**PRINCIPLES OF BIG DATA MANAGEMENT**

**PHASE 2 REPORT**

**Team Member:**

**Jagadheesh Maroju (16297112)**

**Sai Prasad Raju Bollepalli (16301708)**

**Girija Dasoju (16302148)**

**Link**

Tweets data:

[**https://drive.google.com/open?id=18jn55iiWOOTp\_1r\_dPYG3PGkuSUxWnTp**](https://drive.google.com/open?id=18jn55iiWOOTp_1r_dPYG3PGkuSUxWnTp)

**OBJECTIVE**

* Stream tweets from Twitter (we collected more than 300,000 tweets).
* Write at least 10 analytic queries to explore and understand the collected tweets.
* Create Visualizations on the results.
* Create a web page to view the visualizations.

**SOFTWARES PLANNING TO USE**

* Apache Spark
  + Spark Streaming to collect tweets.
  + Spark SQL to store and execute queries.
* Matplot for visualization
* HTML,CSS and Javascript for web application

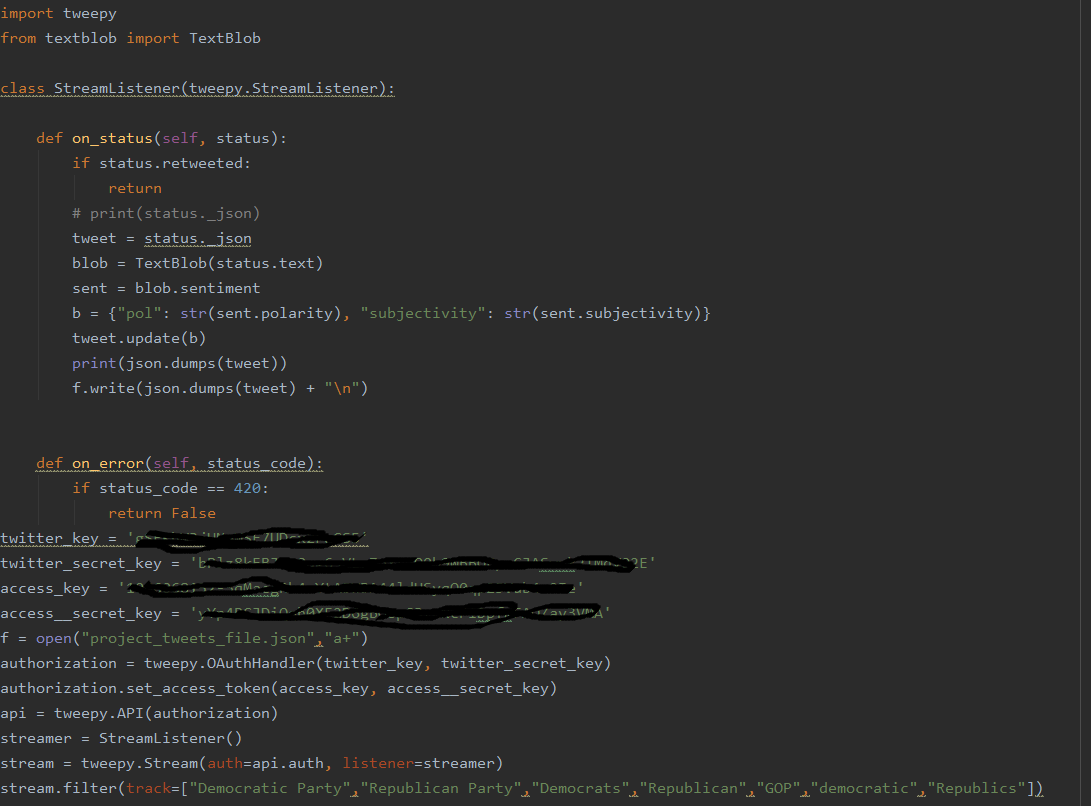
**IMPLEMENTATION**

* Initially, we wrote a python program to stream the tweets and save them into JSON format which will have details of the tweets.
* The extracted JSON tweets are persisted into the Apache SparkSQL in the form of Views.
* We used Pycharm IDE to write queries and visualize the outputs using tables and charts.
* The web application is developed using HTML, CSS and Java script.

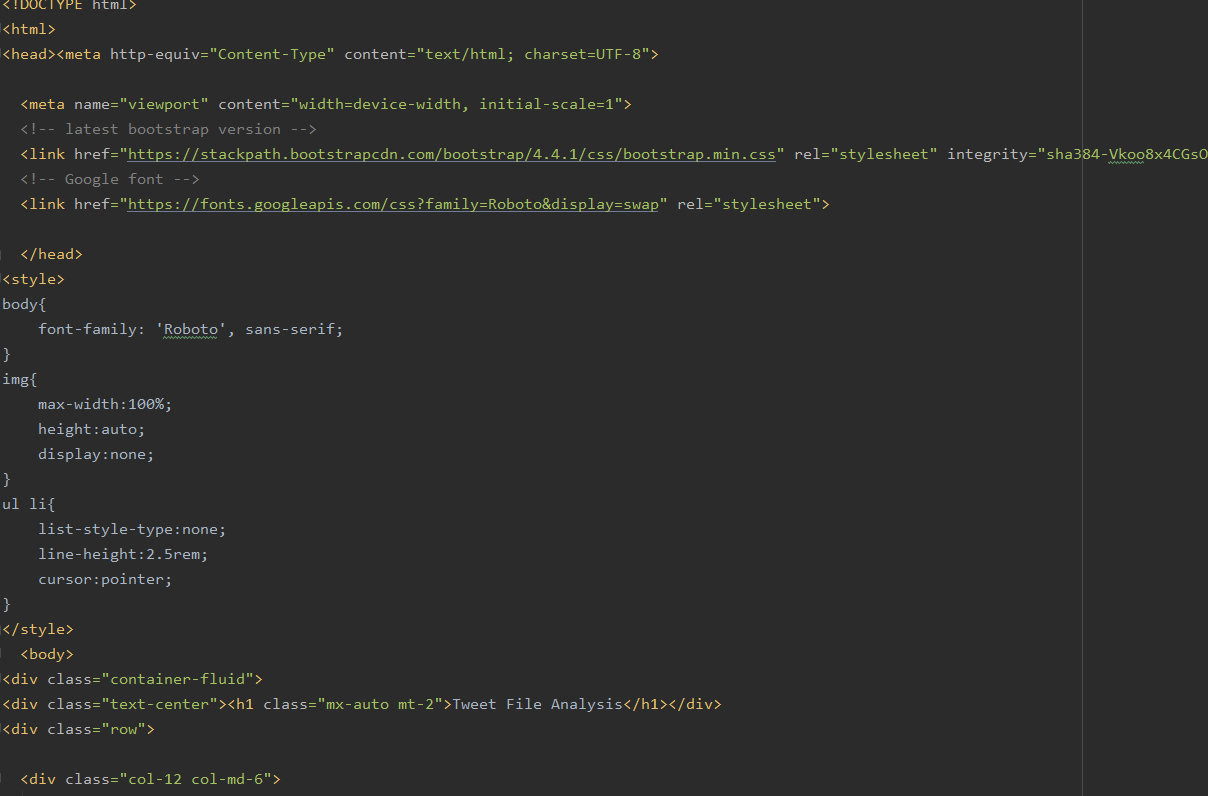
Libraries used:

