Coursera Capstone Project Quiz 1

Abhijit Jantre

19 January 2017

Loading necessary libraries

```
library(utils)
```

Loading the zip file

```
a <- unzip("C:/Data Science/R/Coursera Capstone Project/Project 1/Coursera-SwiftKey.zip")
list(a)
```

```
## [[1]]
##  [1] "./final/de_DE/de_DE.twitter.txt" "./final/de_DE/de_DE.blogs.txt"
##  [3] "./final/de_DE/de_DE.news.txt"    "./final/ru_RU/ru_RU.blogs.txt"
##  [5] "./final/ru_RU/ru_RU.news.txt"    "./final/ru_RU/ru_RU.twitter.txt"
##  [7] "./final/en_US/en_US.twitter.txt" "./final/en_US/en_US.news.txt"
##  [9] "./final/en_US/en_US.blogs.txt"   "./final/fi_FI/fi_FI.news.txt"
## [11] "./final/fi_FI/fi_FI.blogs.txt"   "./final/fi_FI/fi_FI.twitter.txt"
```

Quetion 1

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

200

150

250

100

```
b <- file.size("C:/Data Science/R/Coursera Capstone Project/Project 1/Coursera-SwiftKey/final/en_US/en_
b/1024^2
```

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

Question 2

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

Over 2 million

Around 2 hundred thousand

Around 5 hundred thousand

Around 1 million

```
length(readLines("C:/Data Science/R/Coursera Capstone Project/Project 1/Coursera-SwiftKey/final/en_US/en
```

```
## Warning in readLines("C:/Data Science/R/Coursera Capstone Project/Project
## 1/Coursera-SwiftKey/final/en_US/en_US.twitter.txt"): line 167155 appears to
## contain an embedded nul
```

```
## Warning in readLines("C:/Data Science/R/Coursera Capstone Project/Project
## 1/Coursera-SwiftKey/final/en_US/en_US.twitter.txt"): line 268547 appears to
## contain an embedded nul
```

```
## Warning in readLines("C:/Data Science/R/Coursera Capstone Project/Project
## 1/Coursera-SwiftKey/final/en_US/en_US.twitter.txt"): line 1274086 appears
## to contain an embedded nul
```

```
## Warning in readLines("C:/Data Science/R/Coursera Capstone Project/Project
## 1/Coursera-SwiftKey/final/en_US/en_US.twitter.txt"): line 1759032 appears
## to contain an embedded nul
```

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

Question 3

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

Over 11 thousand in the blogs data set

Over 40 thousand in the news data set

Over 40 thousand in the blogs data set

Over 11 thousand in the news data set

```
twitter <- readLines("C:/Data Science/R/Coursera Capstone Project/Project 1/Coursera-SwiftKey/final/en_
```

```
## Warning in readLines("C:/Data Science/R/Coursera Capstone Project/Project
## 1/Coursera-SwiftKey/final/en_US/en_US.twitter.txt"): line 167155 appears to
## contain an embedded nul
```

```
## Warning in readLines("C:/Data Science/R/Coursera Capstone Project/Project
## 1/Coursera-SwiftKey/final/en_US/en_US.twitter.txt"): line 268547 appears to
## contain an embedded nul
```

```
## Warning in readLines("C:/Data Science/R/Coursera Capstone Project/Project
## 1/Coursera-SwiftKey/final/en_US/en_US.twitter.txt"): line 1274086 appears
## to contain an embedded nul
```

```
## Warning in readLines("C:/Data Science/R/Coursera Capstone Project/Project
## 1/Coursera-SwiftKey/final/en_US/en_US.twitter.txt"): line 1759032 appears
## to contain an embedded nul
```

```
blog <- readLines("C:/Data Science/R/Coursera Capstone Project/Project 1/Coursera-SwiftKey/final/en_US/
news <- readLines("C:/Data Science/R/Coursera Capstone Project/Project 1/Coursera-SwiftKey/final/en_US/
```

```
## Warning in readLines("C:/Data Science/R/Coursera Capstone Project/Project
## 1/Coursera-SwiftKey/final/en_US/en_US.news.txt"): incomplete final line
## found on 'C:/Data Science/R/Coursera Capstone Project/Project 1/Coursera-
## SwiftKey/final/en_US/en_US.news.txt'
```

```
max(nchar(blog))
```

```
## [1] 40835
```

```
max(nchar(news))
```

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

```
max(nchar(twitter))
```

```
## [1] 213
```

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?

0.5

0.25

2

4

```
sum(grepl("love",twitter))/sum(grepl("hate",twitter))
```

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

Question 5

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

It's a tweet about Jeff Leek from one of his students in class

They just enrolled in a biostat program

They need biostats help on their project

They haven't studied for their biostats exam

```
grep("biostats",twitter,value = TRUE)
```

```
## [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.)

3

1

2

0

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

```
##
##   FALSE    TRUE
## 2360145       3
```