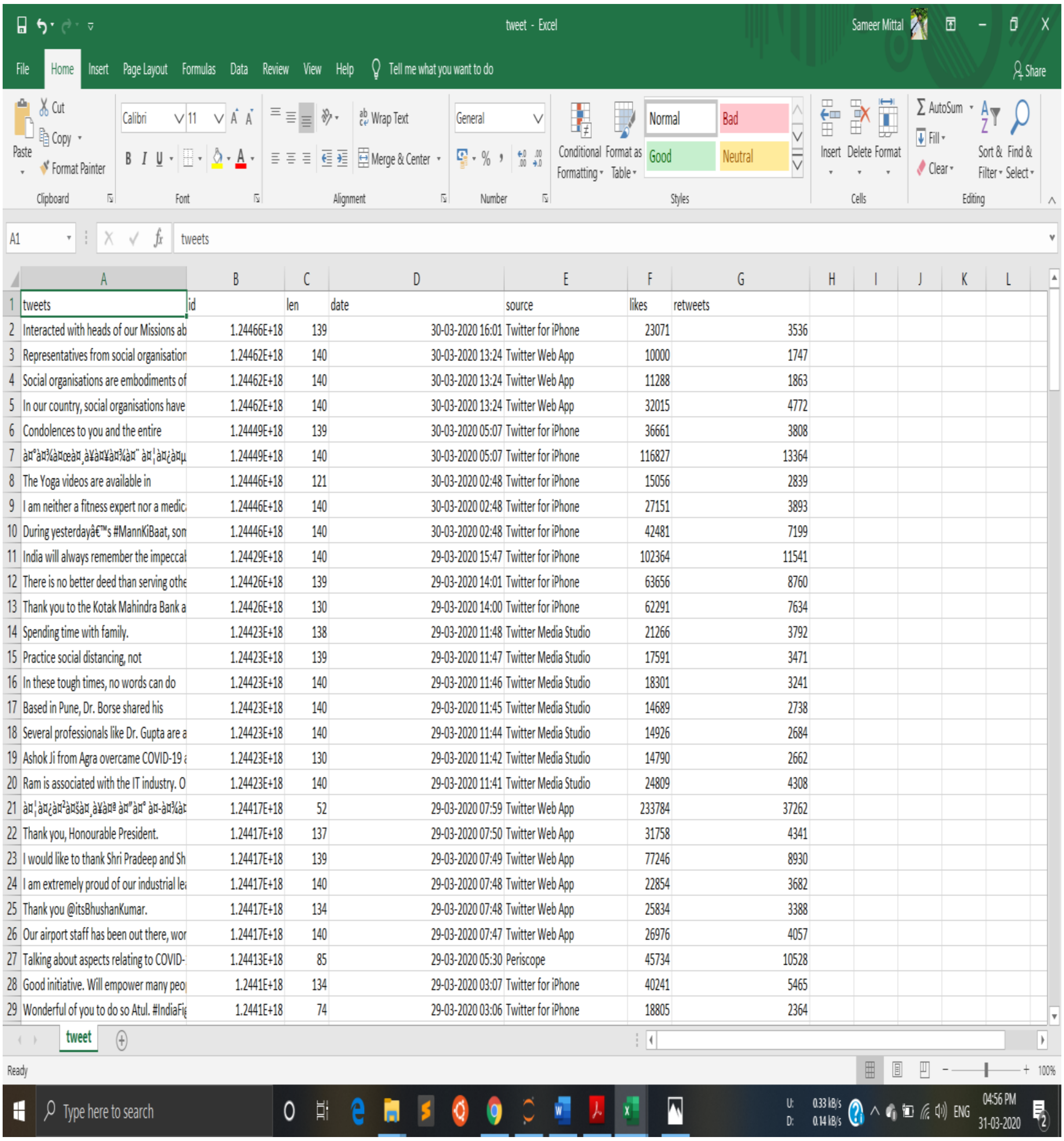
Output:- for getting output first we need to run TwitterStre aming.ipynb them SparkStreaming.ipynb



```
                                              tweets
id  \
0  Interacted with heads of our Missions abroad o...  1244656090404360
192
1  Representatives from social organisations spok...  1244616459382173
702
2  Social organisations are embodiments of compas...  1244616456504872
960
3  In our country, social organisations have a ve...  1244616453896077
312
4  Condolences to you and the entire family, @Oma...  1244491530611470
336
5  राजस्थान दिवस की वहां के सभी भाइयों और बहनों क...  1244491316060229632
6  The Yoga videos are available in different lan...  1244456357647933
440
7  I am neither a fitness expert nor a medical ex...  1244456353956945
921
8  During yesterday’s #MannKiBaat, someone asked ...  1244456351180316
672
9  India will always remember the impeccable serv...  1244290006299987
968

  len                date              source   likes  retweets
0  139 2020-03-30 16:01:51  Twitter for iPhone   23071      3536
1  140 2020-03-30 13:24:22     Twitter Web App   10000      1747
2  140 2020-03-30 13:24:21     Twitter Web App   11288      1863
3  140 2020-03-30 13:24:21     Twitter Web App   32015      4772
4  139 2020-03-30 05:07:57  Twitter for iPhone   36661      3808
5  140 2020-03-30 05:07:05  Twitter for iPhone  116827     13364
6  121 2020-03-30 02:48:11  Twitter for iPhone   15056      2839
7  140 2020-03-30 02:48:10  Twitter for iPhone   27151      [illegible]
```

