Project4

Shariq Mian

11/15/2021

```r
library(tm)
library(knitr)
library(plyr)
library(wordcloud)
library(tidyverse)
library(tm)
library(magrittr)
library(data.table)
library(e1071)
library(caret)
library(randomForest)
```

```r
ham="~/Shariq School/SPS/Data 607/SpamHam/easy_ham"
count_ham=length(list.files(path = ham))
ham_list=list.files(ham)
count_ham
```

```
## [1] 2550
```

```r
spam= "~/Shariq School/SPS/Data 607/SpamHam/spam_2"
count_spam=length(list.files(path = spam))
spam_list=list.files(spam)
count_spam
```

```
## [1] 1397
```

```r
spam_list=list.files(spam)
ham_text = NA
for(i in 1:length(ham_list))
{
  path=paste0(ham, "/", ham_list[1])
  text =readLines(path)
  list= list(paste(text, collapse="\n"))
  ham_text = c(ham_text,list)

}

spam_text = NA
for(i in 1:length(spam_list))
{
  path=paste0(spam, "/", spam_list[1])
  text =readLines(path)
  list= list(paste(text, collapse="\n"))
  spam_text = c(spam_text,list)
}
```

```
email_body <- function(ham_text){
  message = str_split(ham_text,"\n\n") %>% unlist()
  body = paste(message[2:length(message)], collapse=' ' )
  return(body)
}

ham_text <- email_body(ham_text)
```

general filtering opts:

```
# Building a new corpus
ham_corpus =VCorpus(VectorSource(unlist(lapply(ham_text, as.character))))
ham_terms_matrix = TermDocumentMatrix(ham_corpus,control= list(removePunctuation=TRUE, removeNumbers=TRU
ham_corpus = tm_map(ham_corpus, removeNumbers)
ham_corpus = tm_map(ham_corpus, removeWords, stopwords())
ham_corpus = tm_map(ham_corpus, removePunctuation)
ham_corpus = tm_map(ham_corpus, stemDocument)
ham_corpus = tm_map(ham_corpus, stripWhitespace)
ham_terms_matrix = TermDocumentMatrix(ham_corpus)

spam_corpus= VCorpus(VectorSource(spam_text))
spam_terms_matrix= TermDocumentMatrix(spam_corpus,control=list(removePunctuation=TRUE, removeNumbers=TRU
spam_corpus = tm_map(spam_corpus, removeNumbers)
spam_corpus = tm_map(spam_corpus, removeWords, stopwords())
spam_corpus = tm_map(spam_corpus, removePunctuation)
spam_corpus = tm_map(spam_corpus, stemDocument)
spam_corpus = tm_map(spam_corpus, stripWhitespace)
spam_terms_matrix = TermDocumentMatrix(spam_corpus)
```

```
#ham_df =as.data.frame(unlist(ham_text),stringsAsFactors = FALSE)
#ham_df$type = "ham"
#colnames(ham_df) = c("text","Classification")
#ham_df


#spam_df =as.data.frame(unlist(spam_text),stringsAsFactors = FALSE)
#spam_df$type = "spam"
#colnames(spam_df) = c("text","Classification")
#spam_df
```

```
spam_df = as.data.frame(as.table(spam_terms_matrix))
spam_df$spam_ham = "SPAM"
colnames(spam_df) = c('TERM', 'SPAM_DOCS', 'SPAM_FREQ', 'TYPE_SPAM')
spam_df = subset(spam_df, select = -c(2) )
```

```r
spam_df$SPAM_FREQ[is.na(spam_df$SPAM_FREQ)] = '0'
spam_df = ddply(spam_df, .(TERM, TYPE_SPAM), summarize, SPAM_FREQ = sum(as.numeric(SPAM_FREQ)))
head(spam_df, n = 20)
```

```
##                TERM TYPE_SPAM SPAM_FREQ
## 1             above      SPAM      1397
## 2            accept      SPAM      1397
## 3               add      SPAM      1397
## 4          addrarpa      SPAM      2794
## 5           address      SPAM     11176
## 6         addressbr      SPAM      2794
## 7             again      SPAM      1397
## 8              also      SPAM      1397
## 9           america      SPAM      1397
## 10       americanbr      SPAM      1397
## 11  americanexpress      SPAM      1397
## 12              and      SPAM      5588
## 13              ani      SPAM      1397
## 14           answer      SPAM      2794
## 15           appear      SPAM      4191
## 16              are      SPAM      1397
## 17             asbr      SPAM      1397
## 18              ask      SPAM      1397
## 19           author      SPAM      1397
## 20            avail      SPAM      1397
```

```r
ham_df = as.data.frame(as.table(ham_terms_matrix))
ham_df$spam_ham = "HAM"
colnames(ham_df) = c('TERM', 'HAM_DOCS', 'HAM_FREQ', 'TYPE_HAM')
ham_df = subset(ham_df, select = -c(2) )
ham_df$HAM_FREQ[is.na(ham_df$HAM_FREQ)] = '0'
ham_df = ddply(ham_df, .(TERM, TYPE_HAM), summarize, HAM_FREQ = sum(as.numeric(HAM_FREQ)))
head(ham_df, n = 20)
```

```
##                       TERM TYPE_HAM HAM_FREQ
## 1                      abl      HAM     2550
## 2                   actual      HAM     2550
## 3                      ago      HAM     2550
## 4                      and      HAM     2550
## 5                      aug      HAM    33150
## 6                     bulk      HAM     2550
## 7                      cfd      HAM     2550
## 8           charsetusascii      HAM     2550
## 9                    chris      HAM     5100
## 10                    code      HAM     2550
## 11                    come      HAM     5100
## 12                 command      HAM     7650
## 13                  compil      HAM     2550
## 14              contenttyp      HAM     2550
## 15                   creat      HAM     2550
## 16                     cvs      HAM     2550
## 17  cwgdatedfaddeepeddycom      HAM     5100
## 18                   date      HAM     5100
## 19                    day      HAM     2550
## 20                  debug      HAM     2550
```

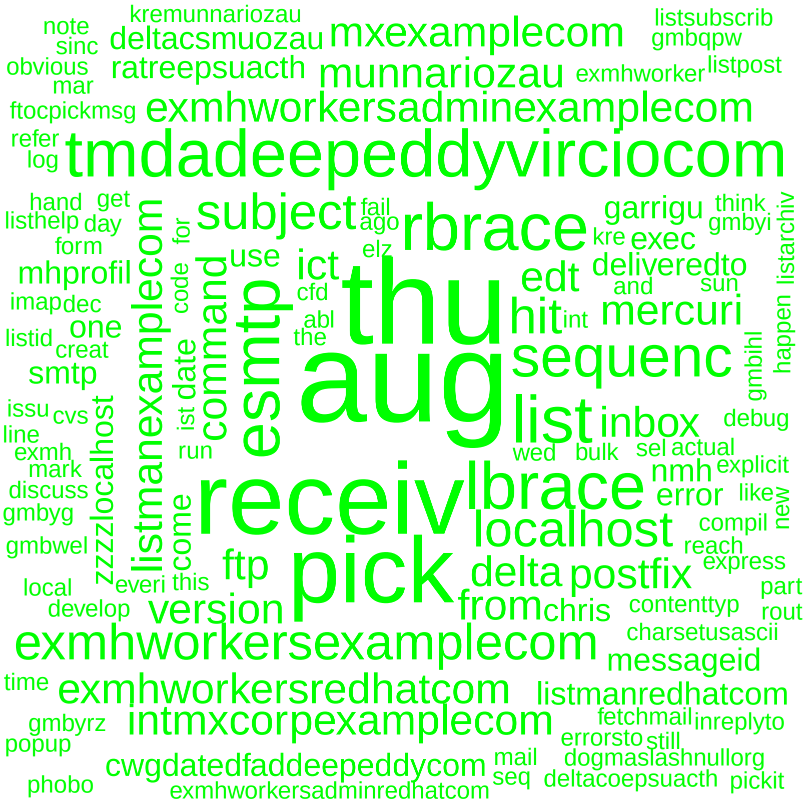
```
# Bind the data frames
spam_ham_df = merge(x = ham_df, y = spam_df, by="TERM", all = TRUE)
nrow(spam_ham_df)
```

```
## [1] 416
```

