1.1 (Unix)

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
```