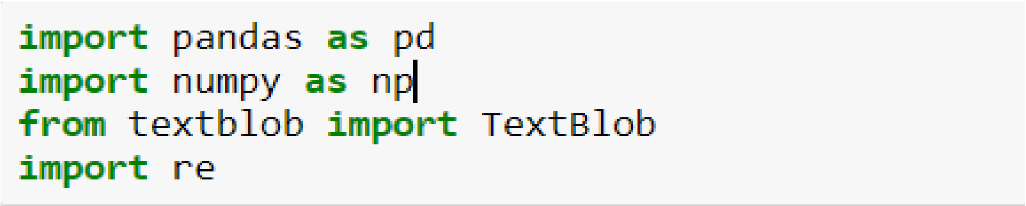
csv file



	A	B	C	D	E	F	G
1	tweets	id	len	date	source	likes	retweets
2	Interacted with heads of our Missions ab	1.24466E+18	139	30-03-2020 16:01	Twitter for iPhone	23071	3536
3	Representatives from social organisation	1.24462E+18	140	30-03-2020 13:24	Twitter Web App	10000	1747
4	Social organisations are embodiments of	1.24462E+18	140	30-03-2020 13:24	Twitter Web App	11288	1863
5	In our country, social organisations have	1.24462E+18	140	30-03-2020 13:24	Twitter Web App	32015	4772
6	Condolences to you and the entire	1.24449E+18	139	30-03-2020 05:07	Twitter for iPhone	36661	3808
7	à¤°à¤¾à¤œà¤¸à¥à¤¥à¤¾à¤¨ à¤¦à¤¿à¤µ	1.24449E+18	140	30-03-2020 05:07	Twitter for iPhone	116827	13364
8	The Yoga videos are available in	1.24446E+18	121	30-03-2020 02:48	Twitter for iPhone	15056	2839
9	I am neither a fitness expert nor a medic	1.24446E+18	140	30-03-2020 02:48	Twitter for iPhone	27151	3893
10	During yesterdayâ€™s #MannKiBaat, son	1.24446E+18	140	30-03-2020 02:48	Twitter for iPhone	42481	7199
11	India will always remember the impeccal	1.24429E+18	140	29-03-2020 15:47	Twitter for iPhone	102364	11541
12	There is no better deed than serving othe	1.24426E+18	139	29-03-2020 14:01	Twitter for iPhone	63656	8760
13	Thank you to the Kotak Mahindra Bank a	1.24426E+18	130	29-03-2020 14:00	Twitter for iPhone	62291	7634
14	Spending time with family.	1.24423E+18	138	29-03-2020 11:48	Twitter Media Studio	21266	3792
15	Practice social distancing, not	1.24423E+18	139	29-03-2020 11:47	Twitter Media Studio	17591	3471
16	In these tough times, no words can do	1.24423E+18	140	29-03-2020 11:46	Twitter Media Studio	18301	3241
17	Based in Pune, Dr. Borse shared his	1.24423E+18	140	29-03-2020 11:45	Twitter Media Studio	14689	2738
18	Several professionals like Dr. Gupta are a	1.24423E+18	140	29-03-2020 11:44	Twitter Media Studio	14926	2684
19	Ashok Ji from Agra overcame COVID-19 a	1.24423E+18	130	29-03-2020 11:42	Twitter Media Studio	14790	2662
20	Ram is associated with the IT industry. O	1.24423E+18	140	29-03-2020 11:41	Twitter Media Studio	24809	4308
21	à¤¦à¤¿à¤²à¤šà¤¸à¥à¤ª à¤”à¤° à¤à¤¾à¤	1.24417E+18	52	29-03-2020 07:59	Twitter Web App	233784	37262
22	Thank you, Honourable President.	1.24417E+18	137	29-03-2020 07:50	Twitter Web App	31758	4341
23	I would like to thank Shri Pradeep and Sh	1.24417E+18	139	29-03-2020 07:49	Twitter Web App	77246	8930
24	I am extremely proud of our industrial le	1.24417E+18	140	29-03-2020 07:48	Twitter Web App	22854	3682
25	Thank you @itsBhushanKumar.	1.24417E+18	134	29-03-2020 07:48	Twitter Web App	25834	3388
26	Our airport staff has been out there, wor	1.24417E+18	140	29-03-2020 07:47	Twitter Web App	26976	4057
27	Talking about aspects relating to COVID-	1.24413E+18	85	29-03-2020 05:30	Periscope	45734	10528
28	Good initiative. Will empower many peo	1.2441E+18	134	29-03-2020 03:07	Twitter for iPhone	40241	5465
29	Wonderful of you to do so Atul. #IndiaFig	1.2441E+18	74	29-03-2020 03:06	Twitter for iPhone	18805	2364

