Unix Command

1.1

In [32]:

```
# Extract fist 500,000 lines into "test_set_tweets_500000.txt "
!head -500000 test_set_tweets.txt > test_set_tweets_500000.txt
```

In [73]:

```
# Extract hashtags words and store them into "hashtags_500000.txt"
!grep -P -o "#[^ \t]+" test_set_tweets_500000.txt > hashtags_500000.txt
```

In [74]:

```
# First 10 lines of "hashtags_500000.txt"
!head -10 hashtags_500000.txt
```

```
#confession.
#worstfeeling.:
#FF
#mm.
#niggas.
#dontjudgeme
#nowplaying.
#nowplaying.
#PersonalBelief
#imjustsayin
```

In [75]:

```
# strip out punctuation and convert uppercase to lowercase
!sed 's/#//g' hashtags_500000.txt | sed 's/[^a-zA-Z]//g' | sed -e 's/\(.*\)/\L
\1/' > keywords_500000.txt
```

In [76]:

```
# Calculate frequence of hashtags and store the result into "result_hastags_5000
00.txt"
!sort keywords_500000.txt | uniq --count | sort -nr > result_hastags_500000.txt
```

In [91]:

```
# Result of top 10 hashtags
!head -10 result_hastags_500000.txt
```

```
   3581 ff
   1809 nowplaying
   1402 fb
   1361
   1029 mm
    686 fail
    622 random
    591 haiti
    529 shoutout
    457 followfriday
```

In [92]:

```
# Shell script of 1.1
!cat 1_1.sh
```

```
#!/bin/sh
sed 's/#//g' hashtags_500000.txt | sed 's/[^a-zA-Z]//g' | sed -e
's/\(.*\)/\L\1/' > keywords_500000.txt
sort keywords_500000.txt | uniq --count | sort -nr > result_hastags
_500000.txt
head -10 result_hastags_500000.txt
```

In [101]:

```
# Runtime of 1.1 using Unix command
!time bash 1_1.sh
```

```
   3581 ff
   1809 nowplaying
   1402 fb
   1361
   1029 mm
    686 fail
    622 random
    591 haiti
    529 shoutout
    457 followfriday
0.33user 0.01system 0:00.25elapsed 133%CPU (0avgtext+0avgdata 6436m
axresident)k
0inputs+2232outputs (0major+2558minor)pagefaults 0swaps
```

1.2

In [13]:

```
# Extract fist 250,000 lines into "training_set_tweets_250000.txt"
!head -250000 training_set_tweets.txt > training_set_tweets_250000.txt
```

In [14]:

```
# First 10 lines in "training_set_tweets_250000.txt"
!head -10 training_set_tweets_250000.txt
```

In [15]:

```
# Join the first 500,000 lines from "test_set_tweets_500000.txt" and the first 2
50,000 lines from "training_set
# _tweets_250000.txt" into "tweets.txt"
!cat test_set_tweets_500000.txt training_set_tweets_250000.txt > tweets.txt
```

In [16]:

```
# Number of lines in "tweets.txt"
!wc -l tweets.txt
```

```
750000 tweets.txt
```

In [84]:

```
# Extract username and store them into "tweets_username.txt"
!grep -P -o "@[^ \t]+" tweets.txt > tweets_username.txt
```

In [85]:

```
# First 10 lines in "tweets_username.txt"
!head -10 tweets_username.txt
```

```
@LovelyJ_Janelle
@Iam_MarkyMark
@Iam_MarkyMark
@Iam_MarkyMark
@seanlamar919
@TRenee3
@LovelyJ_Janelle
@seanlamar919
@seanlamar919
@Iam_MarkyMark:
```

In [86]:

```
# Calculate frequence of hashtags and store the result into "result_username_750
000.txt"
!sort tweets_username.txt | uniq --count | sort -nr > result_username_750000.txt
```

