Coursera Capstone Project: Quiz 1

Roel Peters

19 november 2016

```
suppressWarnings(suppressMessages(twitterData <- readLines('dataset/final/en_US/en_US.twitter.txt')))
suppressWarnings(suppressMessages(newsData <- readLines('dataset/final/en_US/en_US.news.txt')))
suppressWarnings(suppressMessages(blogData <- readLines('dataset/final/en_US/en_US.blogs.txt')))
```

Question 1

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

```
file.info('dataset/final/en_US/en_US.blogs.txt')$size/1000000
```

```
## [1] 210.16
```

Question 2

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

```
length(twitterData)
```

```
## [1] 2360148
```

Question 3

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

```
twitterDataLength <- nchar(twitterData)
newsDataLength <- nchar(newsData)
blogDataLength <- nchar(blogData)
```

Blog file: 40835
News file: 5760
Twitter file: 5760

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?

```
twitterLove <- grepl('love',twitterData,ignore.case=F)
twitterHate <- grepl('hate',twitterData,ignore.case=F)
length(twitterData[twitterLove])/length(twitterData[twitterHate])
```

```
## [1] 4.108592
```

Question 5

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

```
twitterData[grep('biostats',twitterData)]
```

```
## [1] "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.)

```
length(twitterData[grepl('A computer once beat me at chess, but it was no match for me at kickboxing',tw
```

```
## [1] 3
```