4.3 SentimentAnalysis.ipynb

Now I get the csv file with tweets but for sentiment Analysis the data is inappropriate first I need to clean the data then perform sentiment anlysis on that data

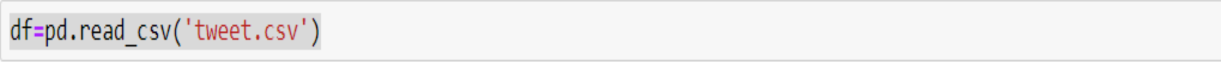


```
import pandas as pd
import numpy as np
from textblob import TextBlob
import re
```

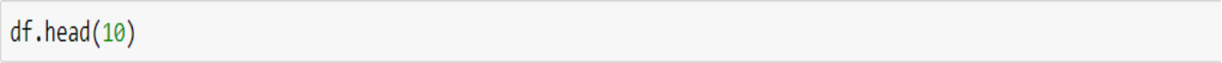
Pandas for dataframe and numpy for analytics

Textblob we use for sentiment analysis

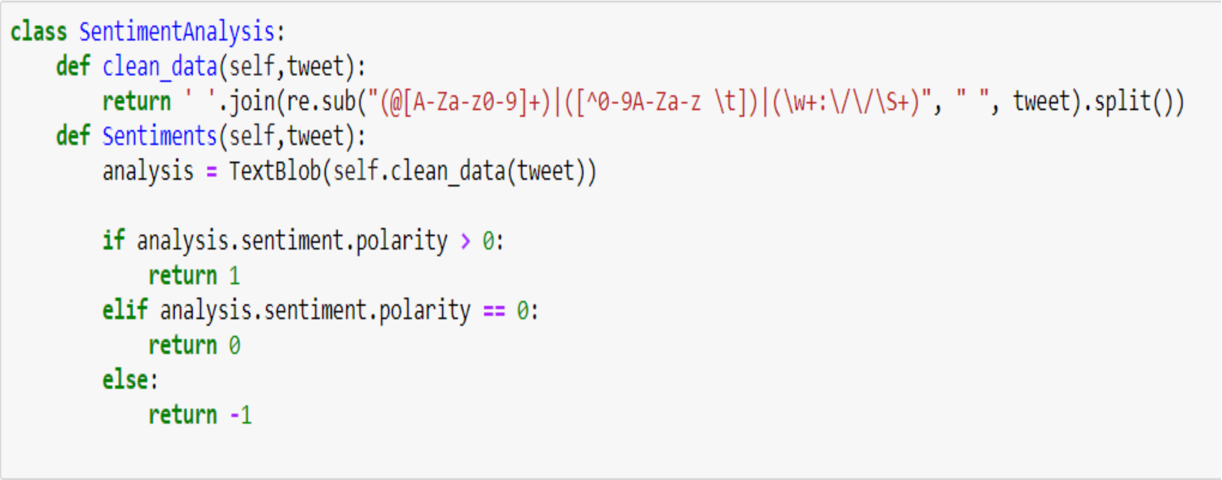
Re is regular expression this helps to clean the data



```
df=pd.read_csv('tweet.csv')
```



```
df.head(10)
```



```
class SentimentAnalysis:
    def clean_data(self,tweet):
        return ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/\S+)", " ", tweet).split())
    def Sentiments(self,tweet):
        analysis = TextBlob(self.clean_data(tweet))

        if analysis.sentiment.polarity > 0:
            return 1
        elif analysis.sentiment.polarity == 0:
            return 0
        else:
            return -1
```

First we are reading csv file with the help of **.read_csv** method

Class SentimentAnalysis helps to clean and find the sentiment of the tweet

def clean_data(self,tweet):

return ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/\S+)", " ", tweet).split())

it will return cleaning all the data by removing any special character converting upper case to lower.