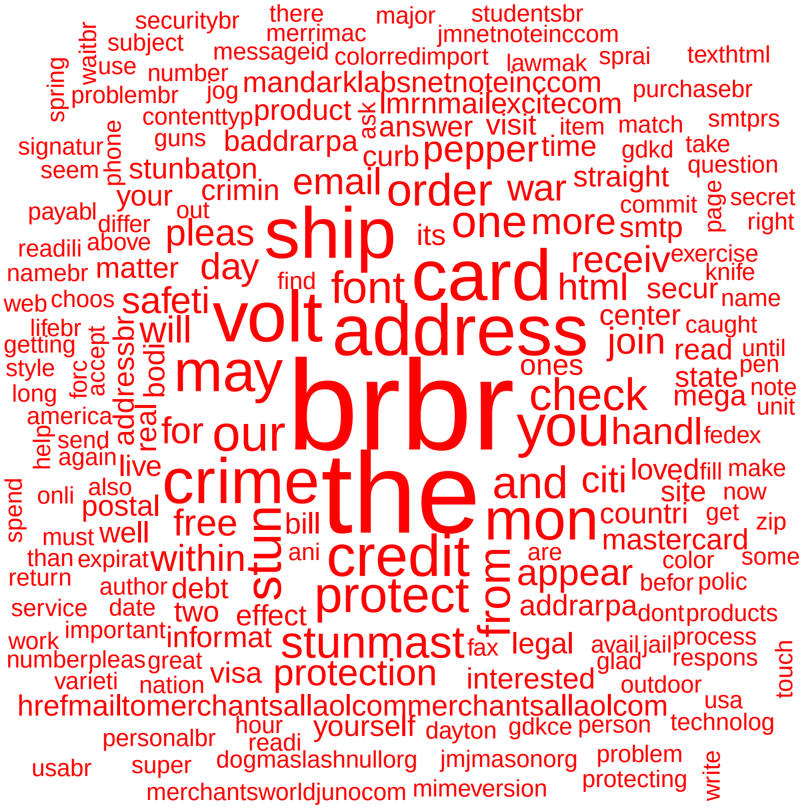
```
spam_ham_df<- spam_ham_df[sample(nrow(spam_ham_df)),]
head(spam_ham_df, n=20)
```

```
##                                               TERM TYPE_HAM HAM_FREQ TYPE_SPAM
## 44                                         explicit      HAM     2550      <NA>
## 277                               lmrnmailexcitecom     <NA>       NA      SPAM
## 245                                            gdkd     <NA>       NA      SPAM
## 132                                      returnpath      HAM     2550      SPAM
## 198                                            bodi     <NA>       NA      SPAM
## 168                                   zzzzlocalhost      HAM     5100      <NA>
## 91                                            local      HAM     2550      <NA>
## 387                                           until     <NA>       NA      SPAM
## 96                                    maillocalhost      HAM     2550      <NA>
## 265                                            jail     <NA>       NA      SPAM
## 104                                       messageid      HAM     5100      SPAM
## 394                                         varieti     <NA>       NA      SPAM
## 65  httpslistmanexamplecommailmanlistinfoexmhwork      HAM     5100      <NA>
## 97                      mailtoexmhworkersexamplecom      HAM     2550      <NA>
## 408                                           worth     <NA>       NA      SPAM
## 176                                           again     <NA>       NA      SPAM
## 296                                            more     <NA>       NA      SPAM
## 72                                              int      HAM     2550      <NA>
## 410                                        xkeyword     <NA>       NA      SPAM
## 19                                              day      HAM     2550      SPAM
##     SPAM_FREQ
## 44         NA
## 277      2794
## 245      1397
## 132      1397
## 198      2794
## 168        NA
## 91         NA
## 387      1397
## 96         NA
## 265      1397
## 104      1397
## 394      1397
## 65         NA
## 97         NA
## 408      1397
## 176      1397
## 296      4191
## 72         NA
## 410      1397
## 19       4191
```

```
wordcloud(ham_corpus, max.words = 200, random.order = FALSE, colors=c('green'))
```



```
wordcloud(spam_corpus, max.words = 200, random.order = FALSE, colors=c('red'))
```


securitybr
there
major
studentsbr
waitbr
merrimac
jmnetnoteinccom
subject
messageid
colorredimport
lawmak
sprai
texthtml
spring
use
number
jog
mandarklabsnetnoteinccom
purchasebr
problembr
lmrnmailexcitecom
contenttyp
product
ask
answer
visit
item
match
smtprs
signatur
phone
guns
baddrarpa
pepper
time
gdkd
take
seem
stunbaton
curb
straight
question
email
order
war
your
crimin
commit
secret
payabl
out
page
ship
one
more
right
differ
pleas
its
smtp
readili
above
card
receiv
exercise
namebr
matter
day
find
font
html
knife
secur
web
choos
safeti
volt
address
center
name
lifebr
will
join
caught
getting
accept
addressbr
read
until
forc
bodi
may
style
ones
pen
state
long
brbr
check
mega
note
america
real
unit
help
send
for
our
you
handl
fedex
again
live
loved
fill
make
crime
and
citi
spend
onli
also
site
now
postal
the
mon
countri
get
must
well
free
bill
zip
stun
mastercard
than
expirat
within
ani
credit
are
color
some
return
author
from
appear
befor
polic
debt
service
date
protect
addrarpa
dont
products
two
effect
work
important
informat
stunmast
legal
avail
jail
process
fax
numberpleas
great
glad
respons
visa
protection
touch
varieti
nation
interested
outdoor
hrefmailtomerchantsallaolcommerchantsallaolcom
usa
hour
yourself
dayton
gdkce
person
technolog
personalbr
readi
usabr
super
dogmaslashnullorg
jmjmasonorg
problem
write
protecting
merchantsworldjunocom
mimeversion