Basic Exploratory Analysis

TARUN

26/08/2020

R Markdown

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/) 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.

```
library(ggplot2)
library(dplyr)
library(tokenizers)
library(stringi)
library(stringr)
library(tm)
library(quanteda)
```

Downloading Data

```
if(!file.exists("./Coursera-SwiftKey")){
  dir.create("./data")
  url <- "https://d396qusza40orc.cloudfront.net/dsscapstone/dataset/Coursera-SwiftKey.zip"
  download.file(Url, destfile="./data/Coursera-SwiftKey.zip", mode = "wb")
  unzip(zipfile="./data/Coursera-SwiftKey.zip", exdir="./data")
}
```

##Reading Data

```
datablog <- readLines("./Coursera-SwiftKey/final/en_US/en_US.blogs.txt")
datanews <- readLines("./Coursera-SwiftKey/final/en_US/en_US.news.txt")
datatwitter <- readLines("./Coursera-SwiftKey/final/en_US/en_US.twitter.txt")

##summary of data
stri_stats_general(datablog)
```

```
##       Lines LinesNEmpty       Chars CharsNWhite
##      899288      899288   208361438   171926076
```

```
stri_stats_general(datanews)
```

```
##       Lines LinesNEmpty       Chars CharsNWhite
##       77259       77259    15683765    13117038
```

```
stri_stats_general(datatwitter)
```

```
##       Lines LinesNEmpty       Chars CharsNWhite
##     2360148     2360148   162384825   134370864
```

##Creating Data

```
subblog <- sample(datablog,size = 1000)
subnews <- sample(datanews,size = 1000)
subtwitter <- sample(datatwitter,size = 1000)
sampledData <- c(subblog,subnews,subtwitter)
head(sampledData)
```

```
## [1] "That brings us to the second problem â\200" problem of bad science. The lack of proper scientifi
## [2] "I'm no scientist and I wouldn't want to insult anybody by attempting to go into any more depth o
## [3] "Today I attended a training session organized by the FBI on the investigation of the Peterson ho
## [4] "Now with restrictions lifted the south carolina mortgage refinancing. Retirement commission shou
## [5] "Soon his friend wanted to leave, so we exchanged numbers, and I gave him one last long kiss and
## [6] "Bromine: Is a disinfectants that can be used as an alternative for chlorine. In swimming pools,
```