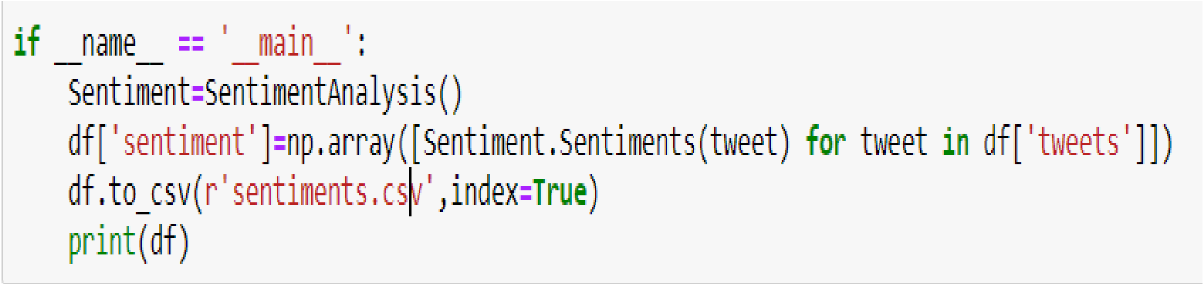
```
def Sentiments(self,tweet):
        analysis = TextBlob(self.clean_data(tweet))
         if analysis.sentiment.polarity > 0:
            return 1
        elif analysis.sentiment.polarity == 0:
            return 0
        else:
            return -1
```

Now this function is responsible for analysing sentiments with respect to tweet text it has a special method called polarity which will check the sentiments

If polarity>0 then tweet is 1 or **positive**

If polarity==0 then tweet is 0 or **neutral**

If polarity<0 then tweet is -1 or **negative**



```
if __name__ == '__main__':
    Sentiment=SentimentAnalysis()
    df['sentiment']=np.array([Sentiment.Sentiments(tweet) for tweet in df['tweets']])
    df.to_csv(r'sentiments.csv',index=True)
    print(df)
```

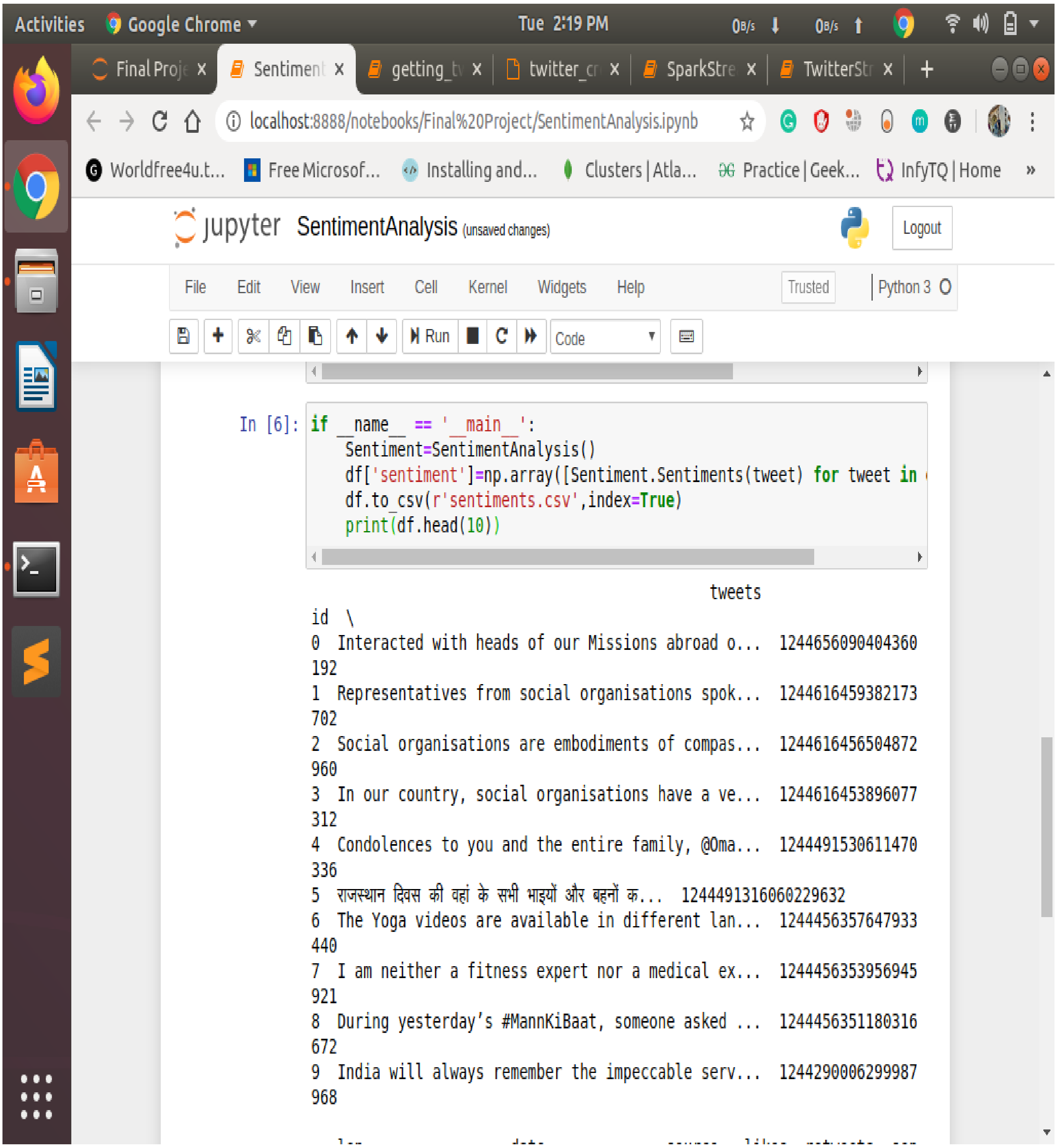
From this part we call our class.