In [87]:

```
# Result of top 10 usernames
! head -10 result_username_750000.txt
```

```
   1234 @RevRunWisdom:
    939 @listensto
    525 @DonnieWahlberg
    441 @OGmuscles
    419 @addthis
    411 @breatheitin
    354 @justinbieber
    347 @MAV25
    303 @karlievoice
    291 @mtgcolorpie
```

In [102]:

```
# Shell script for 1.2 (1)
!cat 1_2.sh
```

```
#!/bin/sh
grep -P -o "@[^ \t]+" tweets.txt > tweets_username.txt
sort tweets_username.txt | uniq --count | sort -nr > result_usernam
e_750000.txt
head -10 result_username_750000.txt
```

In [103]:

```
# Runtime of 1.2 (1) using Unix command
!time bash 1_2.sh
```

```
   1234 @RevRunWisdom:
    939 @listensto
    525 @DonnieWahlberg
    441 @OGmuscles
    419 @addthis
    411 @breatheitin
    354 @justinbieber
    347 @MAV25
    303 @karlievoice
    291 @mtgcolorpie
2.06user 0.04system 0:01.12elapsed 187%CPU (0avgtext+0avgdata 55920
maxresident)k
0inputs+25128outputs (0major+16292minor)pagefaults 0swaps
```

1.2

In [71]:

```
# Extract username and store them into "tweets_username.txt"
!grep -P -o "#[^ \t]+#[^ \t]+" tweets.txt > tweets_twohashtags.txt
```

In [88]:

```
# First 10 lines in "tweets_twohashtags.txt"
!head -10 tweets_twohashtags.txt
```

```
#trueshit...#scaryshit...but
#!%#$^!@%&
#honey.......#imjussayn
#pause.....#megapause
#9:#Virtualization
#8:#Offshore
#thatisall--#agree
#FF--#FF
#LENT....#SMH
#UCanTakeTheKidOutHoodBut#UCanTakeTheKidOutHoodBut#UCanTakeTheKidOu
tHoodBut#UCanTakeTheKidOutHoodBut#UCanTakeTheKidOutHoodBut
```

In [89]:

```
# Calculate frequence of hashtags and store the result into "result_twohastags_7
50000.txt"
!sort tweets_twohashtags.txt | uniq --count | sort -nr > result_twohashtags_7500
00.txt
```

In [90]:

```
# Result of top 10 tweets that have at least two hashtags
! head -10 result_twohashtags_750000.txt
```

```
      8 #affiliate#marketing
      5 #zewdy#zewdy
      5 #BGC#BGC
      5 ####
      4 #??PFoundersday#??PFoundersday
      4 #Celebrity,#Philanthropy
      3 #AKA#AKA
      3 #39;What&#39;s
      3 #39;streaming&#39;
      3 #39;Green&#39;
```

In [104]:

```
# Shell script for 1.2 (2)
!cat 1_3.sh
```

```
#!/bin/sh
grep -P -o "#[^ \t]+#[^ \t]+" tweets.txt > tweets_twohashtags.txt
sort tweets_twohashtags.txt | uniq --count | sort -nr > result_twoh
ashtags_750000.txt
head -10 result_twohashtags_750000.txt
```

In [105]:

```
# Runtime of 1.2 (2) using Unix command
!time bash 1_3.sh
```

```
      8 #affiliate#marketing
      5 #zewdy#zewdy
      5 #BGC#BGC
      5 ####
      4 #??PFoundersday#??PFoundersday
      4 #Celebrity,#Philanthropy
      3 #AKA#AKA
      3 #39;What&#39;s
      3 #39;streaming&#39;
      3 #39;Green&#39;
0.17user 0.01system 0:00.18elapsed 100%CPU (0avgtext+0avgdata 3108m
axresident)k
0inputs+48outputs (0major+828minor)pagefaults 0swaps
```

2

In [ ]:

```
# Extract fist 250,000 lines into "training_set_tweets_250000.txt"
!head -500000  edges.csv > edges_500000.csv
```

