Coursera Data Science: Capstone Project Requirement

The goal of this project is just to display that you've gotten used to working with the data and that you are on track to create your prediction algorithm. Please submit a report on R Pubs (http://rpubs.com/ (http://rpubs.com/)) that explains your exploratory analysis and your goals for the eventual app and algorithm. This document should be concise and explain only the major features of the data you have identified and briefly summarize your plans for creating the prediction algorithm and Shiny app in a way that would be understandable to a non-data scientist manager. You should make use of tables and plots to illustrate important summaries of the data set.

The motivation for this project is to:

1. Demonstrate that you've downloaded the data and have successfully loaded it in.
2. Create a basic report of summary statistics about the data sets.
3. Report any interesting findings that you amassed so far.
4. Get feedback on your plans for creating a prediction algorithm and Shiny app.

Project Development Planning

Analyze three corpora of US English text which are found online through the downloaded file Post analysis, it shows, that the two sets of "blogs" & "news" are similar, but the "twitter" one is different due to the 140 character limit

Synopsis

In this report we look at three corpora of US English text, a set of internet blogs posts, a set of internet news articles, and a set of twitter messages.
Collect following information:

1. File size
2. Number of lines
3. Number of words
4. Distribution of words (quantiles and plot)
5. Number of characters

Tools & required Libraries

Use R environment with required libraries: stringi, ggplot2, magrittr. Compile report using Rmarkdown & Knitr & Rpubs to publish this overall Milestone Report

General sequencing of steps for achieving the required goal

1. Download Dataset
2. Import Dataset
3. Sample Dataset
4. Frequency table Data
5. Train Model Data
6. Predict Data

Data Acquisition

The data is provided at https://d396qusza40orc.cloudfront.net/dsscapstone/dataset/Coursera-SwiftKey.zip
(https://d396qusza40orc.cloudfront.net/dsscapstone/dataset/Coursera-SwiftKey.zip)

Specify Source & Destination files <——Avoided downloading file for every iteration of code check and hence have removed it from R code

destination_file <- "Coursera-SwiftKey.zip" source_file <- "http://d396qusza40orc.cloudfront.net/dsscapstone/dataset/Coursera-SwiftKey.zip (http://d396qusza40orc.cloudfront.net/dsscapstone/dataset/Coursera-SwiftKey.zip)"

Download Dataset

download.file(source_file, destination_file)

Extract files by unzipping the downloaded file

unzip(destination_file)

Review unzipped files

unzip(destination_file, list = TRUE)

Review Data

list.files(“final”) list.files(“final/en_US”) ```

The corpora are contained in three separate plain-text files, out of which one is binary

```
setwd("~/datasciencecoursera")
# import the blogs and twitter datasets in text mode
blogs <- readLines("final/en_US/en_US.blogs.txt", encoding="UTF-8")
twitter <- readLines("final/en_US/en_US.twitter.txt", encoding="UTF-8")
```

```
## Warning in readLines("final/en_US/en_US.twitter.txt", encoding = "UTF-8"):
## line 167155 appears to contain an embedded nul
```

```
## Warning in readLines("final/en_US/en_US.twitter.txt", encoding = "UTF-8"):
## line 268547 appears to contain an embedded nul
```

```
## Warning in readLines("final/en_US/en_US.twitter.txt", encoding = "UTF-8"):
## line 1274086 appears to contain an embedded nul
```

```
## Warning in readLines("final/en_US/en_US.twitter.txt", encoding = "UTF-8"):
## line 1759032 appears to contain an embedded nul
```

```
# Import news dataset in binary mode
con <- file("final/en_US/en_US.news.txt", open="rb")
news <- readLines(con, encoding="UTF-8")
close(con)
rm(con)
```

Basic Statistics

Analyze individual file sizes in Mega Bytes (MB)

```
# Analyze individual file sizes in MegaBytes (MB)
file.info("final/en_US/en_US.blogs.txt")$size   / 1024^2
```

```
## [1] NA
```

```
file.info("final/en_US/en_US.news.txt")$size    / 1024^2
```

```
## [1] NA
```

```
file.info("final/en_US/en_US.twitter.txt")$size / 1024^2
```

```
## [1] NA
```

For our analysis we need two libraries.

```
# library for character string analysis
library(stringi)
```

```
## Warning: package 'stringi' was built under R version 3.1.2
```

```
# library for plotting
library(ggplot2)
```

