Capstone Project

Nobodynobody

2024-7-19

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com/).

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
# import the blogs and twitter datasets in text mode
blogs <- readLines("C:/Users/fUJITSU/Desktop/final/en_US/en_US.blogs.txt", encoding="UTF-8",
skipNul = TRUE)
twitter <- readLines("C:/Users/fUJITSU/Desktop/final/en_US/en_US.twitter.txt",encoding="UTF-
8", skipNul=TRUE)
```

```
# import the news dataset in binary mode
con <- file("C:/Users/fUJITSU/Desktop/final/en_US/en_US.news.txt", open="rb")
news <- readLines(con, encoding="UTF-8", skipNul = TRUE)
close(con)
rm(con)
```

```
#load libraries
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to be
come errors
```

```
library(tidytext)
library(ggplot2)
library(stringi)
```

```
#Basic statistics by determining the file size
file.info("C:/Users/fUJITSU/Desktop/final/en_US/en_US.news.txt")$size / 1024^2
```

```
## [1] 196.2775
```

```
file.info("C:/Users/fUJITSU/Desktop/final/en_US/en_US.blogs.txt")$size   / 1024^2
```

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

```
file.info("C:/Users/fUJITSU/Desktop/final/en_US/en_US.twitter.txt")$size / 1024^2
```

```
## [1] 159.3641
```

```
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   162096241   134082806
```

```
#summary statistics of the words in the text
blogsWords <- stri_count_words(blogs)
newsWords <- stri_count_words(news)
twitterWords <- stri_count_words(twitter)
```

```
summary(blogsWords)
```

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

```
summary(newsWords)
```

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

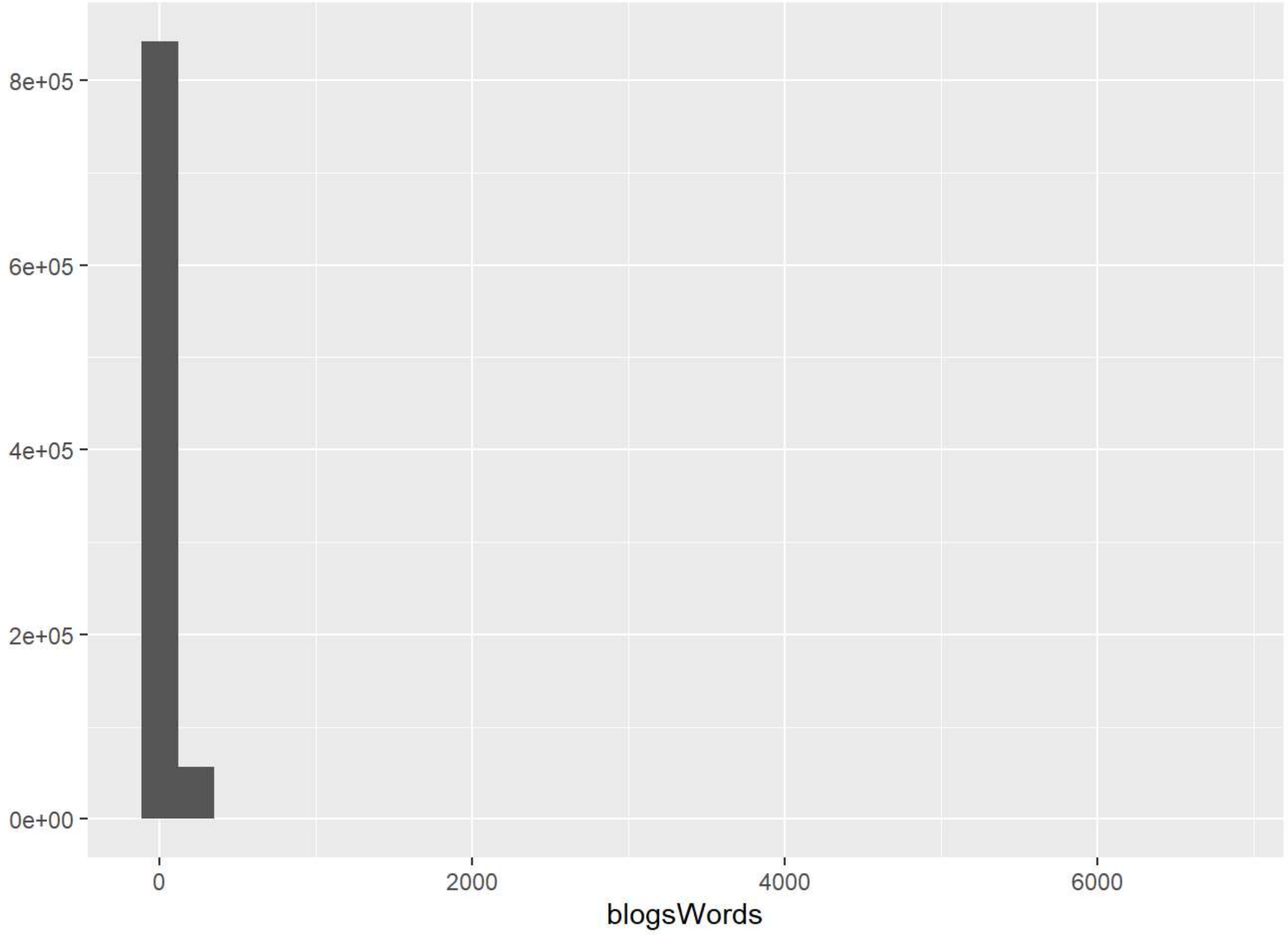
```
summary(twitterWords)
```

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

