title: “607_Fall2021_Project4 Rev2”
author: “Mark Schmalfeld”
date: “11/12/2021”
output: html_document

```
library(tidymodels)
```

```
## Registered S3 method overwritten by 'tune':
##   method                   from
##   required_pkgs.model_spec parsnip
```

```
## -- Attaching packages -------------------------------------- tidymodels 0.1.4 --
```

```
## v broom        0.7.10     v recipes      0.1.17
## v dials        0.0.10     v rsample      0.1.1
## v dplyr        1.0.7      v tibble       3.1.6
## v ggplot2      3.3.5      v tidyr        1.1.4
## v infer        1.0.0      v tune         0.1.6
## v modeldata    0.1.1      v workflows    0.2.4
## v parsnip      0.1.7      v workflowsets 0.1.0
## v purrr        0.3.4      v yardstick    0.0.8
```

```
## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x purrr::discard() masks scales::discard()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## x recipes::step()  masks stats::step()
## * Dig deeper into tidy modeling with R at https://www.tmwr.org
```

```
library(RCurl)
```

```
##
## Attaching package: 'RCurl'
```

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

```
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following objects are masked from 'package:yardstick':
##
##     precision, recall, sensitivity, specificity
## The following object is masked from ’package:purrr’:
##
##     lift
```

```
library(NLP)
```

```
##
## Attaching package: ’NLP’

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

```
library(tidytext)
library(corpus)
library(R.utils)
```

```
## Loading required package: R.oo

## Loading required package: R.methodsS3

## R.methodsS3 v1.8.1 (2020-08-26 16:20:06 UTC) successfully loaded. See ?R.methodsS3 for help.

## R.oo v1.24.0 (2020-08-26 16:11:58 UTC) successfully loaded. See ?R.oo for help.

##
## Attaching package: ’R.oo’

## The following object is masked from ’package:R.methodsS3’:
##
##     throw

## The following object is masked from ’package:RCurl’:
##
##     clone

## The following object is masked from ’package:recipes’:
##
##     check

## The following object is masked from ’package:dials’:
##
##     finalize

## The following objects are masked from ’package:methods’:
##
##     getClasses, getMethods

## The following objects are masked from ’package:base’:
##
##     attach, detach, load, save
## R.utils v2.11.0 (2021-09-26 08:30:02 UTC) successfully loaded. See ?R.utils for help.
```

```
##
## Attaching package: 'R.utils'
```

```
## The following object is masked from 'package:RCurl':
##
##     reset
```

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

```
## The following object is masked from 'package:utils':
##
##     timestamp
```

```
## The following objects are masked from 'package:base':
##
##     cat, commandArgs, getOption, inherits, isOpen, nullfile, parse,
##     warnings
```

```
library(tm)
suppressWarnings(library(wordcloud))
```

```
## Loading required package: RColorBrewer
```

```
library(SnowballC)
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

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

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v readr   2.0.1     v forcats 0.5.1
## v stringr 1.4.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x NLP::annotate()         masks ggplot2::annotate()
## x data.table::between()   masks dplyr::between()
## x readr::col_factor()     masks scales::col_factor()
## x RCurl::complete()       masks tidyr::complete()
## x purrr::discard()        masks scales::discard()
## x R.utils::extract()      masks tidyr::extract()
## x dplyr::filter()         masks stats::filter()
## x data.table::first()     masks dplyr::first()
## x stringr::fixed()        masks recipes::fixed()
## x dplyr::lag()            masks stats::lag()
## x data.table::last()      masks dplyr::last()
## x caret::lift()           masks purrr::lift()
## x readr::spec()           masks yardstick::spec()
## x data.table::transpose() masks purrr::transpose()
```

```
library(tidyr)
library(dplyr)
library(stringr)
library(stats)
library(readtext)
library(caTools)
library(randomForest)
```

```
## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: ’randomForest’

## The following object is masked from ’package:ggplot2’:
##
##     margin

## The following object is masked from ’package:dplyr’:
##
##     combine
```

```
library(tm.plugin.webmining)
```

```
##
## Attaching package: ’tm.plugin.webmining’

## The following objects are masked from ’package:R.utils’:
##
##     extract, parse

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

```
library(tm)
library(RTextTools)
```

```
## Loading required package: SparseM
```

```
##
## Attaching package: 'SparseM'
```

```
## The following object is masked from 'package:base':
##
##     backsolve
```

```
## Registered S3 method overwritten by 'tree':
##   method     from
##   print.tree cli
```

```
##
## Attaching package: 'RTextTools'
```

```
## The following objects are masked from 'package:SnowballC':
##
##     getStemLanguages, wordStem
```

```
library(R.utils)
library(utils)
library(textmineR)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
##
## Attaching package: 'textmineR'
```

```
## The following object is masked from 'package:Matrix':
##
##     update
```

```
## The following object is masked from 'package:recipes':
##
##     update
## The following object is masked from 'package:stats':
##
##     update
```

```
library(SparseM)
```