df['sentiment']=np.array([Sentiment.Sentiments(tweet) for tweet in df['tweets']])

This will create a new column in dataframe called sentiments based on data getting from the class

In this we are passing **tweets** column

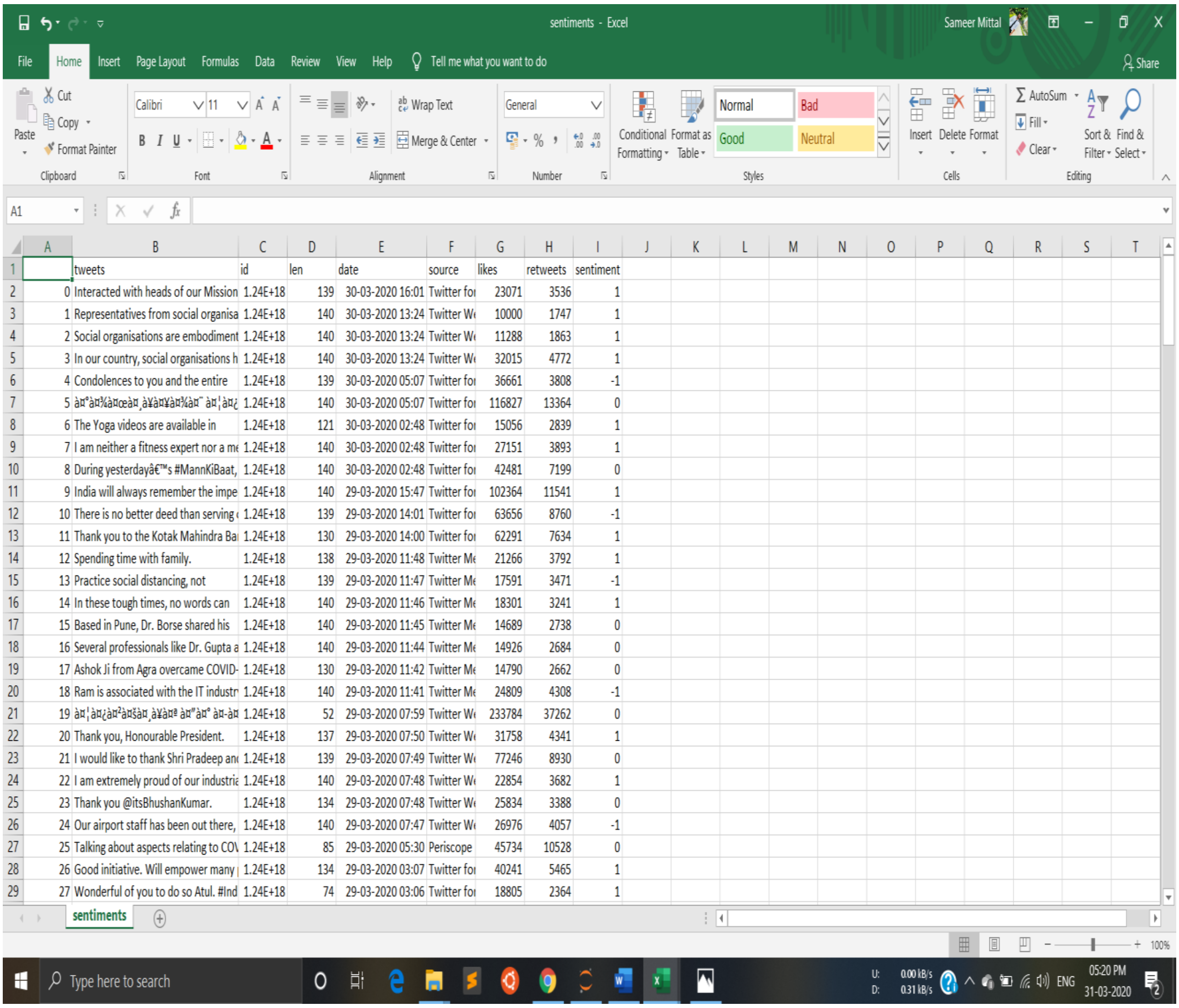
Output :-



```
In [6]: if __name__ == '__main__':
            Sentiment=SentimentAnalysis()
            df['sentiment']=np.array([Sentiment.Sentiments(tweet) for tweet in
            df.to_csv(r'sentiments.csv',index=True)
            print(df.head(10))
```

```
                                              tweets
id  \
0  Interacted with heads of our Missions abroad o...  1244656090404360
192
1  Representatives from social organisations spok...  1244616459382173
702
2  Social organisations are embodiments of compas...  1244616456504872
960
3  In our country, social organisations have a ve...  1244616453896077
312
4  Condolences to you and the entire family, @Oma...  1244491530611470
336
5  राजस्थान दिवस की वहां के सभी भाइयों और बहनों क...  1244491316060229632
6  The Yoga videos are available in different lan...  1244456357647933
440
7  I am neither a fitness expert nor a medical ex...  1244456353956945
921
8  During yesterday’s #MannKiBaat, someone asked ...  1244456351180316
672
9  India will always remember the impeccable serv...  1244290006299987
968
```

