**RTD Mini Project – Group 5**

**Detailed Design :**

There are two files - **Real\_Time\_Analysis\_Twitter.ipynb** and **Spark\_Streaming.ipynb** where the former file contains the **Stream Listener** and **Transmitter** code which transmits the twitter data through the created socket, the latter file contains the required methods for analysing live twitter data. The created classes and methods are listed below -

* Class TweetsListener - The class that contains method to create the socket for streaming data.
* Class Transmitter - The class that contains sendData() method to transmit the streaming data via the binded socket.
* get\_json(myjson) - The method that returns the JSON object of the passed argument
* updateTotalCount(currentCount, countState) - The method that sums up the current count with the total count
* lang(lang) - The method that returns the complete language name from the existing language dictionary.
* country(country): - The method that parses the “place” field of the tweet JSON and returns the country name.
* tokenize(text): - The method that removes the stop words , special characters from the tweet text and tokenizes the words which would be useful in sentimental analysis.
* get\_tweet\_sentiment(tokens): - The method that gets the sentiment of the tokens using the polarity of the TextBlob object
* createGraph(): - The method that generates graph dynamically over the data collected from the streaming twitter data.
* Threading concept is used to refresh the graph dynamically over the live streaming data collecting for every 30 seconds.

**Assumptions :**

* The tweets that has valid value for “place” field is considered for Geo analysis and Geo Plot is created. Also, the coordinates for most of the tweets were found to be null , hence “coordinates” are not used for plotting but they are aggregated and printed in the output.

**Steps to Execute :**

Please follow below steps sequentially to see the outputs,

*Step 1:* Open File named “Real\_Time\_Analysis\_Twitter”

Change the Host and Port.

*Step 2:* Now open File named “Spark\_Streaming\_Final”

Change the Host and Port to the one in step 1.

*Step 3:* Run the “Real\_Time\_Analysis\_Twitter” file first after clearing output.

Ensure you got the output as “Listening on port: YOUR PORT NUMBER”

Step 4: If possible, please supply your Twitter secret access tokens

*Step 5:* Run the “Spark\_Streaming\_Final” file after clearing the output.

Wait for ~30 to 50 seconds (socket window).

*Step 6:* On successful execution of the code, the aggregated output of the analysis will be displayed as mentioned in the below section.

If individual output is required, then uncomment the pprint() part for each dstream and execute. Also, the main Dstream object can be viewed by uncommenting the “dstream\_tweets.pprint()” line.

**Output Format:**

**Influential people in twitter**

You should see the output as,

Influential People: Follower's Count:

If you want to see the top 20, uncomment the influential.pprint(20)

You should also see a barplot of ‘User vs Followers Count'

You should see Output for,

**Trending Topics in twitter right now**

As,

Top Trend: Count:

If you want to see the top 10, uncomment the trends.pprint(10)

You should also see a barplot of ‘Twitter Trend Plot’

And for the word occurring most of the times, you should see below,

Top Word: Count:

Then, you should see Output for,

**Top tweeting user**

As,

Top Tweeting User: Tweet Count:

If you want to see the top 50, uncomment the top\_tweeting\_user.pprint(50)

You should also see a barplot of ‘User vs Tweets Count’

Then, you should see Output for,

**Language based analysis**

As,

Popular Language: Count:

If you want to see the top 7, uncomment the language.pprint(7)

You should also see a barplot of ‘Most Tweeted Language’

For,

**Geo Location with high tweet volume:**

You should see Output as,

Top Geo: Tweet Volume:

Note: In the above you’ll see Outputs for Sentiment Analysis and Popular tweets but the tweets are not of iphone.

*Step 6:* For Popular tweets and Sentiment Analysis about iphone,

Go back to “Real\_Time\_Analysis\_Twitter” file and in both the sendData() method, **comment the line - “twitter\_stream.filter(track=['a','is','the','I','are','you'])” and uncomment the line - “twitter\_stream.filter(track=['iphone'])”** to track iphone tweets.

*Step 7:* Clear the Output of both the files.

Run the “Real\_Time\_Analysis\_Twitter” file first.

Ensure you got the output as “Listening on port: YOUR PORT NUMBER”

Run the “Spark\_Streaming\_Final” file then.

Along with the above responses, you’ll now see below for the iphone tweets.

**Sentiment Analysis:**

Top Sentiment: Count:

You should also see a barplot of ‘Tweets Sentiment’

For,

**Most popular tweets:**

You should see Output as,

Most Retweeted: Count: