Quiz 1

Xinchen Pan

December 10, 2015

Load the data

```
#using skipNul to avoid problem like line xxx appears to contain an embedded nul

twitter <- readLines("en_US.twitter.txt",encoding = "UTF-8",skipNul=T)
blogs <- readLines("en_US.blogs.txt",encoding = "UTF-8",skipNul=T)
news <- readLines("en_US.news.txt",encoding = "UTF-8",skipNul=T)
```

```
## Warning in readLines("en_US.news.txt", encoding = "UTF-8", skipNul = T):
## incomplete final line found on 'en_US.news.txt'
```

The warning is caused by the empty line in the last line of news dataset. I can delete that line instead, since it does not affect analysis I will leave as it is.

1

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

```
#There might be an option to change from bytes to megabytes.
file.size("en_US.blogs.txt")/1024/1024
```

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

2

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

```
length(twitter)
```

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

3

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

```
library(stringr)
max(str_length(twitter))
```

```
## [1] 140
```

```
max(str_length(blogs))
```

```
## [1] 40833
```

```
max(str_length(news))
```

```
## [1] 5760
```

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?

```
love <- length(grep('love', twitter))
hate <- length(grep('hate', twitter))
love / hate
```

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

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 =/"
```

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.)

```r
grep("A computer once beat me at chess, but it was no match for me at kickboxing", twitter)
```

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