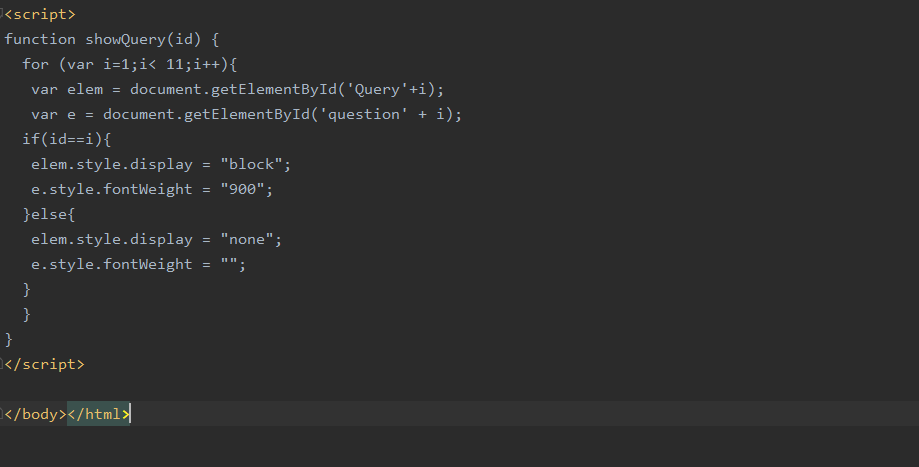
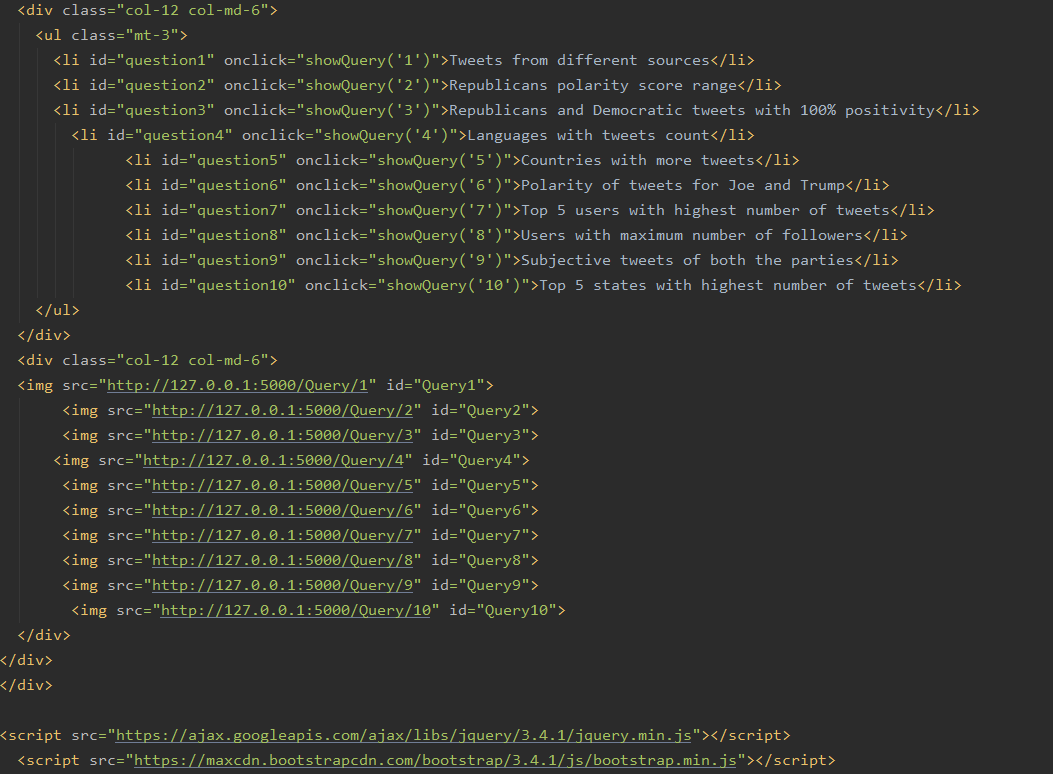
* Matplot for visualizations
* TextBlob to get sentimental score
* Flask for web application

Code for Collecting the Tweets:

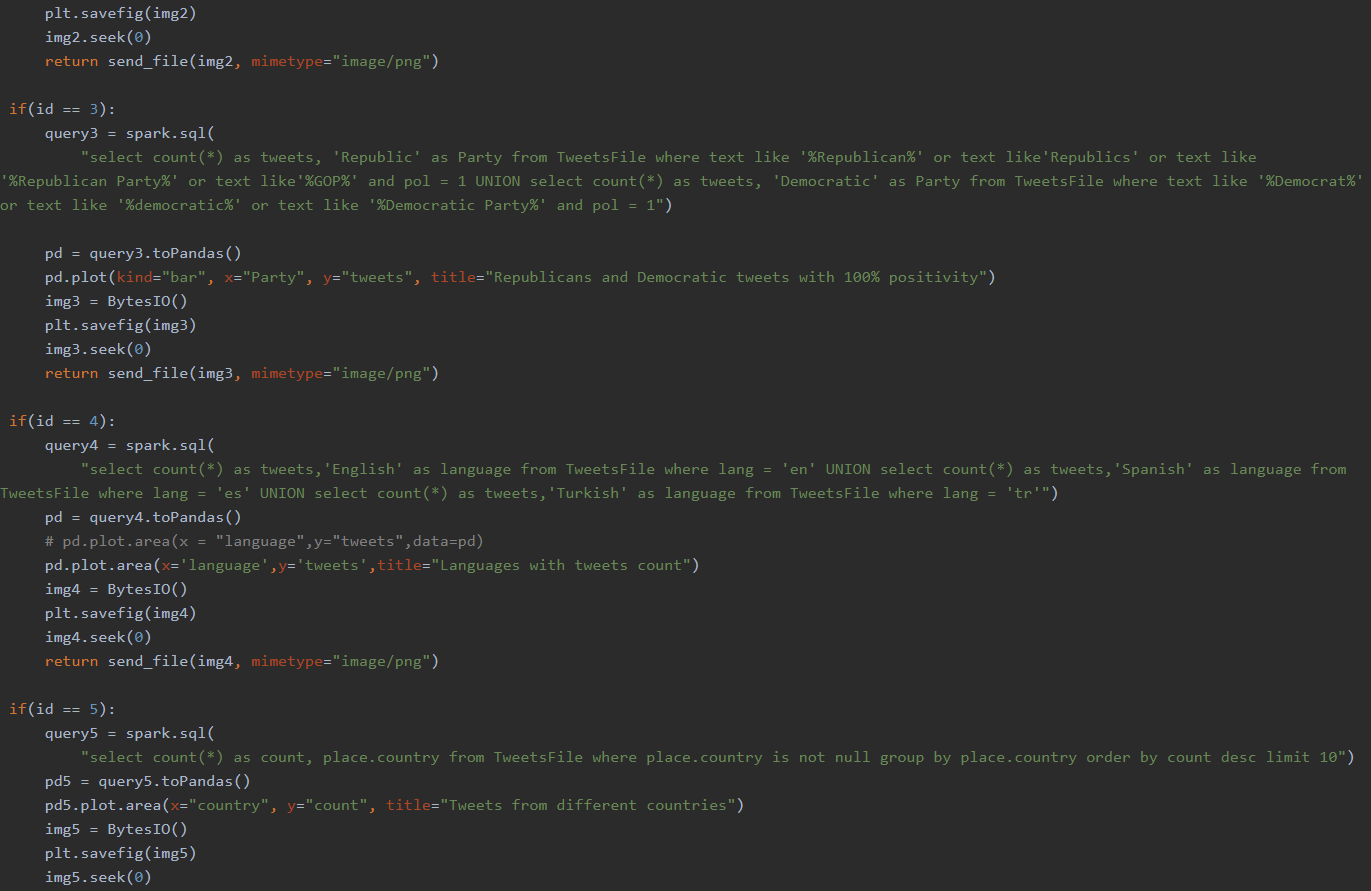
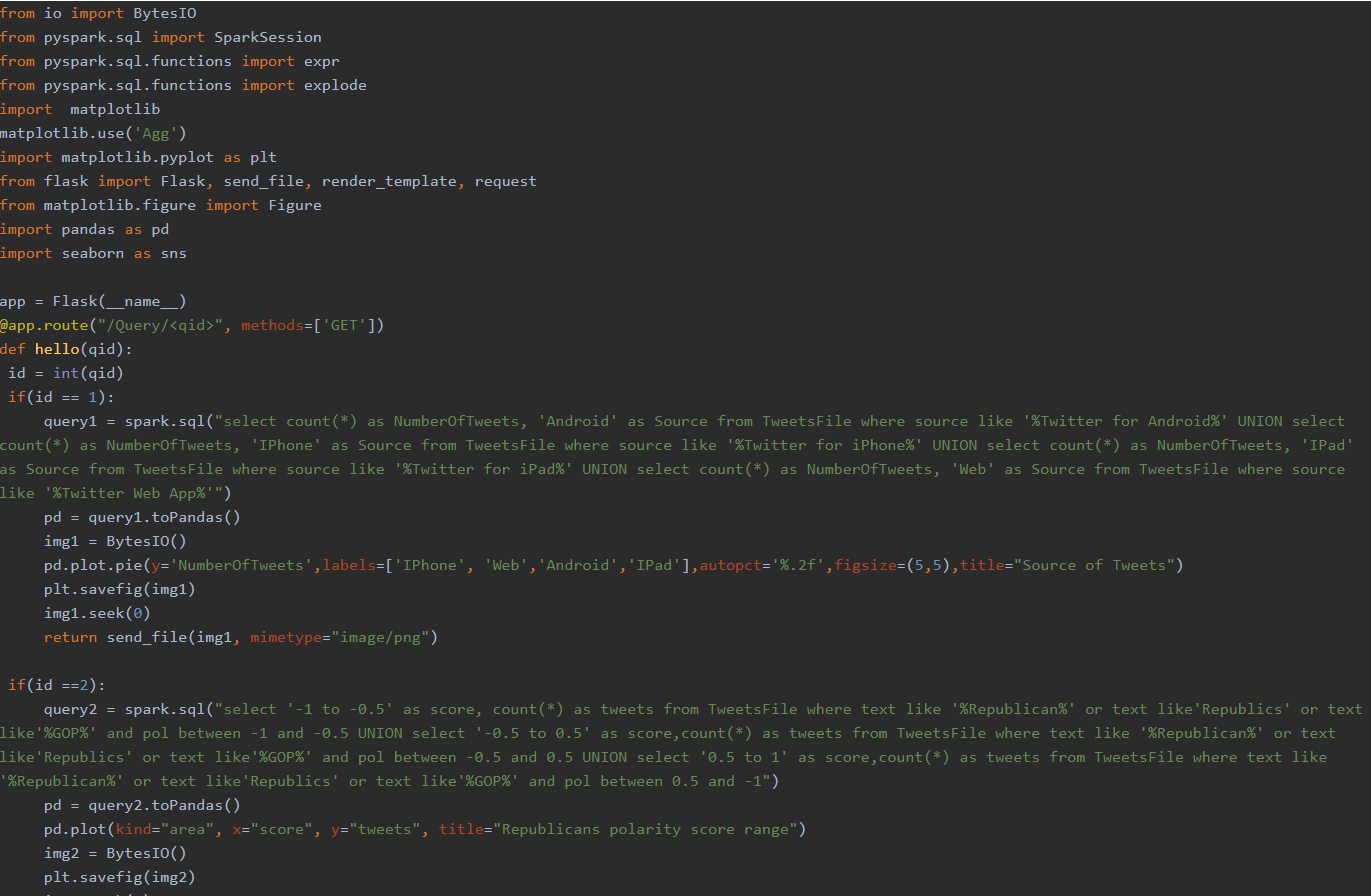


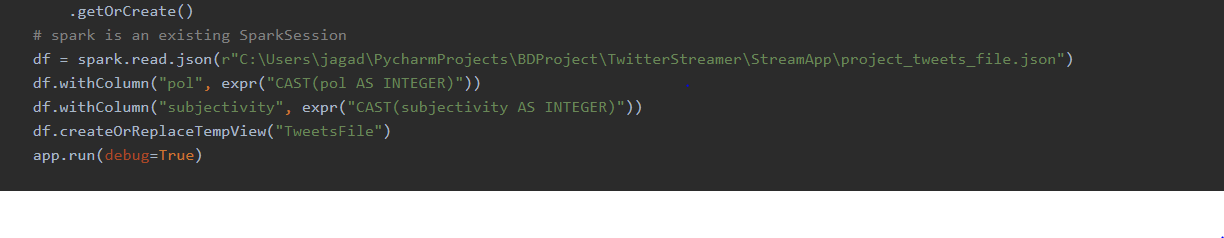
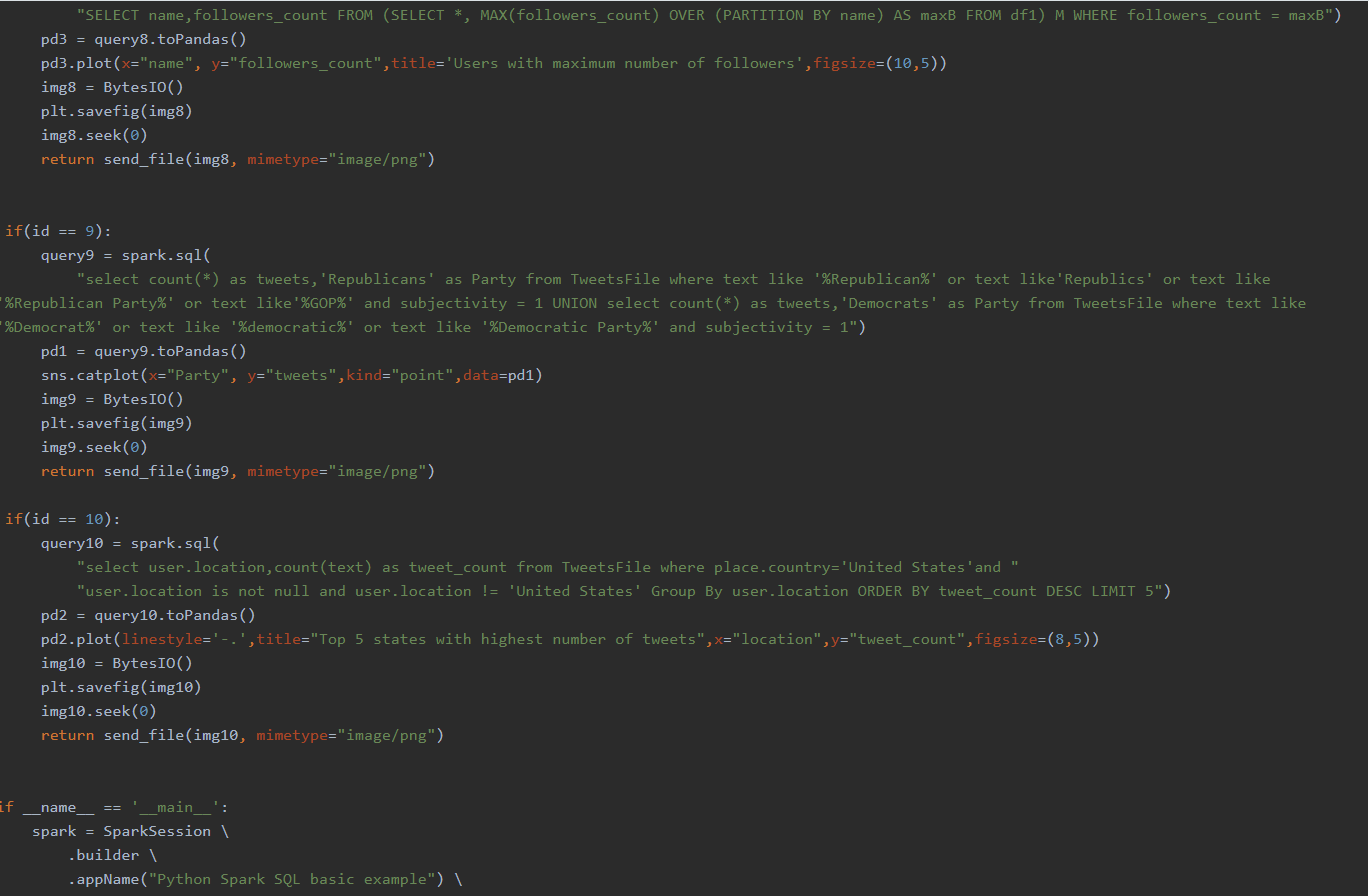
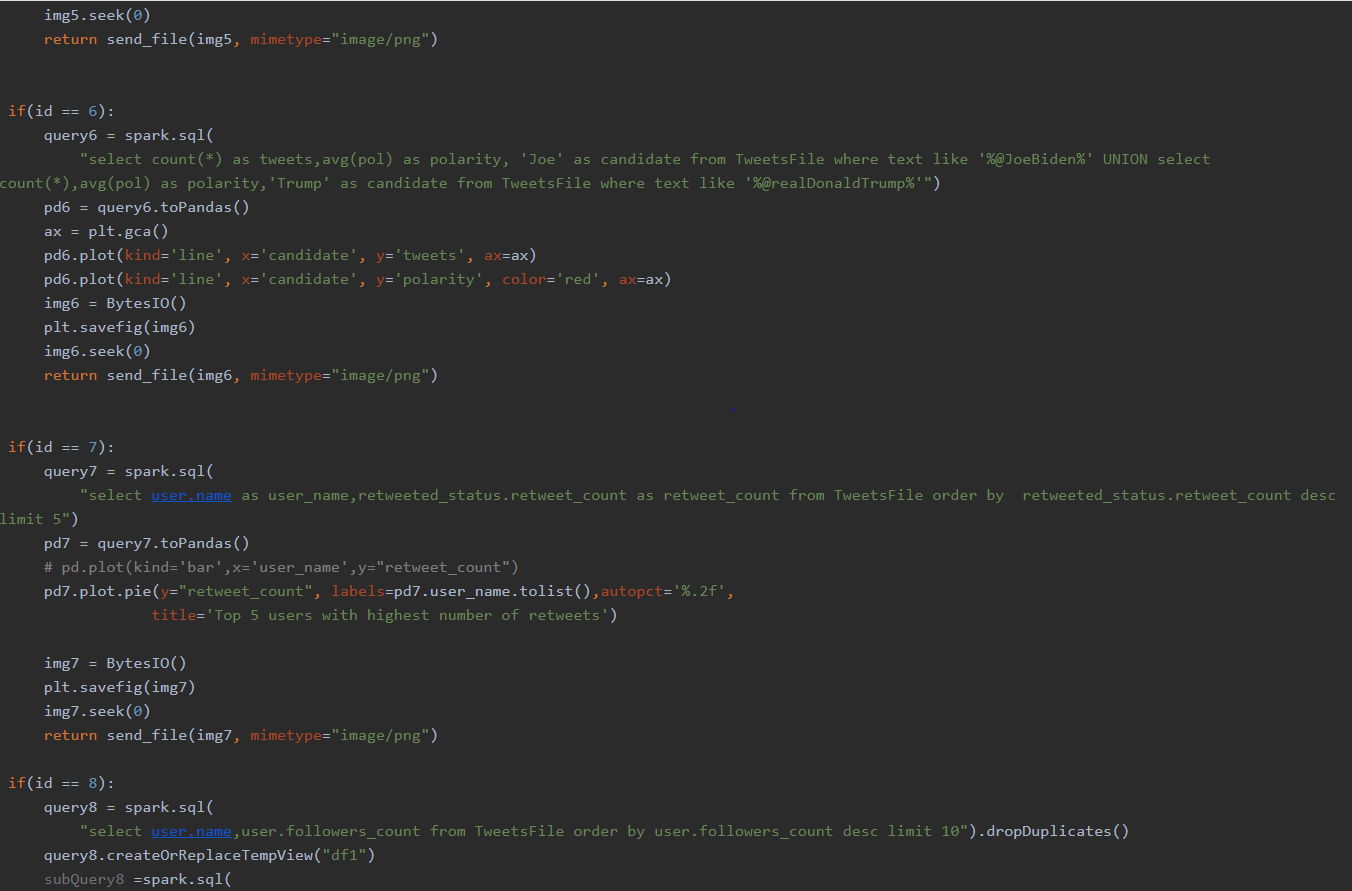
HTML and JavaScript code for UI:





Code For 10 Queries:



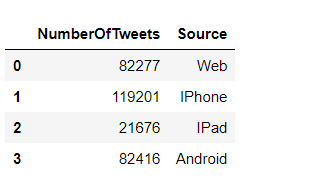


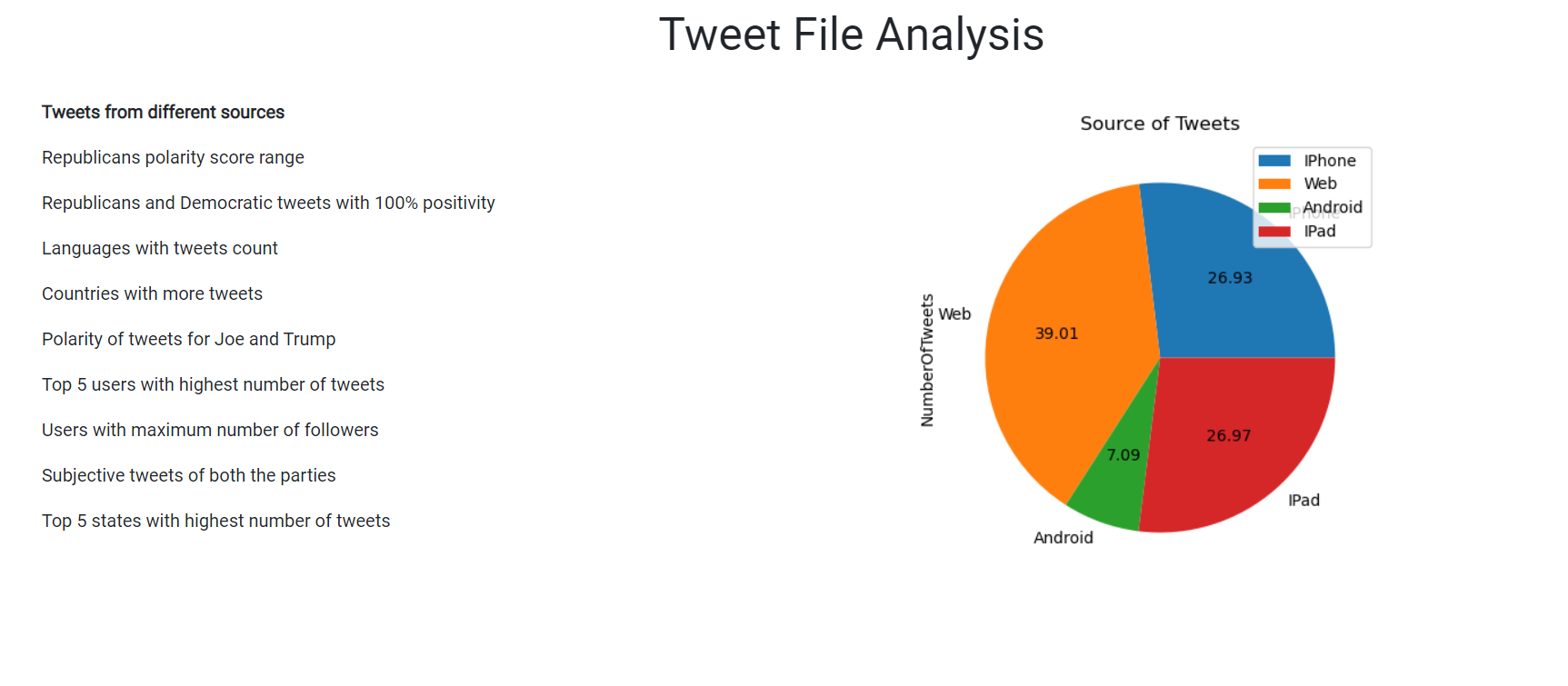
**Query Execution and Results:**

We have executed and visualized totally 10 queries.

1). Tweets originated from Different sources using Pie Chart.

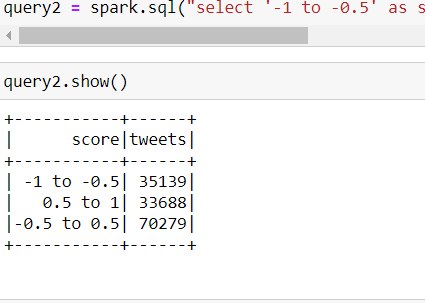
Query result in table form :

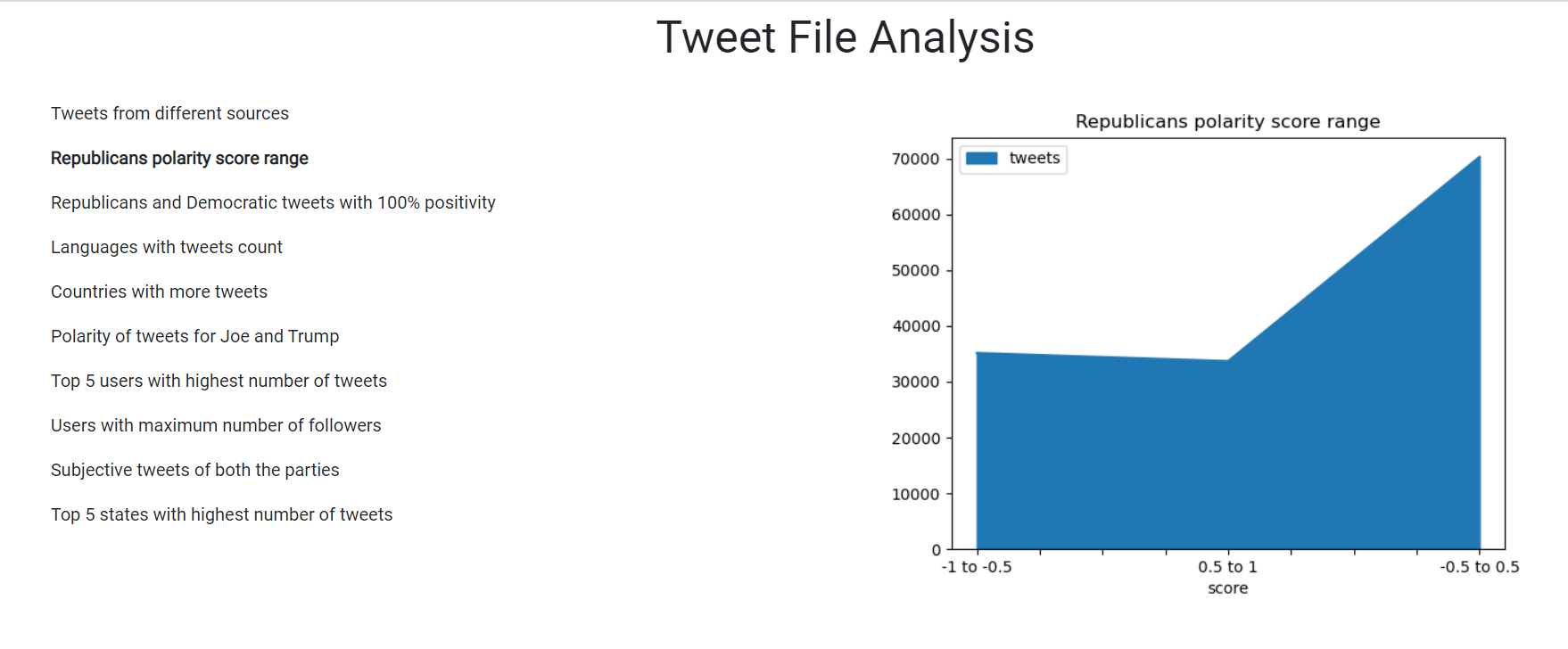




2) Query to display the Republicans with different ranges of polarity score using “Area plot”.