```
#plot frequecy distributions of the words in the text
qplot(blogsWords)
```

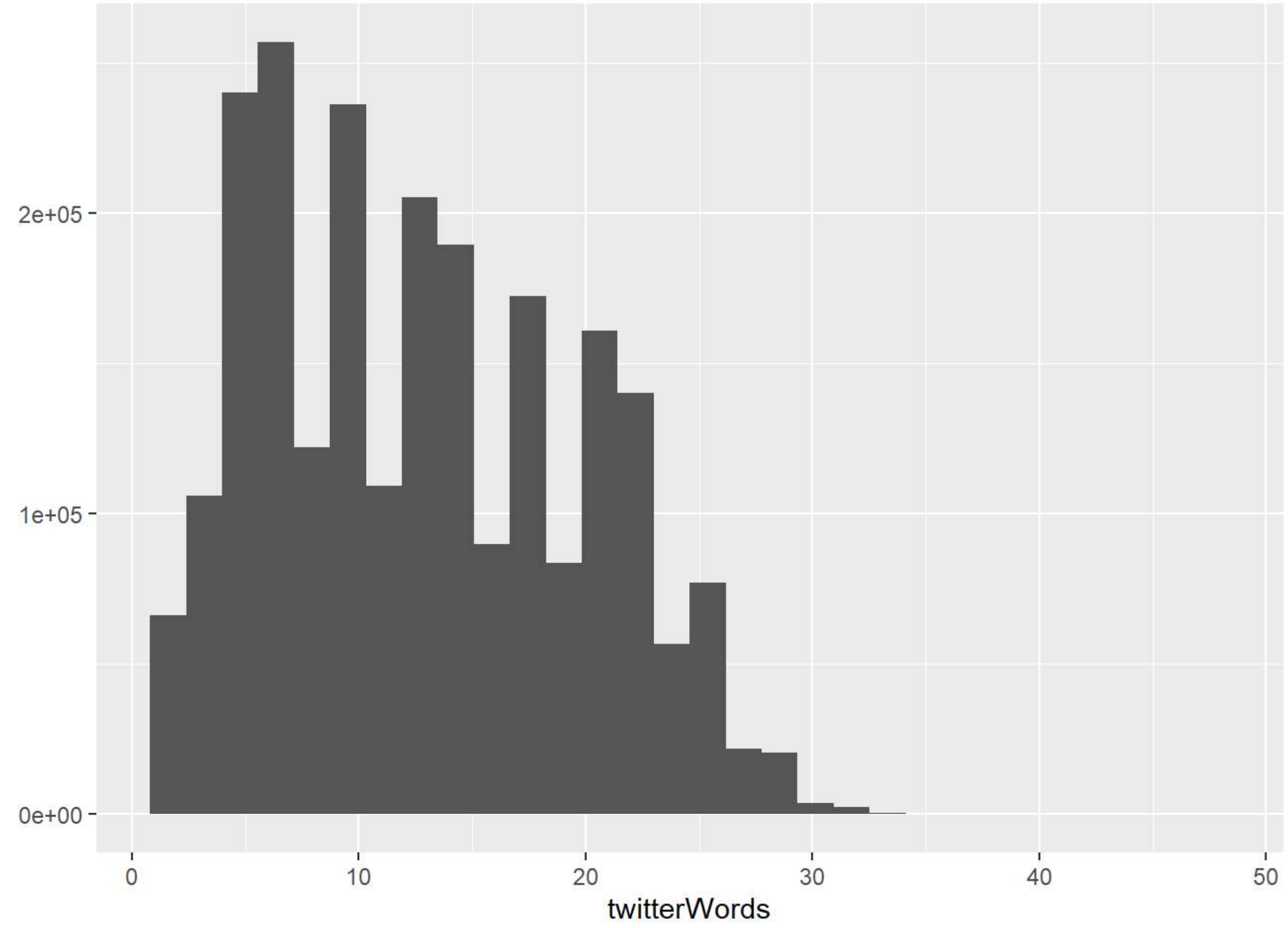
```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



