

1.2.1 (Unix)

In [13]:

```
# Extract first 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 frequency of hashtags and store the result into "result_username_750000.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_username_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
```