Csv file



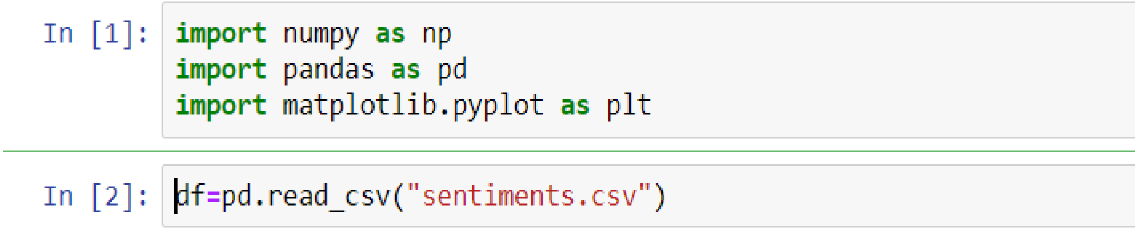
	A	B	C	D	E	F	G	H	I
1		tweets	id	len	date	source	likes	retweets	sentiment
2	0	Interacted with heads of our Mission	1.24E+18	139	30-03-2020 16:01	Twitter foi	23071	3536	1
3	1	Representatives from social organisa	1.24E+18	140	30-03-2020 13:24	Twitter W	10000	1747	1
4	2	Social organisations are embodiment	1.24E+18	140	30-03-2020 13:24	Twitter W	11288	1863	1
5	3	In our country, social organisations h	1.24E+18	140	30-03-2020 13:24	Twitter W	32015	4772	1
6	4	Condolences to you and the entire	1.24E+18	139	30-03-2020 05:07	Twitter foi	36661	3808	-1
7	5	à¤°à¤¾à¤œà¤¸à¥à¤¥à¤¾à¤¨ à¤¦à¤¿	1.24E+18	140	30-03-2020 05:07	Twitter foi	116827	13364	0
8	6	The Yoga videos are available in	1.24E+18	121	30-03-2020 02:48	Twitter foi	15056	2839	1
9	7	I am neither a fitness expert nor a me	1.24E+18	140	30-03-2020 02:48	Twitter foi	27151	3893	1
10	8	During yesterdayâ€™s #MannKiBaat,	1.24E+18	140	30-03-2020 02:48	Twitter foi	42481	7199	0
11	9	India will always remember the impe	1.24E+18	140	29-03-2020 15:47	Twitter foi	102364	11541	1
12	10	There is no better deed than serving	1.24E+18	139	29-03-2020 14:01	Twitter foi	63656	8760	-1
13	11	Thank you to the Kotak Mahindra Ba	1.24E+18	130	29-03-2020 14:00	Twitter foi	62291	7634	1
14	12	Spending time with family.	1.24E+18	138	29-03-2020 11:48	Twitter Me	21266	3792	1
15	13	Practice social distancing, not	1.24E+18	139	29-03-2020 11:47	Twitter Me	17591	3471	-1
16	14	In these tough times, no words can	1.24E+18	140	29-03-2020 11:46	Twitter Me	18301	3241	1
17	15	Based in Pune, Dr. Borse shared his	1.24E+18	140	29-03-2020 11:45	Twitter Me	14689	2738	0
18	16	Several professionals like Dr. Gupta a	1.24E+18	140	29-03-2020 11:44	Twitter Me	14926	2684	0
19	17	Ashok Ji from Agra overcame COVID-	1.24E+18	130	29-03-2020 11:42	Twitter Me	14790	2662	0
20	18	Ram is associated with the IT industr	1.24E+18	140	29-03-2020 11:41	Twitter Me	24809	4308	-1
21	19	à¤¦à¤¿à¤²à¤šà¤¸à¥à¤ª à¤”à¤° à¤à¤	1.24E+18	52	29-03-2020 07:59	Twitter W	233784	37262	0
22	20	Thank you, Honourable President.	1.24E+18	137	29-03-2020 07:50	Twitter W	31758	4341	1
23	21	I would like to thank Shri Pradeep an	1.24E+18	139	29-03-2020 07:49	Twitter W	77246	8930	0
24	22	I am extremely proud of our industria	1.24E+18	140	29-03-2020 07:48	Twitter W	22854	3682	1
25	23	Thank you @itsBhushanKumar.	1.24E+18	134	29-03-2020 07:48	Twitter W	25834	3388	0
26	24	Our airport staff has been out there,	1.24E+18	140	29-03-2020 07:47	Twitter W	26976	4057	-1
27	25	Talking about aspects relating to CO\	1.24E+18	85	29-03-2020 05:30	Periscope	45734	10528	0
28	26	Good initiative. Will empower many	1.24E+18	134	29-03-2020 03:07	Twitter foi	40241	5465	1
29	27	Wonderful of you to do so Atul. #Ind	1.24E+18	74	29-03-2020 03:06	Twitter foi	18805	2364	1