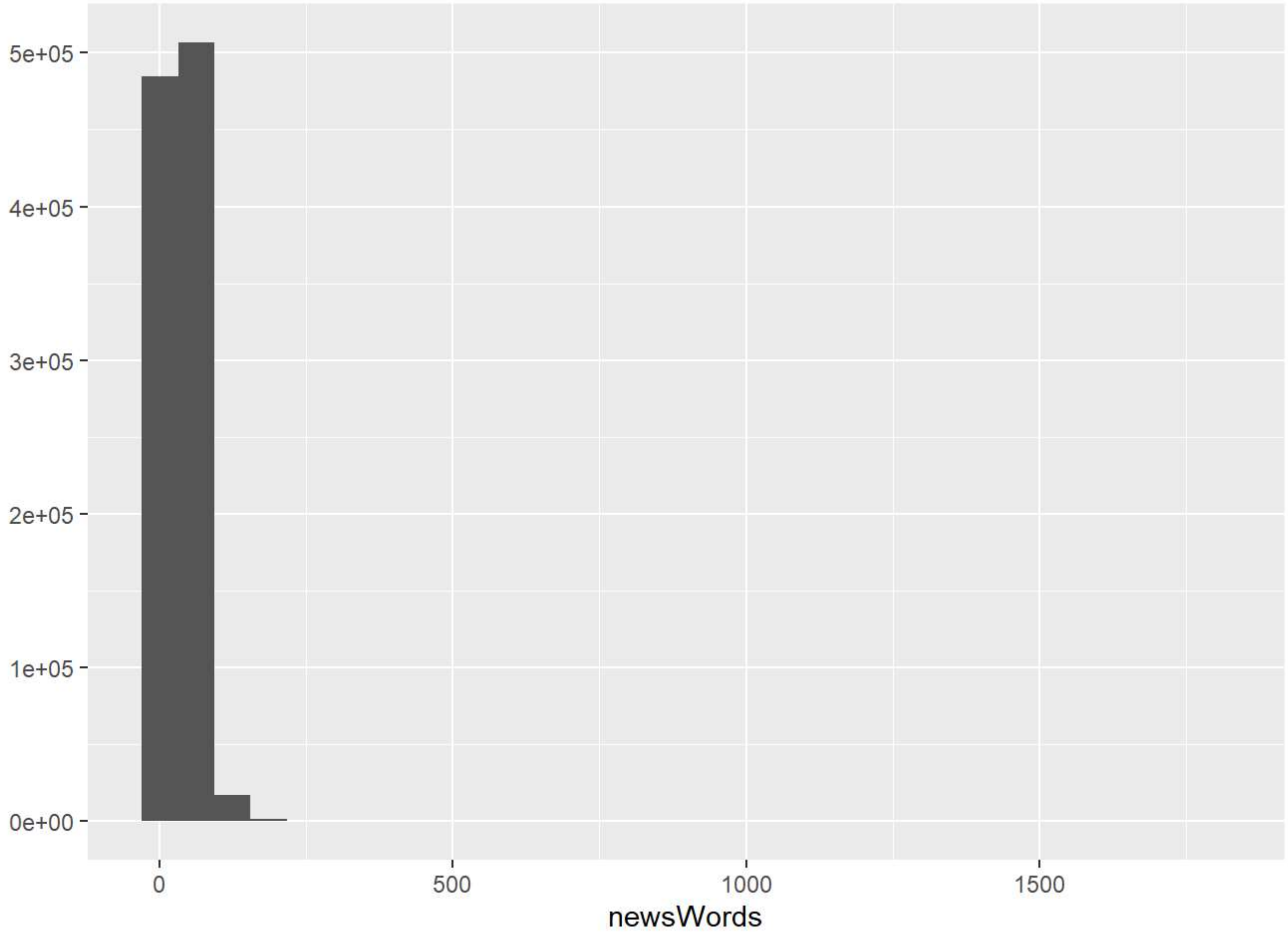
```
qplot(twitterWords)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
qplot(newsWords)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
# create samples from the text (twitter, news, blogs)
blogs_sample <-sample(blogs, 2500)
news_sample <- sample(news, 2500)
twitter_sample <-sample(twitter, 2500)
```

```
# concatenation the sample into one document
sample <- c(blogs_sample, news_sample, twitter_sample)
```

```
# load libraries
library(dplyr)
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(stringr)
library(NLP)
```

```
##
## Attaching package: 'NLP'
```

```
## The following object is masked from 'package:ggplot2':
##
##     annotate
```

