RITI CHAKRABORTY, 17231417

Assignment-03: Data Visualization

This assignment deals with visualizing twitter data generated from twitter apps using anaconda prompt. The twitter data fetched are related to the tweets which were made on web mining. The data is obtained and saved in json format.

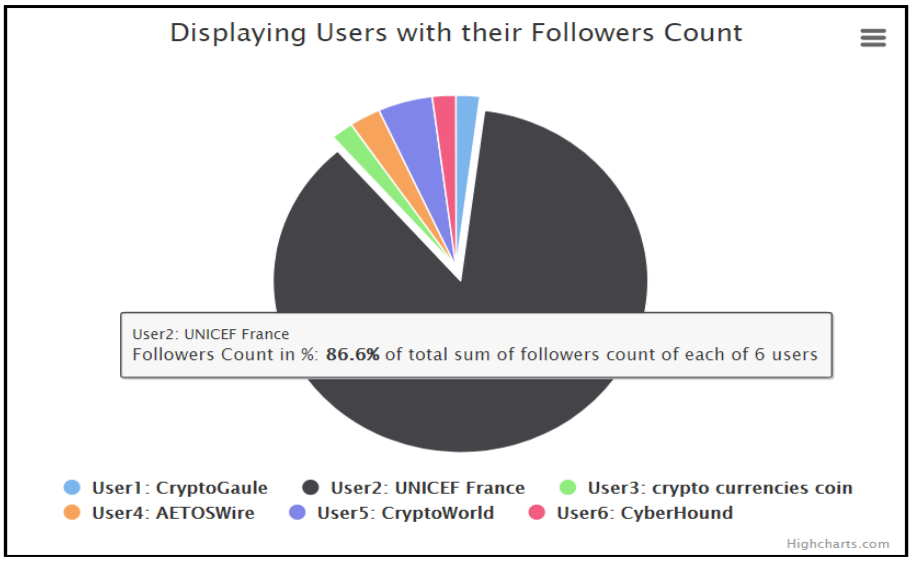
Data Preprocessing:

1. The json file is first converted into .csv file using 'open refine' tool.
2. The csv file is read into pandas data frame.
3. The cells with missing values were removed.
4. The data fetched from the twitter app did not have coordinates, therefore a separate dataset has been used to get the coordinates.
5. From the link: https://developers.google.com/public-data/docs/canonical/countries_csv, countries.csv file has been downloaded to perform a merge with the twitter data so as to obtain the coordinates.
6. However, after performing the merge function only six records were extracted from the subset of records that was generated after removing the NA values.
7. All these six records are used to generate the visualization.
8. The attributes chosen are: Username, followers count, Location, Retweet Count, Favorites count, Latitudes and Longitudes.
9. Each of the attributes are then converted into json scripts using json.dumps() function.

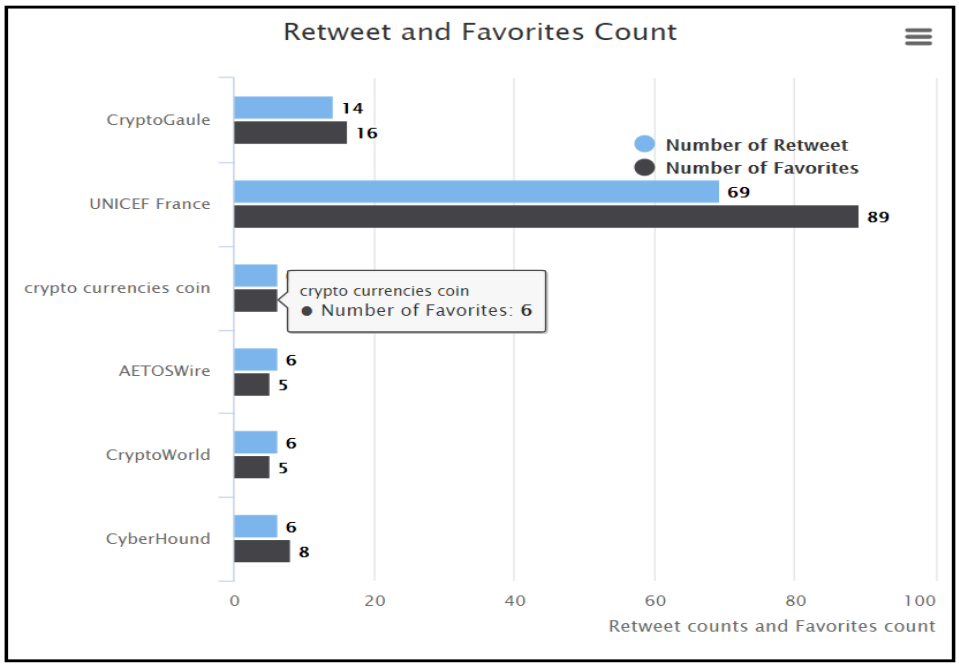
Design of the web page:

The page has two charts and one map which helps in visualizing the data which is fetched from the twitter apps.

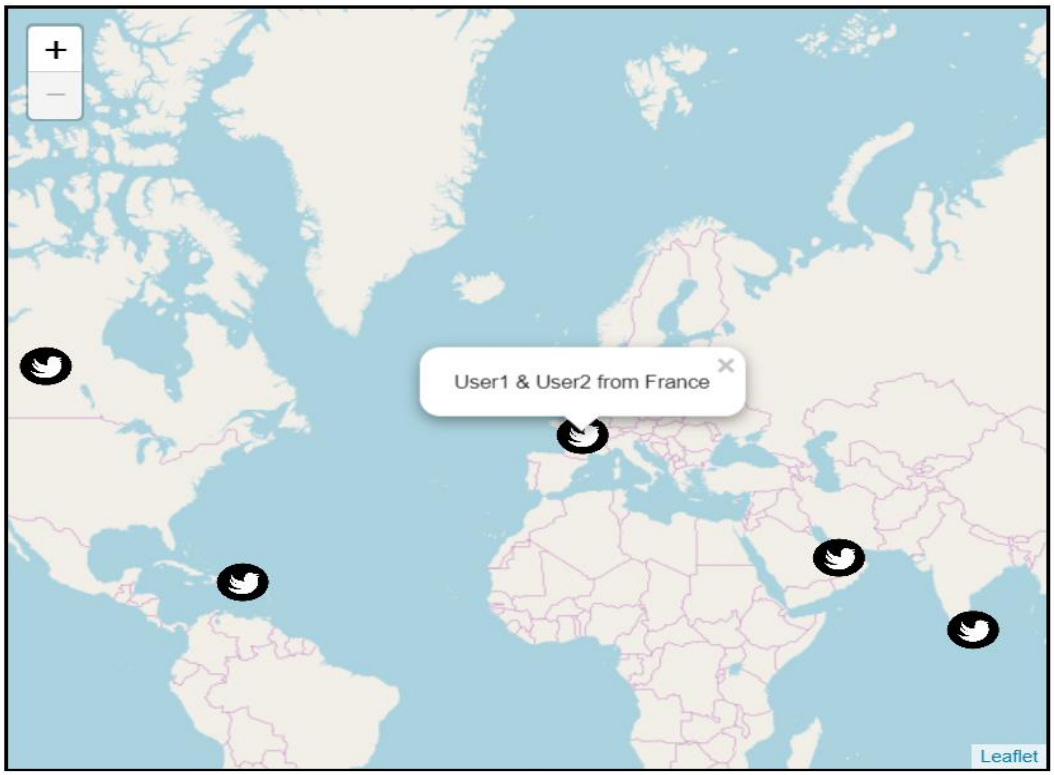
1. The first is a pie chart showing the percentage of followers count each user has. The value is calculated depending upon the summed-up values followers of each user. On clicking each section on the pie-chart, the values are revealed.



2. The second is a reversed bar graph. It illustrates the number of retweets made by each of the six users and the number of favorites marked by the user. The values of retweet count and favorites count were over a wide range therefore some lower values were invisible on the bar graph as the max value was too high. Therefore, the values have been normalized by adding a random constant number so as to make the data visible on the graph.



3. The third is a leaflet map. The coordinates obtained by performing merge are used to map the location of each user on the leaflet.



Visualization Decisions:

1. The file tweets.json is converted to csv and read in pandas as mentioned above. For creating the charts and the map-based visualization, the format of data required is json. Therefore, after cleaning the data and extracting the useful records and attributes using pandas, I have converted the data in the json format. Then copy pasted the data from pandas output to the charts and maps scripts.

Note: The tweets.json file and the python files are present in the folder for further reference.