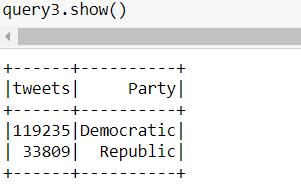
Result:

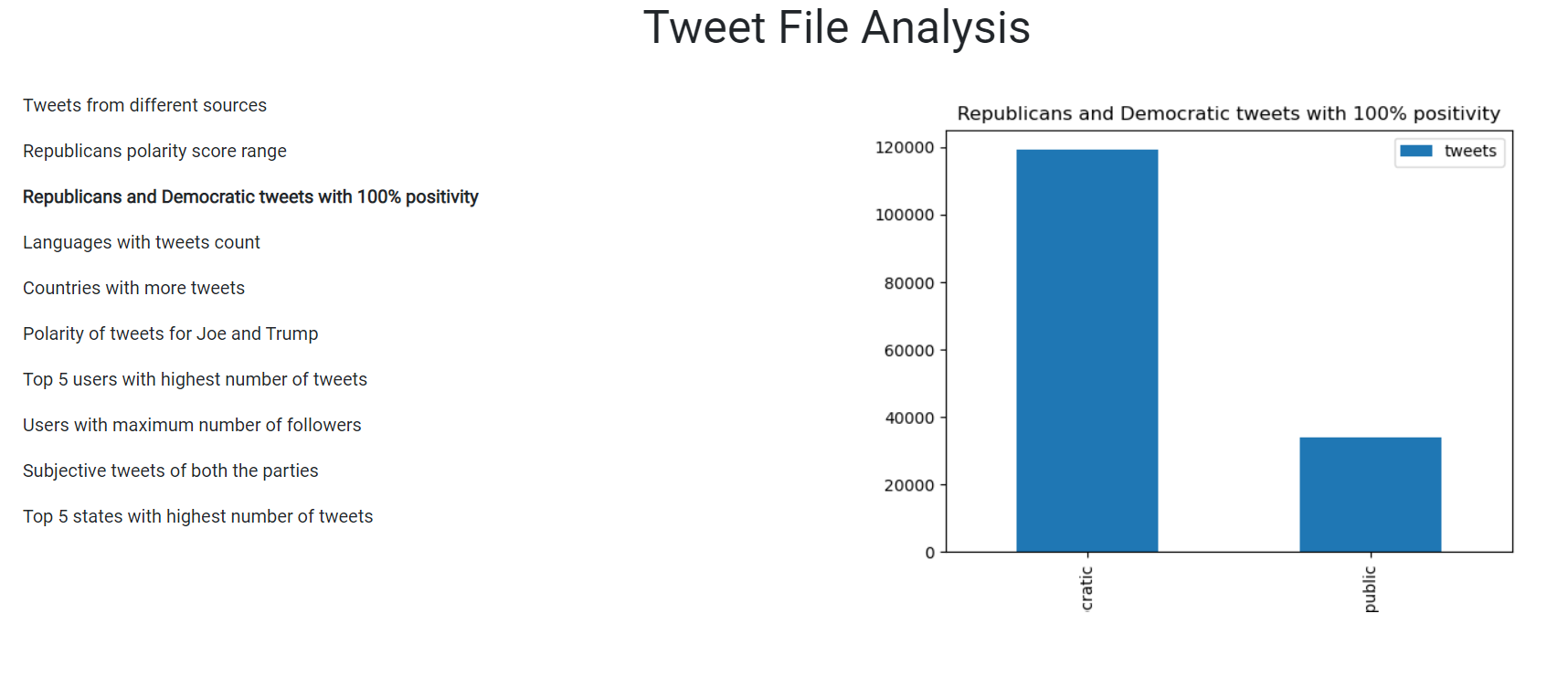




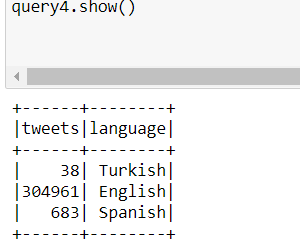
3). Tweets with 100% positivity( i.e polarity = 1) among Republicans and Democratic’s using

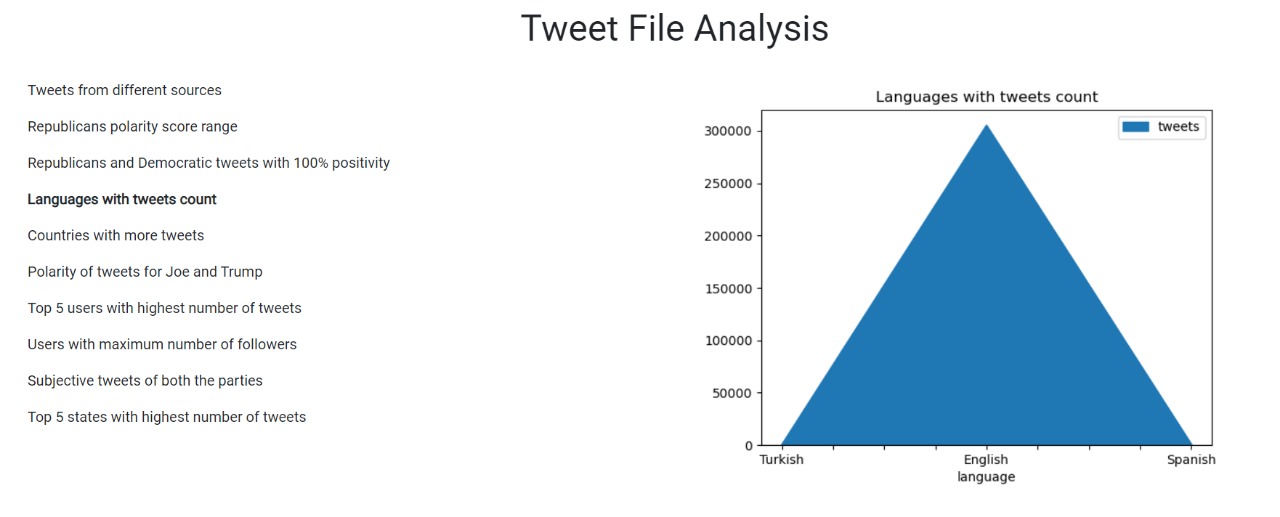
“Bar Chart”