New Column Created with a sentiment column having 0,1,-1

According to sentiments of the tweets

4.4 Analytics.ipynb

In this notebook we are performing analytics on data like max likes on tweets, retweets ploting time series with respect to likes and retweets.



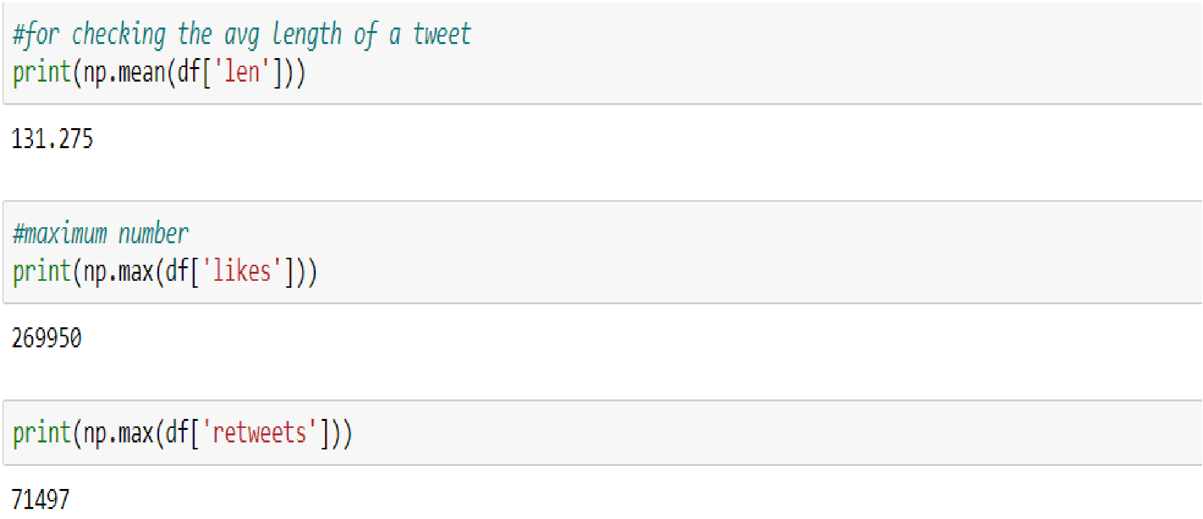
```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
```

```
In [2]: df=pd.read_csv("sentiments.csv")
```

Importing the libraries

Matplotlib is used for visualising

Reading the **sentiments.csv** file which we get from **SentimentAnalysis.ipynb**



```
#for checking the avg length of a tweet
print(np.mean(df['len']))
```

```
131.275
```

```
#maximum number
print(np.max(df['likes']))
```

```
269950
```

```
print(np.max(df['retweets']))
```

```
71497
```

