

## 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)

```
[shuowang@n0 ~]$ hcat /scr/shuowang/lab5/exp1/output/part-r-00000
Oscars 2788
cute 2805
UniteBlue 2823
business 2885
Greece 3013
disabled 3016
entrepreneur 3226
disability 3418
p2 3988
blind 4456
[shuowang@n0 ~]$
```

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

```
^[A^[[A[shuowang@n0 ~]$ hcat /scr/shuowang/lab5/exp2/output/part-r-00000
bdonesem 473712
Exposure4All 640109
talkSPORT 697481
empireofthekop 737711
Footy_Jokes 852120
WFP 877471
Brodalumab 1334286
diggy_simmons 1713295
billboard 2391490
ENews 2853955
[shuowang@n0 ~]$
```

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)

```
[shuowang@n0 ~]$ hcat /scr/shuowang/lab5/exp3/output/part-r-00000
tweeter: ssjr24 tweets: 2353 top hashtag: soles frequency: 1983
tweeter: redditfunnybot tweets: 2516 top hashtag: funny frequency: 2516
tweeter: trollbust3r tweets: 2923 top hashtag: FunnyGif frequency: 2886
tweeter: sexytoes247 tweets: 2944 top hashtag: footfetish frequency: 2572
tweeter: madpepper_ tweets: 3330 top hashtag: lingerie frequency: 3330
tweeter: _23soles tweets: 3458 top hashtag: frequency: 3420
tweeter: 23Cute_Cupcakes tweets: 3798 top hashtag: frequency: 2166
tweeter: coinok tweets: 3811 top hashtag: doge frequency: 3811
tweeter: DurbsTiger tweets: 6837 top hashtag: toes frequency: 6651
tweeter: fun4sads tweets: 36652 top hashtag: fun frequency: 37281
[shuowang@n0 ~]$
```

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).