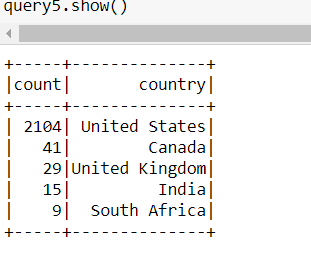


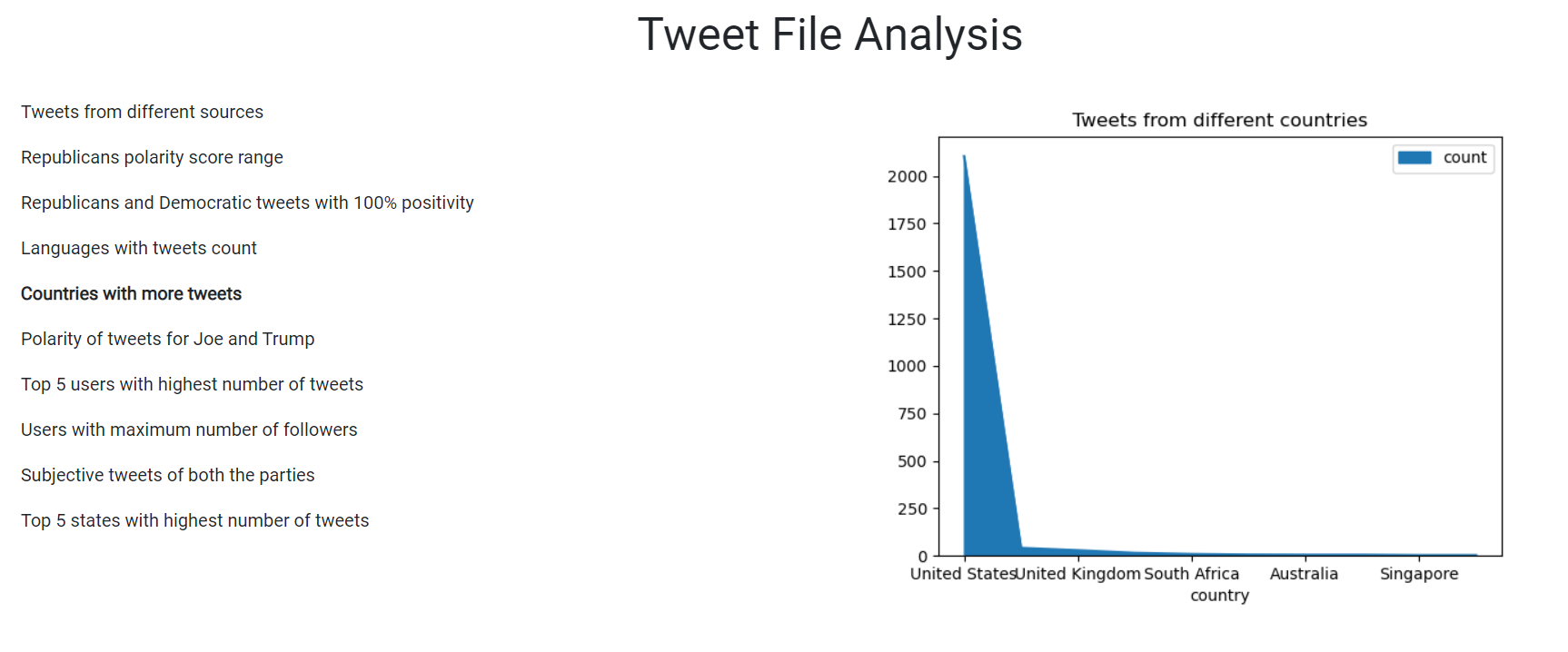
4). Tweets collected from different languages using “Area plot”



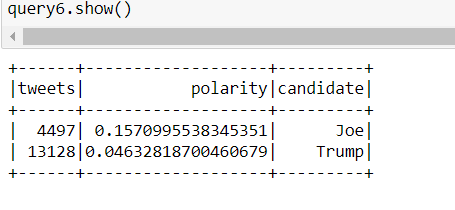


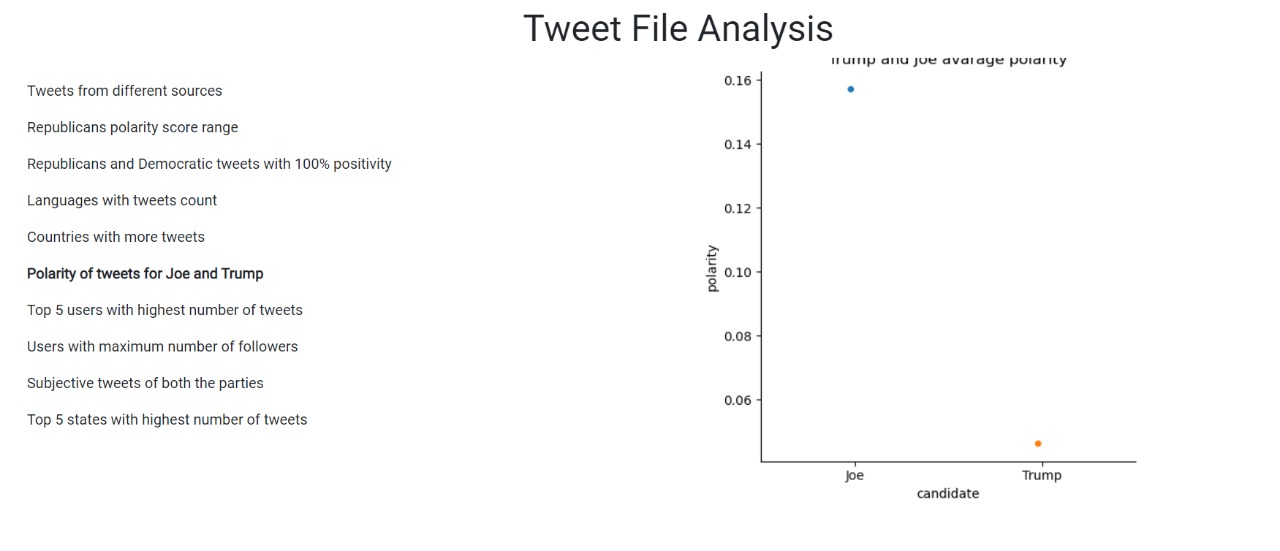
5). Top 5 Countries with Highest number of Tweets using “Area Code”



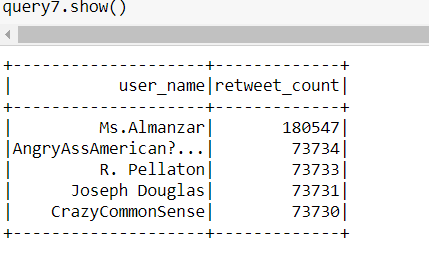


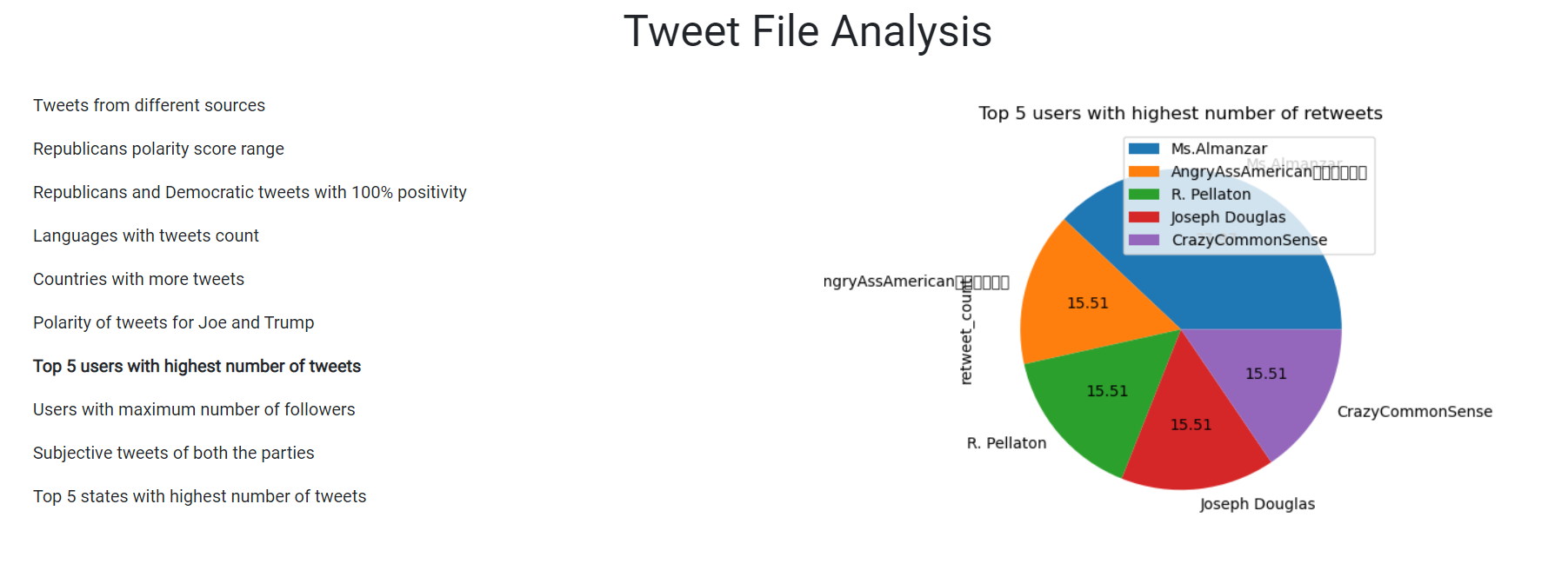
6). Comparing the Polarity of Tweets for Joe and Trump using “Line Graph”