Analyze lines and characters:

```
stri_stats_general( blogs )
```

```
##       Lines LinesNEmpty       Chars CharsNWhite
##      899288      899288   206824382   170389539
```

```
stri_stats_general( news )
```

```
##       Lines LinesNEmpty       Chars CharsNWhite
##     1010242     1010242   203223154   169860866
```

```
stri_stats_general( twitter )
```

```
##       Lines LinesNEmpty       Chars CharsNWhite
##     2360148     2360148   162096031   134082634
```

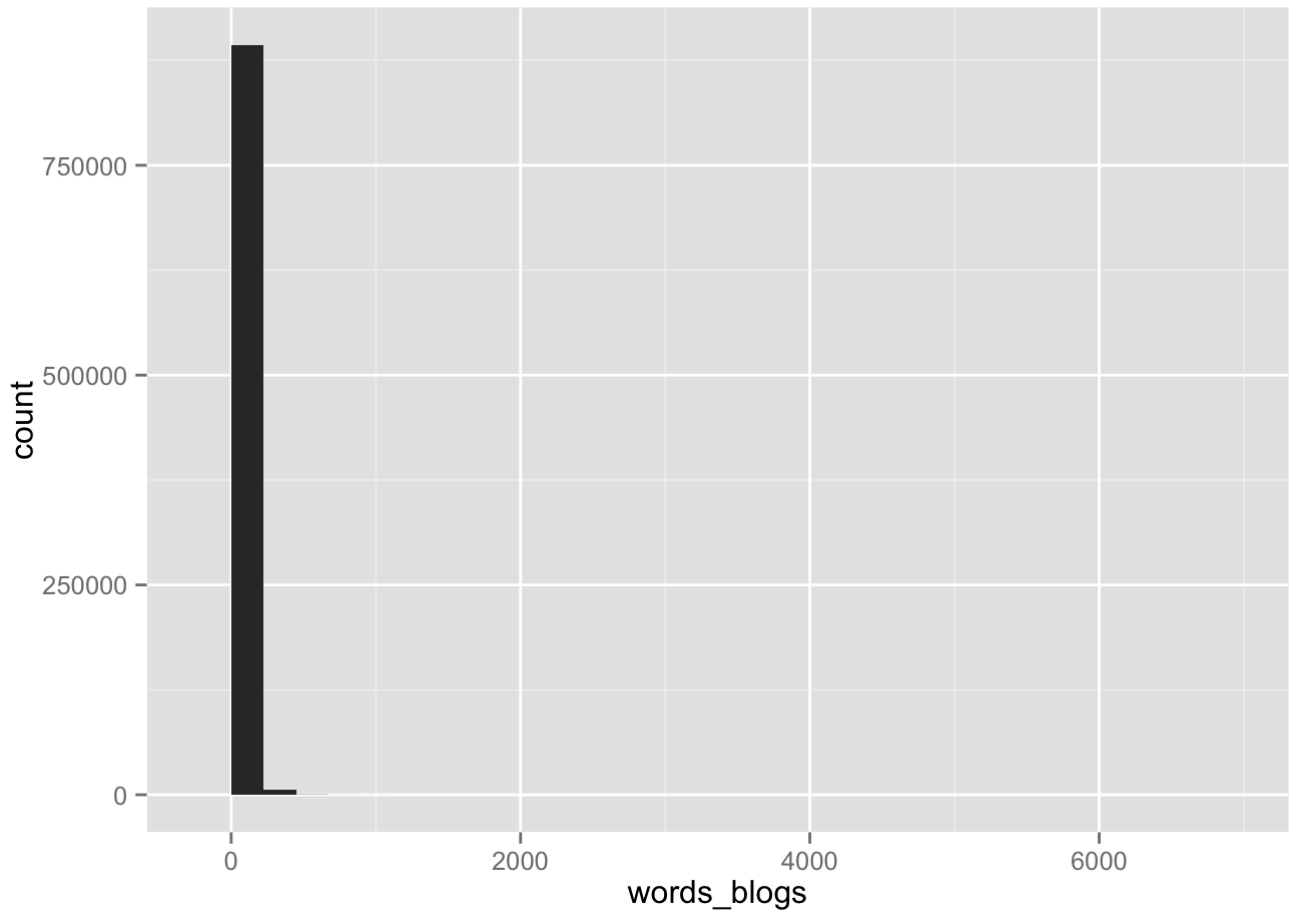
