Data Science Capstone - Quiz 1

Shane Kao

Sunday, February 15, 2015

```
setwd("C:/Users/asus/Downloads/final/en_US")
twitter=readLines("en_US.twitter.txt")
blogs=readLines("en_US.blogs.txt")
news=readLines("en_US.news.txt")
```

Question 1

The `en_US_blogs.txt` file is how many megabytes?

```
file.info(list.files(pattern="*.txt"))
```

```
##                        size isdir mode               mtime
## en_US.blogs.txt   210160014 FALSE  666 2014-07-22 10:13:06
## en_US.news.txt    205811885 FALSE  666 2015-02-15 23:01:23
## en_US.twitter.txt 167105338 FALSE  666 2014-07-22 10:12:58
##                                  ctime               atime exe
## en_US.blogs.txt   2015-02-15 19:12:37 2015-02-15 19:12:37  no
## en_US.news.txt    2015-02-15 19:12:29 2015-02-15 19:12:29  no
## en_US.twitter.txt 2015-02-15 19:12:22 2015-02-15 19:12:22  no
```

```
file.info(list.files(pattern="*.txt"))["en_US.blogs.txt","size"]/1024/1024
```

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

Question 2

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

```
length(twitter)
```

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

Question 3

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

```
summary(nchar(twitter,allowNA = TRUE))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    2.00   36.00   64.00   68.34   99.00  141.00   73348
```

```
summary(nchar(blogs,allowNA = TRUE))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     1.0    37.0   106.0   184.9   265.0 40830.0  256178
```

```
summary(nchar(news,allowNA = TRUE))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     1.0   104.0   180.0   195.3   262.0 11380.0  123543
```

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?

```
length(grep("love",twitter))/length(grep("hate",twitter))
```

```
## [1] 4.109063
```

Question 5

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

```
twitter[grep("biostats",twitter)]
```

```
## [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.)

```
text="A computer once beat me at chess, but it was no match for me at kickboxing"
length(twitter[grep(text,twitter)])
```

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