In this part we are analysing :-

1. Average length of tweet text by **np.mean()**

   Note:- we have one column in our file which calculates the length of each tweet

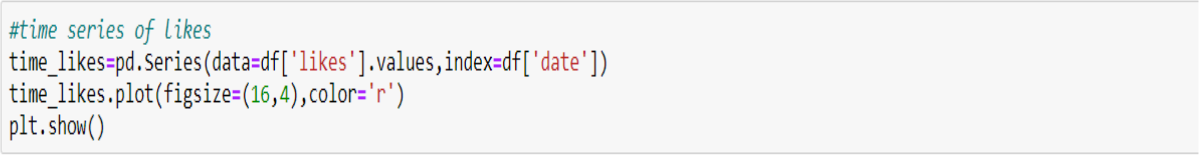
   Output avg tweet length=131.275

2. Maximum number of likes **np.max()** on likes column

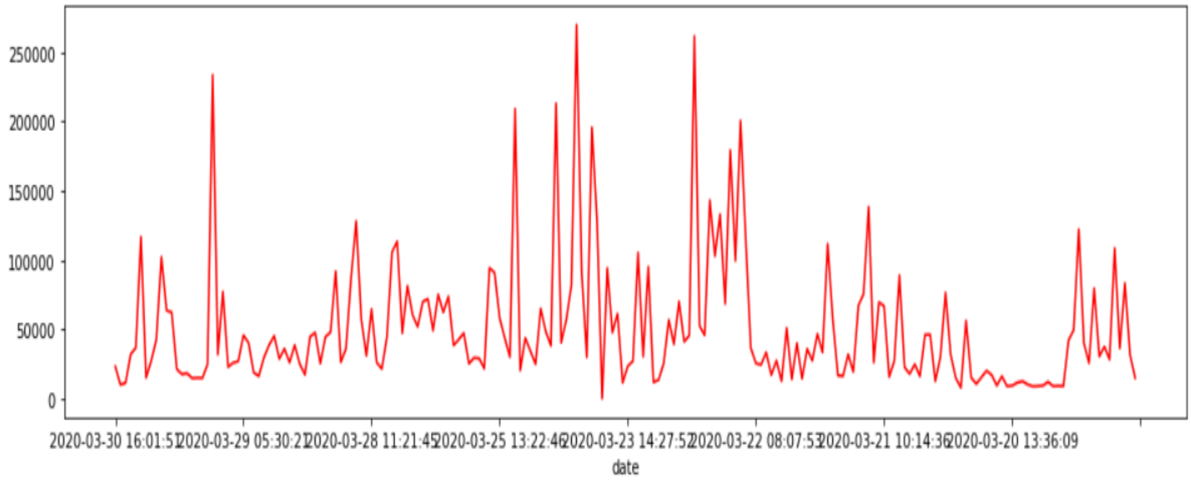
   Output max likes=269950

3. Maximum number of retweets again **np.max()** on retweet column

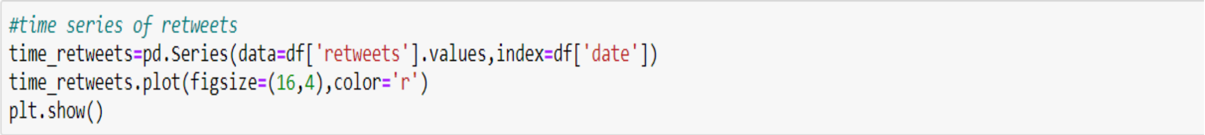
   Output max retweets=71497



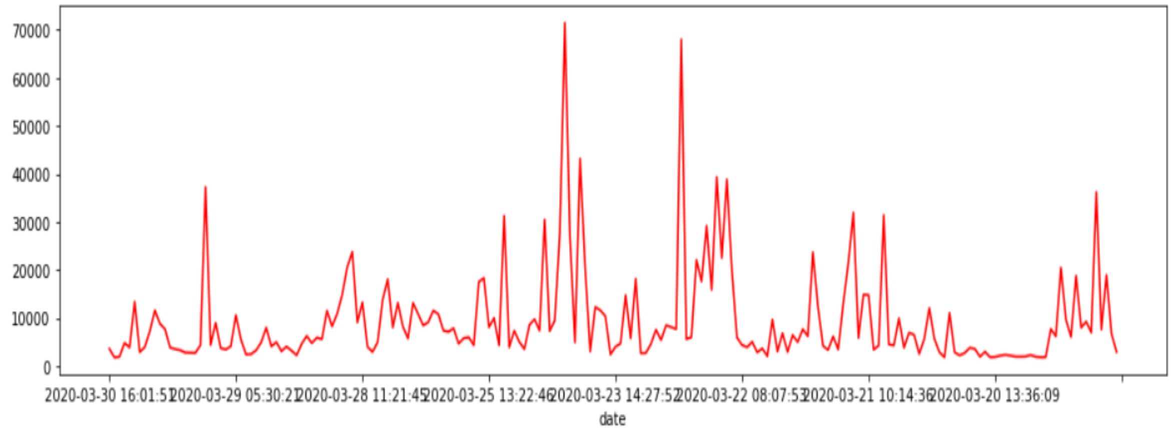
```
#time series of likes
time_likes=pd.Series(data=df['likes'].values,index=df['date'])
time_likes.plot(figsize=(16,4),color='r')
plt.show()
```



This is likes time series plot which tells number of likes with respect to date.



```
#time series of retweets
time_retweets=pd.Series(data=df['retweets'].values,index=df['date'])
time_retweets.plot(figsize=(16,4),color='r')
plt.show()
```



This is retweets time series plot which tells number of retweets with respect to date.

5. **Summary:-**

Sentiment Analysis is very vast topic by getting the sentiments of the data it can help in market research , Knowing the individual opinion about any topic. We can also do some analytics on it to know the popularity of someone.

This project tries to do a part of it by using one of the modern technology of big data i.e spark by using it twiiter data is streamed and sentiment analysis is performed on it. Later this will help to make a classification model in machine learning to make this project more robust. The main goal of this project is to analyse the sentiments but in future this project will analyse the sentiments while the streaming itself. To save more computational time and that makes the real use of Big Data.