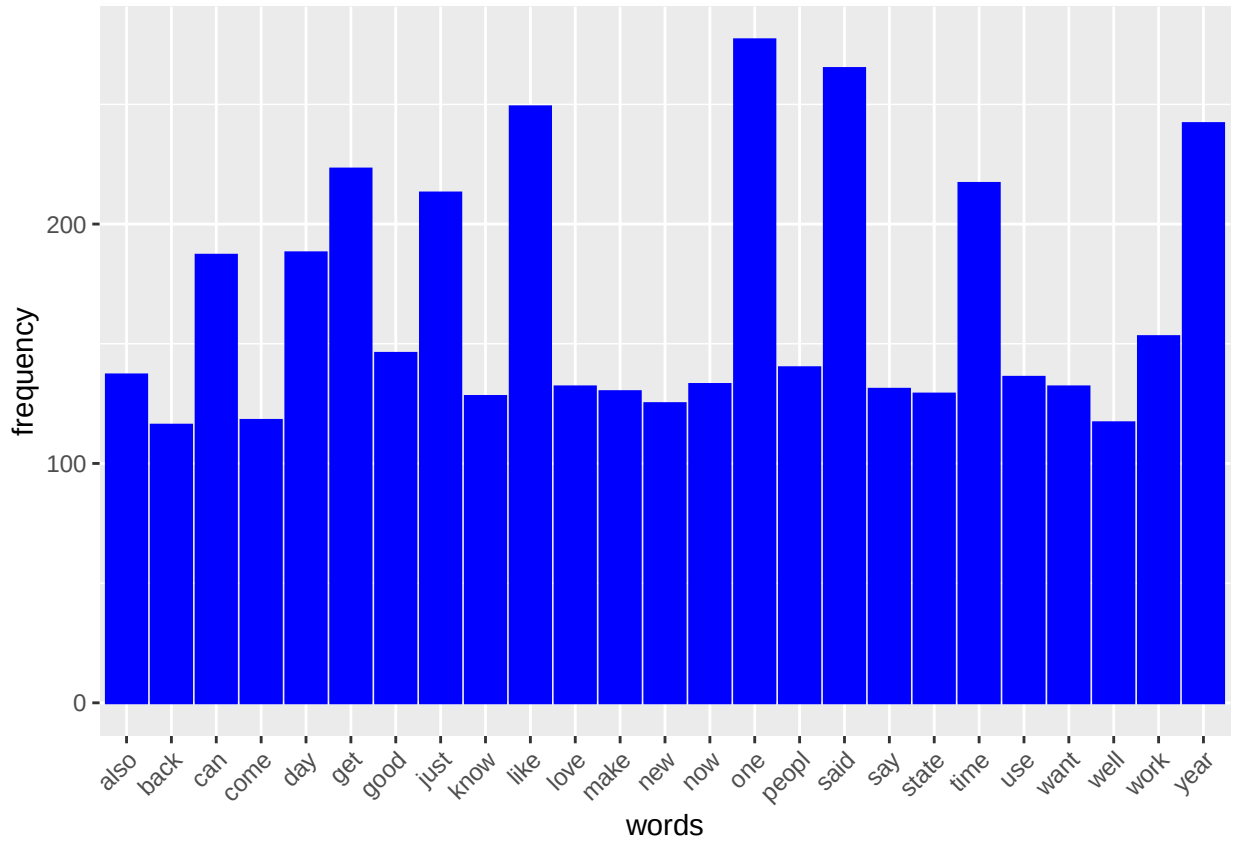
```
corpora <- VCorpus(VectorSource(sampledData))
```

##Transforming Data

```
toSpace <- content_transformer(function(x, pattern) gsub(pattern, " ", x))
corpora <- tm_map(corpora, toSpace, "/|@|//|$|:|:)|*|&|!|?|_|-|#|¦|€|â||ã|$")
corpora <- tm_map(corpora,content_transformer(tolower))
corpora <- tm_map(corpora,removePunctuation)
corpora <- tm_map(corpora,stemDocument)
corpora <- tm_map(corpora,stripWhitespace)
corpora <- tm_map(corpora,removeWords,stopwords("english"))
corpora <- tm_map(corpora,removeNumbers)
corpora <- tm_map(corpora,stripWhitespace)
```

