Twitter Data Analysis

# Scenario:

You are provided with two files to perform basic data analysis:

* tweet.json – Twitter feed
* Dictionary.txt – Contains words with a rating for each word

You are required to analyze user comments in the tweet.json file and assign it a positive or negative polarity based on the words available in the Dictionary file. (Please feel free to add more words to the dictionary file if needed). You might have to clean the JSON file before importing it in Hive.

Solve the following problems in question 1. (90 points)

1. Using the files provided, answer the following:
   1. What are the hashtags used and how many times each hashtag is used?
   2. Which State have the most active users and how many tweets are posted by State?
   3. Based on the user’s followers count, who are the top ten users who have tweeted?
   4. What is the polarity score for each tweet that was posted? Does the tweet have a positive or negative sentiment?

Hint: Use the dictionary.txt for the score

Note: Include the date in the format ’yyyy-mm-dd’, with tweet id, user name and the score.

1. Do you find any problem in the way sentiment analysis was performed in the previous question? If so, how will you improve it? (10 + 10 Bonus points)

Note: You need not solve the proposed solution with the twitter file. Also, note that the Bonus points can only be used towards missed assignment points.

# Deliverables:

1. You must provide the queries along with the screenshots of each question. The screenshot should include Query-ID, job ID, mappers and reducers used.
2. Final deliverable must be submitted in a Word file.