In [18]:

```
# Swap order is userID is larger than followerID and store the result into "edge
s_500000_dup.csv"
!awk -F "," '{if($1<$2) printf("%d,%d\n", $1,$2);if($1>$2) printf("%d,%d\n",
 $2,$1)}' edges_500000.csv > edges_500000_dup.csv
```

In [19]:

```
# Find pairs that appear twice (reciprocal follower) and store it into "output.c
sv"
!sort edges_500000_dup.csv | uniq --count --repeated > output.csv
```

In [34]:

```
# Report reciprocal followers
!grep -E -o " [0-9]+,[0-9]+$" output.csv | awk -F "," '{printf("%d,%d\n%d,%d\n",
$1,$2, $2,$1)}' > result_reciprocalFollowers.txt
```

```
100591,100721
100721,100591
102898,122546
122546,102898
13232,18205
18205,13232
13232,63255
63255,13232
134409,134410
134410,134409
135546,135684
135684,135546
15574,15926
15926,15574
192865,192899
192899,192865
19628,19821
19821,19628
19628,20033
20033,19628
201063,40997
40997,201063
201078,201607
201607,201078
22196,76473
76473,22196
23503,41422
41422,23503
31866,32002
32002,31866
32173,32452
32452,32173
33099,62167
62167,33099
33884,34046
34046,33884
33884,34101
34101,33884
3682,5276
5276,3682
40704,40997
40997,40704
40704,41039
41039,40704
40997,41039
41039,40997
40997,62623
62623,40997
58783,58875
58875,58783
60887,70696
70696,60887
63255,65435
65435,63255
65411,65435
65435,65411
65435,93260
93260,65435
70696,70772
70772,70696
78182,78464
```

```
78464,78182
80092,80096
80096,80092
89222,89350
89350,89222
93260,93427
93427,93260
```

In [106]:

```
# Number of reciprocal followers: 34 * 2
!grep -E -o " [0-9]+,[0-9]+$" output.csv | awk -F "," '{printf("%d,%d\n%d,%d\n",
$1,$2, $2,$1)}' | wc -l
```

```
68
```

In [115]:

```
# Shell script for 2
!cat 2.sh
```

```
#!/bin/sh
awk -F "," '{if($1<$2) printf("%d,%d\n", $1,$2);if($1>$2) printf("%
d,%d\n", $2,$1)}' edges_500000.csv > edges_500000_dup.csv
sort edges_500000_dup.csv | uniq --count --repeated > output.csv
grep -E -o " [0-9]+,[0-9]+$" output.csv | awk -F "," '{printf("%d,%
d\t%d,%d\t", $1,$2, $2,$1)}'
echo "\n"
```

In [116]:

```
# Runtime of 2 using Unix command
!time bash 2.sh
```

```
100591,100721   100721,100591   102898,122546   122546,102898   132
32,18205        18205,13232     13232,63255     63255,13232     134
409,134410      134410,134409   135546,135684   135684,135546   155
74,15926        15926,15574     192865,192899   192899,192865   196
28,19821        19821,19628     19628,20033     20033,19628     201
063,40997       40997,201063    201078,201607   201607,201078   221
96,76473        76473,22196     23503,41422     41422,23503     318
66,32002        32002,31866     32173,32452     32452,32173     330
99,62167        62167,33099     33884,34046     34046,33884     338
84,34101        34101,33884     3682,5276       5276,3682       407
04,40997        40997,40704     40704,41039     41039,40704     409
97,41039        41039,40997     40997,62623     62623,40997     587
83,58875        58875,58783     60887,70696     70696,60887     632
55,65435        65435,63255     65411,65435     65435,65411     654
35,93260        93260,65435     70696,70772     70772,70696     781
82,78464        78464,78182     80092,80096     80096,80092     892
22,89350        89350,89222     93260,93427     93427,93260     \n
1.53user 0.03system 0:00.71elapsed 219%CPU (0avgtext+0avgdata 55792
maxresident)k
0inputs+13344outputs (0major+14257minor)pagefaults 0swaps
```