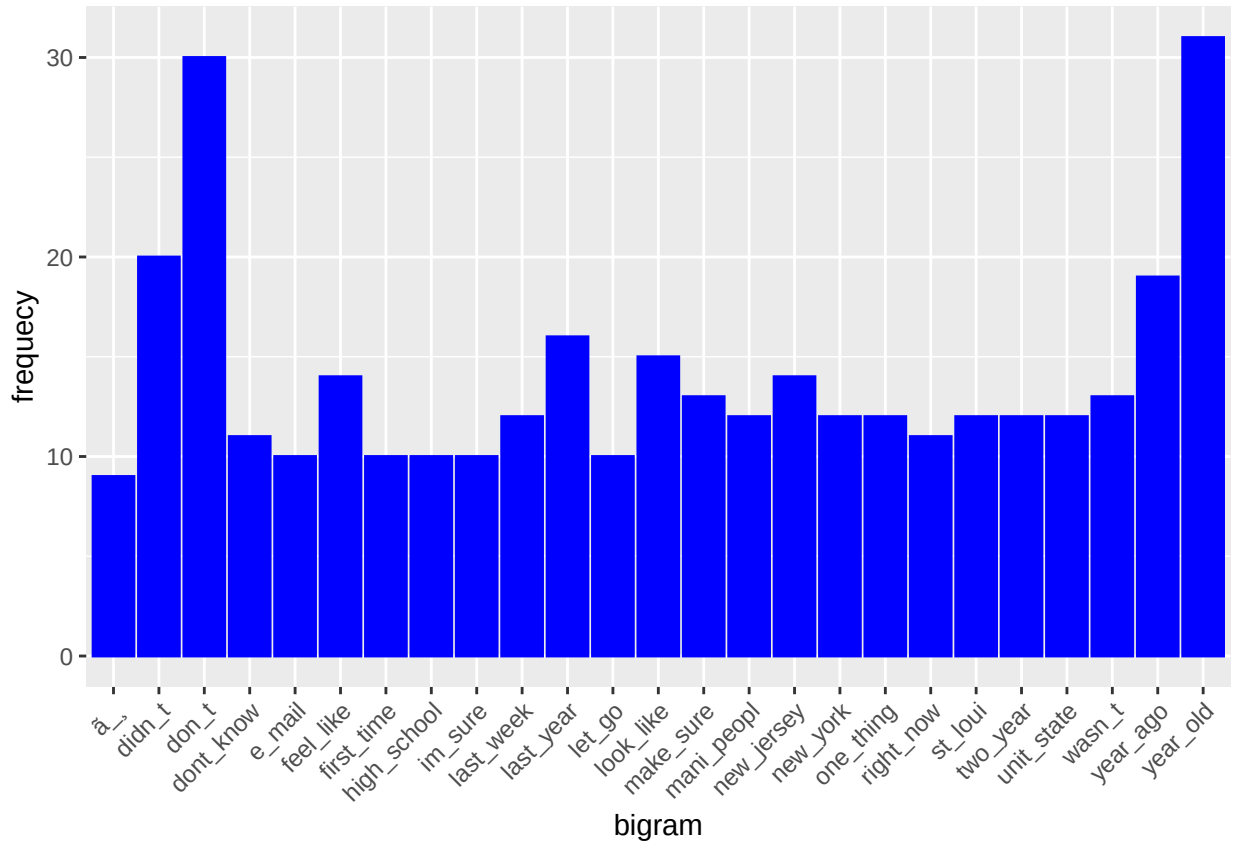
##Data Exploration

```
dtm1 <- TermDocumentMatrix(corpora)
freq <- apply(as.matrix(dtm1),1,sum)
freq <- sort(freq,decreasing = TRUE)
freqdf <- data.frame(words = names(freq),frequency = freq)
## Plot to see distribution of words
ggplot(freqdf[1:25,],aes(words,frequency)) +
  geom_bar(stat = "identity",fill = "blue",color = "blue") +
  theme(axis.text.x = element_text(angle = 45,hjust = 1))
```