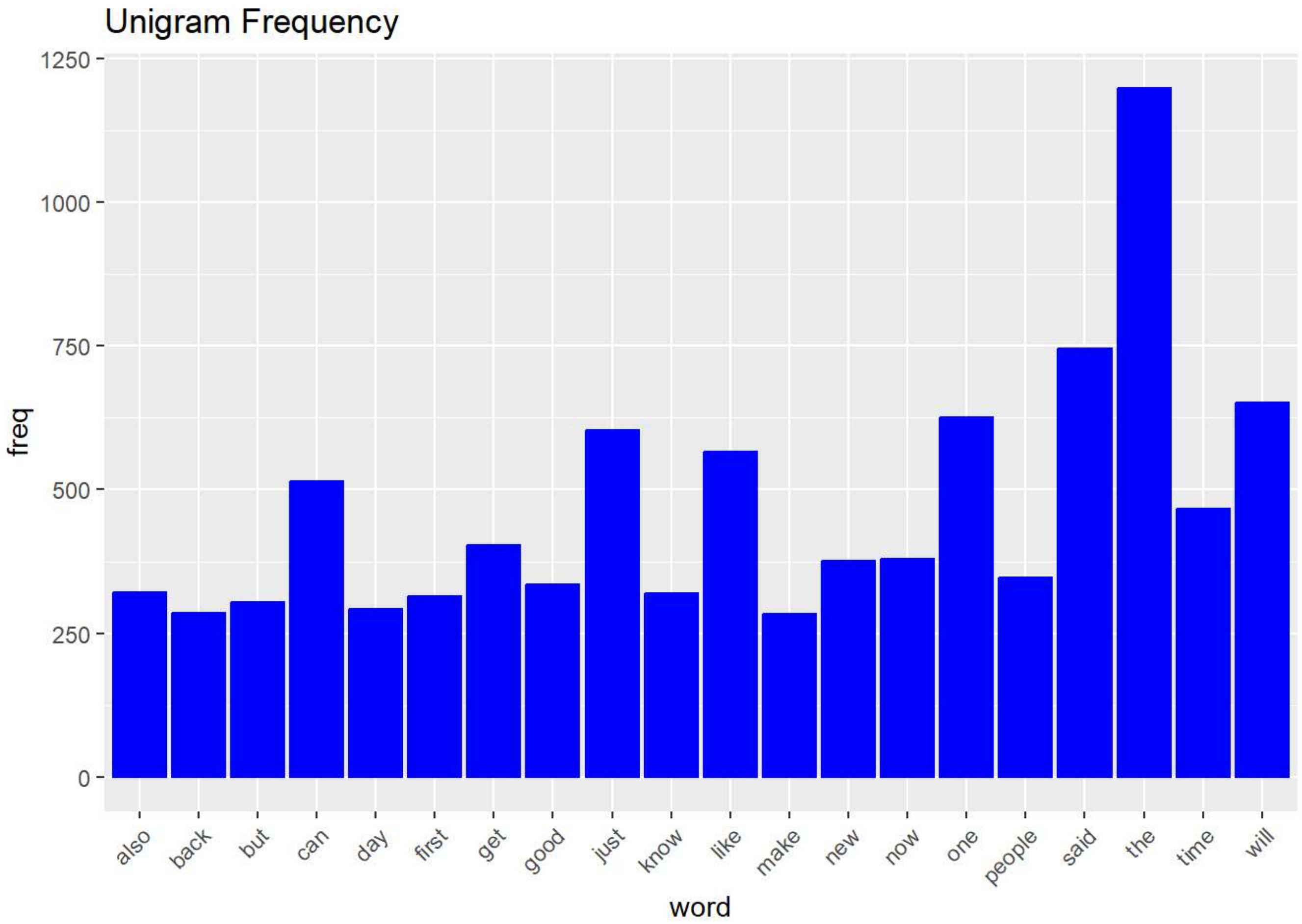
```
library(tm)
library(SnowballC)
library(knitr)
```

```
corpus <- VCorpus(VectorSource(sample))
```

```
corpus <- tm_map(corpus, removePunctuation)
corpus <- tm_map(corpus, stripWhitespace)
corpus <- tm_map(corpus, removeWords, stopwords("english"))
corpus <- tm_map(corpus, content_transformer(tolower))
corpus <- tm_map(corpus, removeNumbers)
```

```
# Analysing the text document.
dtm <- TermDocumentMatrix(corpus)
```

```
# Unigram frequency
freq <- rowSums(as.matrix(dtm))
freq <- sort(freq, decreasing = TRUE)
dfFreq <- data.frame(word = names(freq), freq=freq)
ggplot(dfFreq[1:20, ], aes(word, freq)) + geom_bar(stat="identity", fill="blue", colour="blu
e") +
  theme(axis.text.x=element_text(angle=45, hjust=1)) + ggtitle("Unigram Frequency")
```


Unigram Frequency
freq
1250
1000
750
500
250
0
word
also
back
but
can
day
first
get
good
just
know
like
make
new
now
one
people
said
the
time
will