Data Science Capstone Quiz 1

Terence Lim Zheng Wei

6 April 2016

```
library(R.utils)
```

Question 1

The en_US.blogs.txt file is how many megabytes?

```
bytes<-file.info('../en_US/en_US.blogs.txt')$size
megabytes <- bytes/1000000
```

The file is 210.160014 megabytes.

Question 2

The en_US.twitter.txt has how many lines of text?

```
num_lines <- countLines('../en_US/en_US.twitter.txt')
```

The file has 2360148 number of lines

Question 3

What is the length of the longest line seen in any of the three en_US data sets?

This is achieved by using the `wc -L <filename>` command in bash.

The length of the longest line is over 40 thousand in the blogs data set

Question 4

In the en_US twitter data set, if you divide the number of lines where the word "love" (all lowercase) occurs by the number of lines the word "hate" (all lowercase) occurs, about what do you get?

grep -c "love" en_US.twitter.txt => 90956

grep -c "hate" en_US.twitter.txt => 22138

Value = 4.1085916

Question 5

The one tweet in the en_US twitter data set that matches the word "biostats" says what?

grep "biostats" en_US.twitter.txt

i know how you feel.. i have biostats on tuesday and i have yet to study =/

Question 6

How many tweets have the exact characters “A computer once beat me at chess, but it was no match for me at kickboxing”. (I.e. the line matches those characters exactly.)

grep -c “A computer once beat me at chess, but it was no match for me at kickboxing” en_US.twitter.txt => 3