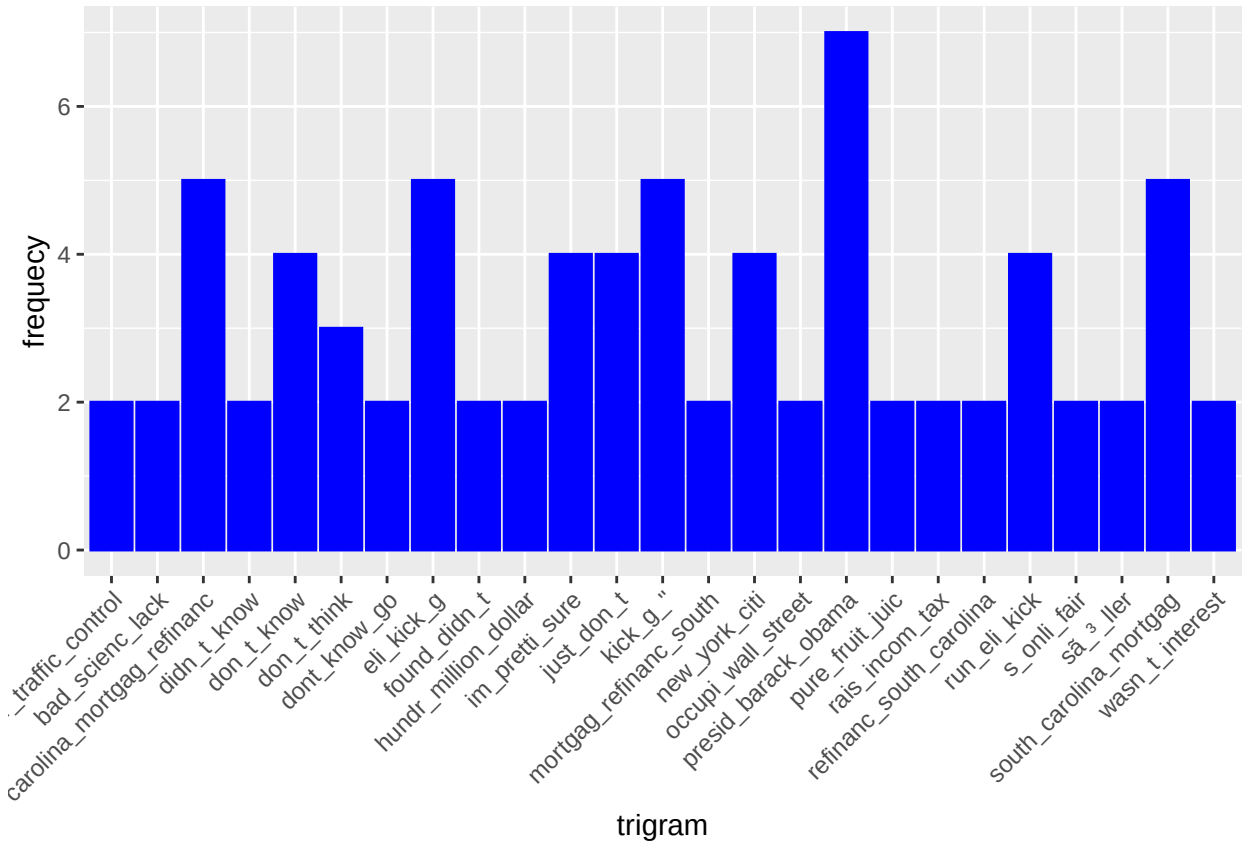

```
## Now lets try 2 gram model
corpora <- corpus(corpora)
tk <- tokens(corpora)
tk1 <- tokens_ngrams(tk)
dfm1 <- dfm(tk1)
bifreq <- colSums(dfm1)
bifreq <- sort(bifreq,decreasing = TRUE)
bifreqdf <- data.frame(bigram = names(bifreq),frequecy = bifreq)

ggplot(bifreqdf[1:25,],aes(bigram,frequecy)) +
  geom_bar(stat = "identity",fill = "blue",color = "blue") +
  theme(axis.text.x = element_text(angle = 45,hjust = 1))
```




```
## Now lets try 3 gram model
tk2 <- tokens_ngrams(tk,3)
dfm2 <- dfm(tk2)
trifreq <- colSums(dfm2)
trifreq <- sort(trifreq,decreasing = TRUE)
trifreqdf <- data.frame(trigram = names(trifreq),frequecy = trifreq)

ggplot(trifreqdf[1:25,],aes(trigram,frequecy)) +
  geom_bar(stat = "identity",fill = "blue",color = "blue") +
  theme(axis.text.x = element_text(angle = 45,hjust = 1))
```




```
## lets look at a word graph
textplot_wordcloud(dfm1,max_words = 20,color = RColorBrewer::brewer.pal(8,"Dark2"))
```

Future plans

```
#Planning to create a model using 2 gram and 3 gram models and coming up with a
#effective ml algorithm in a way it predicts with great accuracy.
```