CprE 419 Lab 5: Analyzing Twitter Data using MapReduce

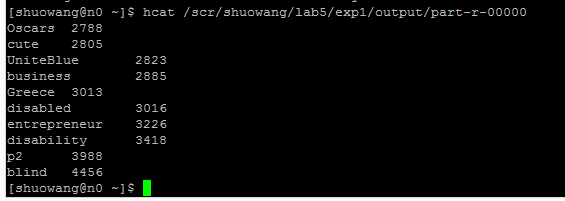
**Shuo Wang**

All the commented codes, job files and output logs are included in the submission.

The input output paths were hard coded.

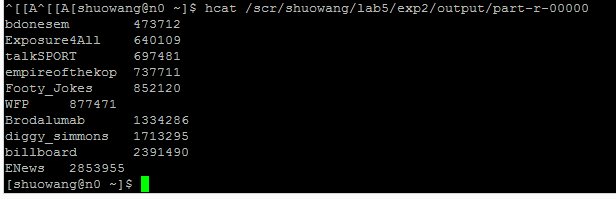
Exp1: Top 10 most common hashtags in oscars.json

(for each tweet, the duplicate hashtags were removed)



So the top 1 most common hashtag in oscars.json is #blind which the frequency of 4456. The number 2 to 10 can be seen in the screen capture (ascending order).

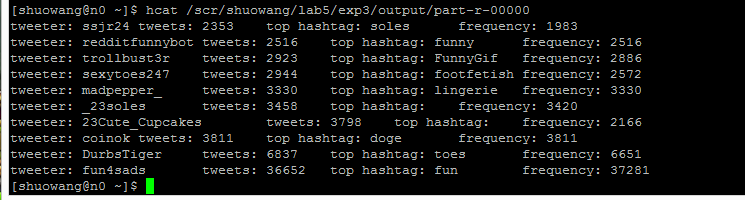
Exp2: top ten most followed tweeters in oscars.json



So the top 1 most followed tweeter in oscars.json is ENews who has 2,853,955 followers at the time the file was queried. The number 2 to 10 can be seen in the screen capture (ascending order).

Exp3: top ten most prolific users and their most commonly used hashtag

(This exp doesn’t require to remove the duplicate hashtags for each tweet, comparing to exp1)



So the top 1 most prolific tweeter in usa.json is fun4sads who has 36,652 tweets at the time the file was queried and its most commonly used hashtag is #fun with frequency of 37,281. The number 2 to 10 can be seen in the screen capture (ascending order).