Github data pull and unpack files

Pull spam and ham files from Github Unzipe and untar the files

```
# download and unzip spam document from github
download.file('https://raw.githubusercontent.com/schmalmr/607_Fall2021_Project_4/main/20021010_spam.tar
bunzip2("spam_zip.tar.bz2", remove = F, overwrite = T)
untar("spam_zip.tar") #set up a spam folder

# download and unzip spam document from github

download.file('https://raw.githubusercontent.com/schmalmr/607_Fall2021_Project_4/main/20030228_easy_ham_
bunzip2("ham_zip.tar.bz2", remove = F, overwrite = T)
untar("ham_zip.tar") #setup a ham folder
```

Data clean up and DF for classification of the files

Identify extraneous ham/ spam file and delete. Create ham file with class 0 as "not spam" and spam file class at 1 foro "spam" Print out the number of ham and spam files we are working with in the datasets.

```
# identify unneded spam file and delete
remove_spam <-list.files(path="spam/", full.names=T, recursive=FALSE, pattern="0000.7b1b73cf36cf9dbc3d64
file.remove(remove_spam)
```

```
## [1] TRUE
```

```
# list of spam files
spam_files <- (list.files(path="spam/", full.names=T, recursive=FALSE))

# identify extraneous ham file and delete
remove_ham <- list.files(path="easy_ham_2/", full.names=T, recursive=FALSE, pattern="cmds")
file.remove(remove_ham)
```

```
## [1] TRUE
```

```
# list of ham files
ham_files <- list.files(path="easy_ham_2/",full.names=T, recursive=FALSE)

spam<-data.frame()
spam<-as.data.frame(unlist(spam_files),stringsAsFactors = FALSE)
spam$class<-1
colnames(spam)<-c("text","class")
spam_num <- nrow(spam) # Total Number of Spam Emails
print(paste0("The Total Number of Emails in the Spam Data-Set is : ", spam_num))
```

```
## [1] "The Total Number of Emails in the Spam Data-Set is : 500"
```

```
ham<-data.frame()
ham<-as.data.frame(unlist(ham_files),stringsAsFactors = FALSE)
ham$class<-0
colnames(ham)<-c("text","class")
ham_num <- nrow(ham) # Total Number of Ham Emails
print(paste0("The Total Number of Emails in the Ham Data-Set is : ", ham_num))
```

```
## [1] "The Total Number of Emails in the Ham Data-Set is : 1400"
```

Combining the ham and spam files

Combined spam and ham files to create a ham_spam file. Shuffle the files names using "2021" as seed number. Print out the head and the team data.

```
ham_spam <- c(ham_files,spam_files)
str(ham_spam)
```

```
##  chr [1:1900] "easy_ham_2//00001.1a31cc283af0060967a233d26548a6ce" ...
```

```
#shuffle file names
set.seed(2021)
ham_spam <- sample(ham_spam,length(ham_spam))

print(paste0("The initial 15 email headers in the ham_spam Data-Set"))
```

```
## [1] "The initial 15 email headers in the ham_spam Data-Set"
```

```
head(ham_spam,15)
```

```
##  [1] "easy_ham_2//00903.ec8820827b3b3b89e471ba86d3ec88c8"
##  [2] "easy_ham_2//00166.8525e30f5b1574a4cb08d5fc8cb740e5"
##  [3] "spam//0054.839a9c0a07f13718570da944986a898a"
##  [4] "easy_ham_2//00442.38508150d13a53c035c15edb79ec4f9d"
##  [5] "easy_ham_2//00743.7787f0f8205e4ff2226a563c39b81039"
##  [6] "easy_ham_2//00908.fa150b0b994587469112fbcb7e8cc2bc"
##  [7] "easy_ham_2//01012.235d771c2e3094e4b4310a86ac7e7352"
##  [8] "easy_ham_2//01094.5bd0918274c1c243e77e44a2987b851c"
##  [9] "easy_ham_2//00192.b1c13f7caac54fca99993a3478d603d9"
## [10] "easy_ham_2//00934.76cd57955d5efc3a84d965b91fb1548e"
## [11] "spam//0387.c2b993b46377256bdcb2314c2553b6f0"
## [12] "spam//0238.7d0de37650a0c0e2d99e52eef4042602"
## [13] "easy_ham_2//00622.1d8e9e4c3e8e00a382595b6a2e6954ab"
## [14] "easy_ham_2//00325.419046d511bd4b995fdec3057ae996b1"
## [15] "easy_ham_2//00495.727ea275e5530758b79884779603b7e0"
```

```
print(paste0(""))
```

```
## [1] ""
```

```
print(paste0("The initial 15 email tails in the ham_spam Data-Set"))
```

```
## [1] "The initial 15 email tails in the ham_spam Data-Set"
```

```
# tail of 1st email
tail(ham_spam,15)
```

```
##  [1] "easy_ham_2//00038.fba603f864720b7894b7b05e6e3f93c0"
##  [2] "easy_ham_2//00863.d9ae47fc90d47d17f9765634e5950c37"
##  [3] "easy_ham_2//00654.7e84d693f6d2dc216aa501c47db607f7"
##  [4] "easy_ham_2//00995.5171f58b6df2d565f3bca02ee548e013"
##  [5] "spam//0069.a0b6cfde0e477af7f406ee756ba53826"
##  [6] "easy_ham_2//00189.959922d0363f85a2a6e7cc689b05b75c"
##  [7] "easy_ham_2//00026.1757d50d495d41e8a5eb30a2f371019c"
##  [8] "easy_ham_2//01140.67e11f7533ac73ebeb728c6fdb86eeff"
##  [9] "spam//0147.65cf30538f09402e4d1bd4aa91d9532a"
## [10] "easy_ham_2//01392.6a9e94b131381aa631022fc1b6c9bdab"
## [11] "easy_ham_2//01076.8a6274c2c970dc8f0a72c34b67a1475b"
## [12] "easy_ham_2//00298.516883ac42f693de96cc953cf59d720b"
## [13] "easy_ham_2//00133.035335262a159fef46fff7499a8ac28f"
## [14] "spam//0222.6ad799703d958681d6e427762f86f179"
## [15] "easy_ham_2//00216.53a04d271ae7b0752fef521c2d5709f7"
```