Count words per line & summarize distribution of these counts per corpus, using summary statistics and a distibution plot

```
words_blogs   <- stri_count_words(blogs)
summary( words_blogs )
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    9.00   28.00   41.75   60.00 6726.00
```

```
qplot(   words_blogs )
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



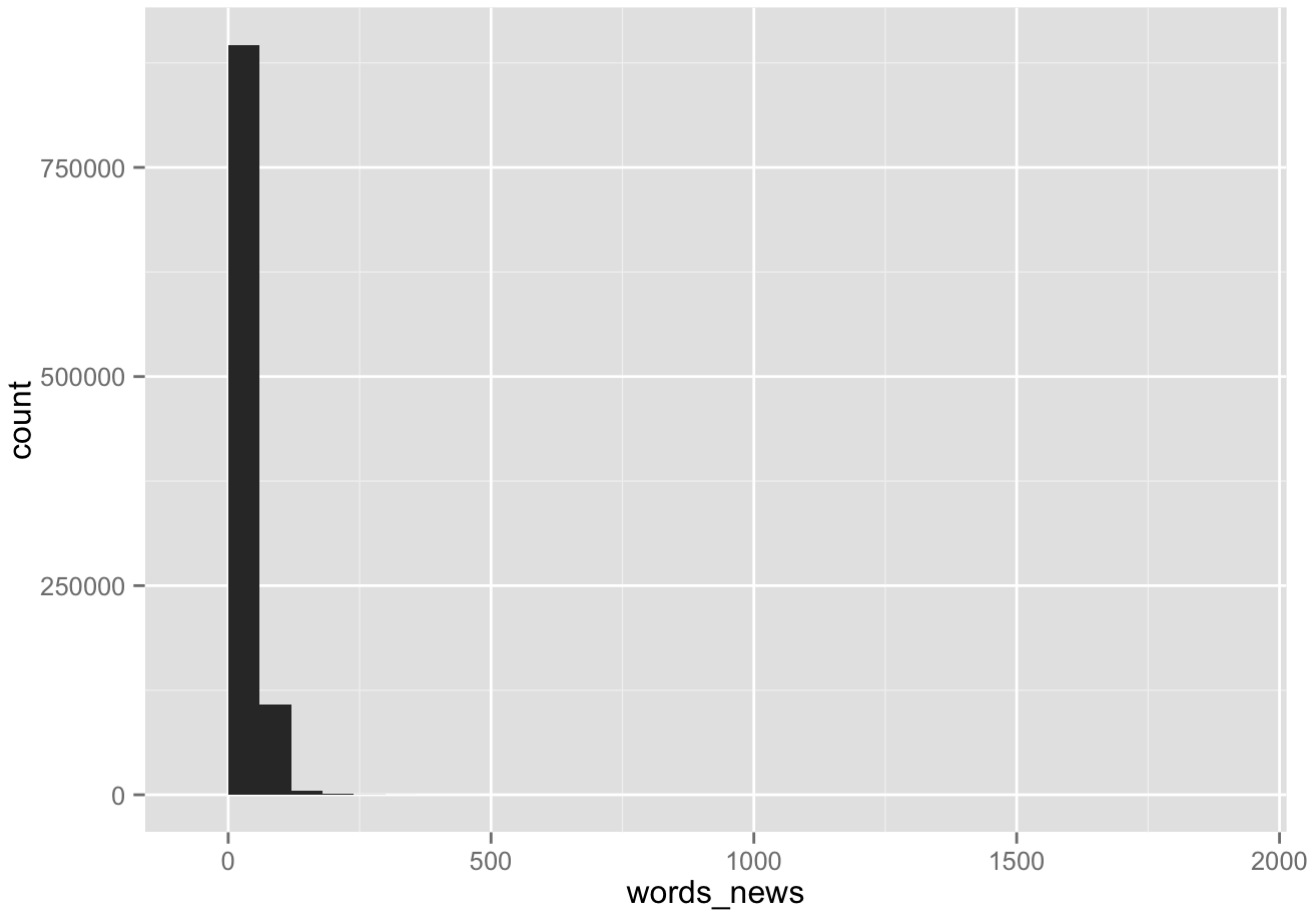
Analyze the “news” corpus:

```
words_news    <- stri_count_words(news)
summary( words_news )
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   19.00   32.00   34.41   46.00 1796.00
```