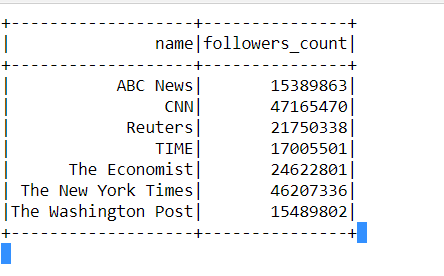


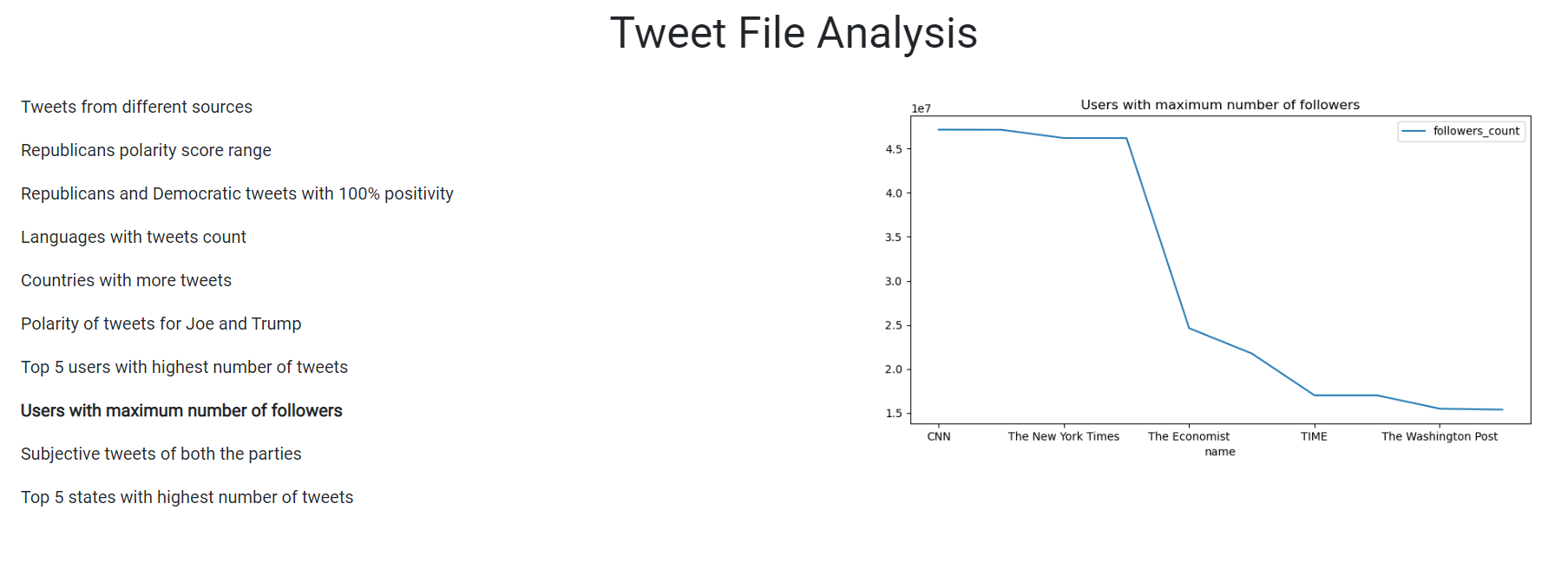
7). Top 5 Users with highest number of tweets from the collected data .



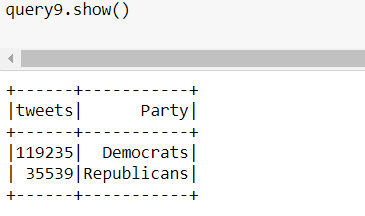


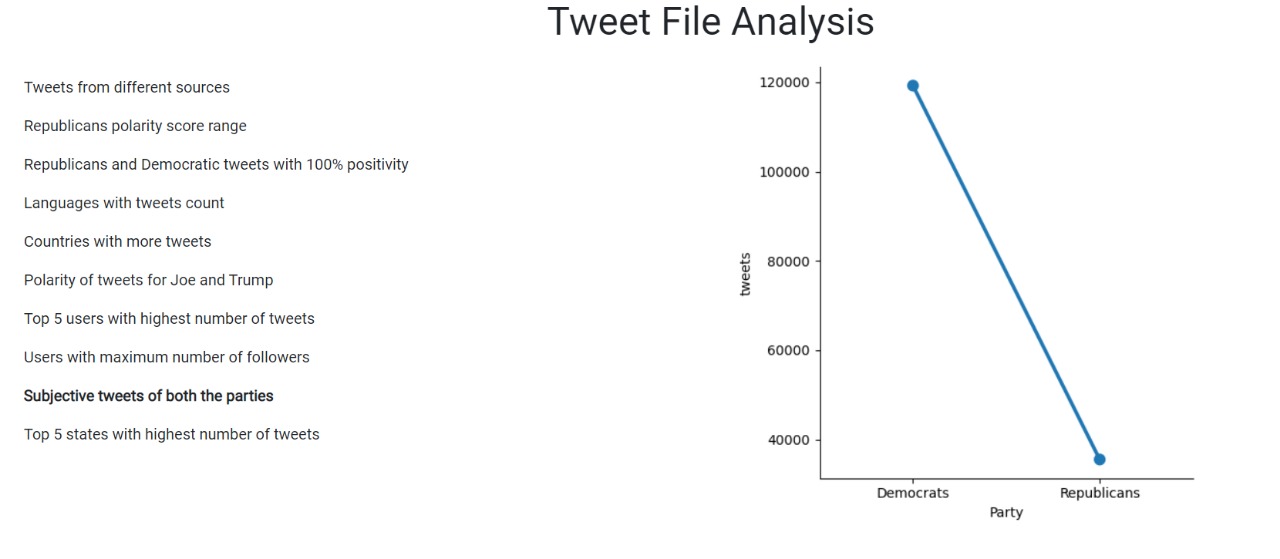
8). Top 5 users with maximum number of followers using “Line Graph”





9). Subjective tweets of Democratic and Republicans parties.





10). Top 5 states in United states with highest number of tweets.