Corpus creation and clean up

Create corpus vector source Convert to lower case Remove numbers Remove stop-words Remove white space

```
corpus = VCorpus(VectorSource(ham_spam))
corpus = tm_map(corpus, content_transformer(tolower))
corpus = tm_map(corpus, removeNumbers)
corpus = tm_map(corpus, removePunctuation)
corpus = tm_map(corpus, removeWords, stopwords())
corpus = tm_map(corpus, stemDocument)
corpus = tm_map(corpus, stripWhitespace)
corpus = tm_map(corpus,stemDocument)

wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per=0.15, min.freq=5, colors = brewer.pal(
```

```
## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spamccccdaeddec could not be fit on page. It will not be plotted.
```

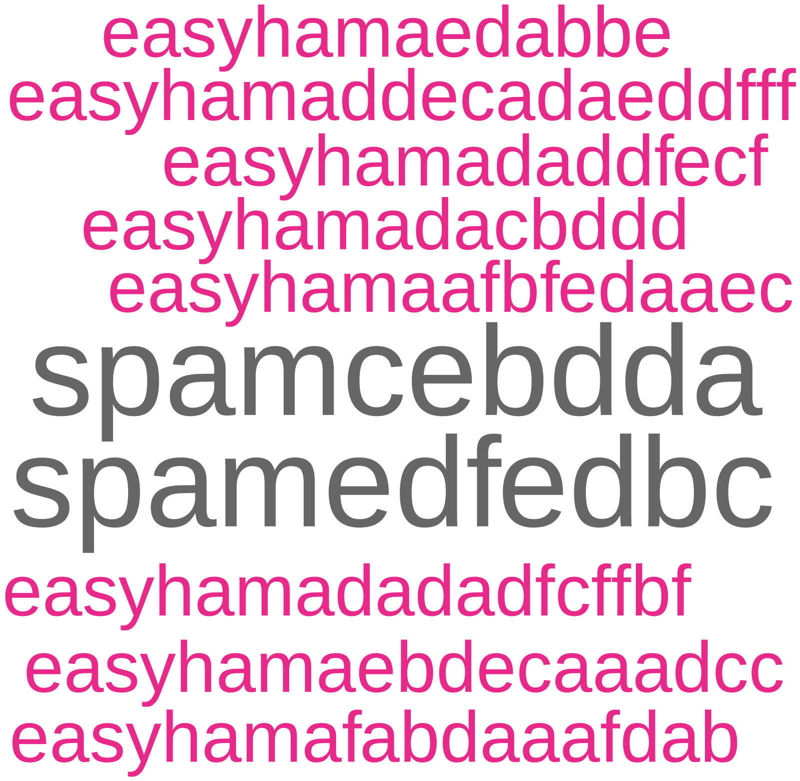
```
## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spamecbabbddfbdf could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamabdefcdcfbcda could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamadbdbdebceedfa could not be fit on page. It will not be plotted.
## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamaebcdfddcd could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per
## = 0.15, : easyhamaecdafebdefbacdc could not be fit on page. It will not be
## plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamaeecccdddddbc could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamafdcdeada could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamafefefcaeaddb could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhambaaadbbfdeaadfd could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhambbacdcba could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhambdcdffbaea could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per
## = 0.15, : easyhambfaeabbeefebeadeb could not be fit on page. It will not be
## plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamcbadcfdbfbcb could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamcbddebecbab could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamccadbfccddfd could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamccbffffcedc could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamcccfafebaebd could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamccdbeedfcabb could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamcdddbebf could not be fit on page. It will not be plotted.
## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamcefdebbdffdef could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamcfccbdfdcf could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamcfcebcdfdead could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamdaffaffebbb could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamdbafbacbee could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamdbfaeeeaff could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamdcbaddecdbfdeec could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamdccebbabdfcac could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per
## = 0.15, : easyhamdcdbedbeabffcdbcf could not be fit on page. It will not be
## plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamdcdceafd could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per
## = 0.15, : easyhamdcfdaaabffabcfea could not be fit on page. It will not be
## plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamddbdadcdfc could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamddcfebcedcdbad could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamddfafddfefbbb could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamdedabbaf could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamdeddaebfdbbeb could not be fit on page. It will not be plotted.
## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamdeffbbacdddb could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamdfcfbdaaeafaeff could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhameabfbfbbbfeea could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamebacabdebeaaeb could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamedaebfbd could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhameeacdec could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhameeddadabbeafeb could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamefafbcefdbeac could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamefbaffaefff could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamefbcfefd could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamefbecbdaeffc could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamefbfaeffcaafb could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamefbfccea could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamfaaafbffcfca could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamfacdbcbdddfdd could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamfadbbdfefbecff could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamfadbbfaccca could not be fit on page. It will not be plotted.
## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamfadebceccacd could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamfaefddbcbab could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamfbbcadceabebcb could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamfbdfecdedcdfbd could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per
## = 0.15, : easyhamfbeecaffefaffbca could not be fit on page. It will not be
## plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamfccfefeecffc could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamfcfdadddfcdfad could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamfeafcfbfbfdcbbf could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamfedacbfbaafa could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamffacccebddfd could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamffbacfeaeefa could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamffebcffefaffdb could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : easyhamffeffacb could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spamadcaaabbdfdcbd could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spamafabecad could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spambbbceaedaecda could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spambeaecacceedaa could not be fit on page. It will not be plotted.
## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spambffcaebefbabfb could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spamceaabefdbcc could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spamcefbfbdbcaaacda could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spamcfafffedfbf could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spamcfddaddffedfb could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spamcfebedaaaaab could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spamcffedbdaada could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spamdabcddf could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spamdccbebedeb could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spamddfbfeeeeda could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spamdfcbfdc could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spameabdebdfdfef could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spamebfdbabbefec could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spamebfecafaceca could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spamedadcadefacd could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spameddeaedb could not be fit on page. It will not be plotted.

## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spamededfdcfc could not be fit on page. It will not be plotted.
## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spameefbeeedbdcc could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spamefdcddbbdfaafa could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(corpus, max.words = 100, random.order = FALSE, rot.per =
## 0.15, : spamfcfdedaceaea could not be fit on page. It will not be plotted.
```



## Document Term Matrix setup

Create Document Term Matrix remove sparse terms

```
dtm<- DocumentTermMatrix(corpus)
dtm<- removeSparseTerms(dtm, 0.90)
dtm
```

```
## <<DocumentTermMatrix (documents: 1900, terms: 0)>>
## Non-/sparse entries: 0/0
## Sparsity           : 100%
## Maximal term length: 0
## Weighting          : term frequency (tf)
```

Work on model classifier

Work incompleted

{r Model size and setup for testing}

number of emails in corpus

N <- length(spam_files) N # set up model container; 70/30 split between train and test data #container <- create_container(dtm, labels = (spam_file$text), trainSize = 1:(0.7*N), testSize = (0.7*N+1):N, virgin = FALSE)

svm_model <- train_model(container, "SVM") #tree_model <- train_model(container, "TREE") #maxent_model <- train_model(container, "MAXENT")

svm_out <- classify_model(container, svm_model) #tree_out <- classify_model(container, tree_model) #maxent_out <- classify_model(container, maxent_model)

{r Model size and setup for testing}

Resources

The data wrangling methods used are representative of various methods covered in previous readings and exercises. However, the building of the classifier models draws significantly on the outlined procedures in Chapter 10 of Automated Data Collection with R, with particular emphasis on pages 310-312. Here is the full citation:

Munzert, Simon et. "Chapter 10: Statistical Text Processing." Automated Data Collection with R: a Practical Guide to Web Scraping and Text Mining, 1st ed., John Wiley & Sons Ltd., Chichester, UK, 2015, pp. 295-321.