Data Science Capstone - Quiz 01

1. Question

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

```
setwd("C:/Coursera/10_Data_Science_Capstone")
FileInfo <- file.info("en_US.blogs.txt")
sizeB <- FileInfo$size
sizeKB <- sizeB/1024
sizeMB <-  sizeKB/1024
sizeMB
```

```
## [1] 200.4242
```

*** 200**
150
250
100

2. Question

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

```
linesTwitter <- readLines("en_US.twitter.txt")
length(linesTwitter)
```

```
## [1] 2360149
```

Around 1 million
*** Over 2 million**
Around 2 hundred thousand
Around 5 hundred thousand

3. Question

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

Counting Blogs file

```
blogsFile <- file("en_US.blogs.txt")
blogsFileLines<-readLines(blogsFile)
close(blogsFile)
summary(nchar(blogsFileLines))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##     1.0    47.0   157.0   231.7   331.0 40840.0
```

Counting News file

```
newsFile <- file("en_US.news.txt")
newsFileLines<-readLines(newsFile)
close(newsFile)
summary(nchar(newsFileLines))
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       2     111     186     203     270    5760
```

Counting Twitter file

```
twitterFile <- file("en_US.twitter.txt")
twitterFileLines<-readLines(twitterFile)
close(twitterFile)
summary(nchar(twitterFileLines))
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     2.0    37.0    64.0    68.8   100.0   213.0
```

*** Over 40 thousand in the blogs data set**
Over 40 thousand in the news data set
Over 11 thousand in the blogs data set
Over 11 thousand in the news data set

4. Question

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?

```
lovelines <- (grepl(" love ", readLines("en_US.twitter.txt")))
numlove <- table(lovelines)["TRUE"]

hatelines <- (grepl(" hate ", readLines("en_US.twitter.txt")))
numhate <- table(hatelines)["TRUE"]

proportion <- numlove/numhate

proportion
##     TRUE
## 4.616849
```

0.5
2
0.25
*** 4**

5. Question

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

```
twitterLines <- readLines("en_US.twitter.txt")
```

```
lineTarget <- grep("biostats",twitterLines)
```

```
twitterLines[lineTarget]
```

```
## [1] "i know how you feel.. i have biostats on tuesday and i have yet to study =/"
```

They just enrolled in a biostat program
It's a tweet about Jeff Leek from one of his students in class
*** They haven't studied for their biostats exam**
They need biostats help on their project

6. Question

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("A computer once beat me at chess, but it was no match for me at kickboxing", twitte
rLines)
```

```
## [1]  519059  835824 2283423
```

*** 3**
0
1
2