```
qplot(   words_news )
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



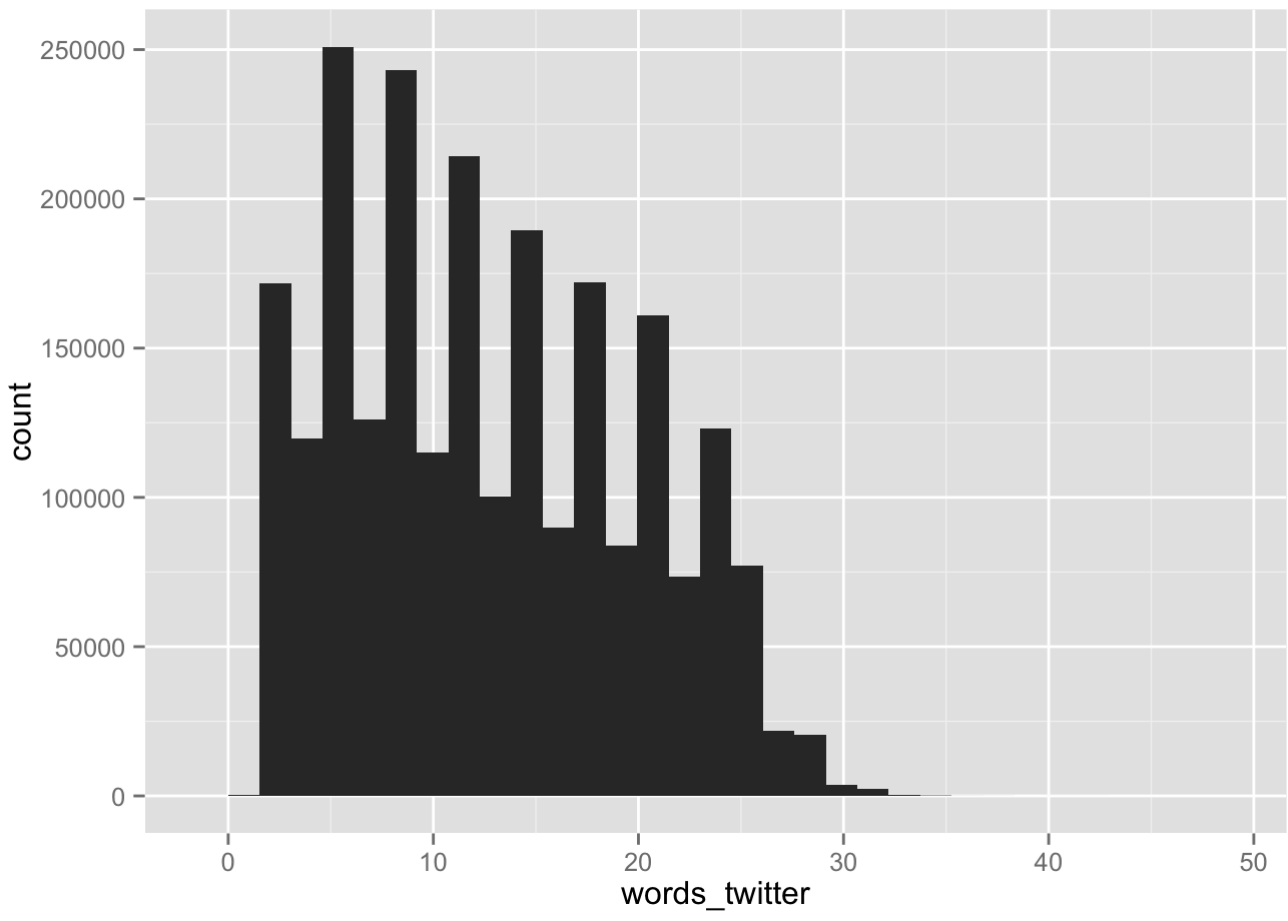
Analyze “twitter” corpus:

```
words_twitter <- stri_count_words(twitter)
summary( words_twitter )
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   12.00   12.75   18.00   47.00
```

```
qplot(   words_twitter )
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



Intermediate Conclusion for Milestone Report

The frequency distributions of the “blogs” and “news” corpora are similar (appearing to be log-normal). The frequency distribution of the “twitter” corpus is again different, as a result of the 140 character limit.

Final Conclusion for Milestone report

For final project, it will be required to work on the training predictive models using training data sets within the corpora. It will be required to compare respective models for each type of text - blogs, news & Twitter & to understand and how these perform against aggregate models trained off the entire corpus that spans the 3 types of text files.