Ejercicio4

Javier

3/11/2020

```
library(tm) #text mining package from R community, tm_map(), content_transformer()
```

```
## Warning: package 'tm' was built under R version 4.0.3
```

```
## Loading required package: NLP
```

```
## Warning: package 'NLP' was built under R version 4.0.3
```

```
library(SnowballC) #used for stemming, wordStem(), stemDocument()
```

```
## Warning: package 'SnowballC' was built under R version 4.0.3
```

```
library(RColorBrewer)
```

```
## Warning: package 'RColorBrewer' was built under R version 4.0.3
```

```
library(wordcloud) #wordcloud generator
```

```
## Warning: package 'wordcloud' was built under R version 4.0.3
```

```
library(e1071) #Naive Bayes
```

```
## Warning: package 'e1071' was built under R version 4.0.3
```

```
library(gmodels) #CrossTable()
```

```
## Warning: package 'gmodels' was built under R version 4.0.3
```

```
library(caret) #ConfusionMatrix()
```

```
## Warning: package 'caret' was built under R version 4.0.3
```

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

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.0.3
```

```
##
## Attaching package: 'ggplot2'
```

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

```r
library(klaR)
```

```
## Warning: package 'klaR' was built under R version 4.0.3
```

```
## Loading required package: MASS
```

```r
file <- read.csv(file = "sms_spam.csv", header = TRUE, sep = ",",
encoding="UTF-8")
file$type <- as.factor(file$type)

#Haremos la nube de spam y de ham por separados
#Mas adelante se mostrara una mezcla entre los dos
spam <- subset(file, type == "spam")
wordcloud(spam$text,scale = c (5.5,1.5), min.freq = 15, max.words = 60,
rot.per = 0.2,colors = brewer.pal(8,"Dark2"))
```

```
## Warning in tm_map.SimpleCorpus(corpus, tm::removePunctuation):
transformation
## drops documents
```

```
## Warning in tm_map.SimpleCorpus(corpus, function(x) tm::removeWords(x,
## tm::stopwords())): transformation drops documents
```

```
## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : code could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : claim could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : every could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : please could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : guaranteed could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : service could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : tone could not be fit on page. It will not be plotted.
```

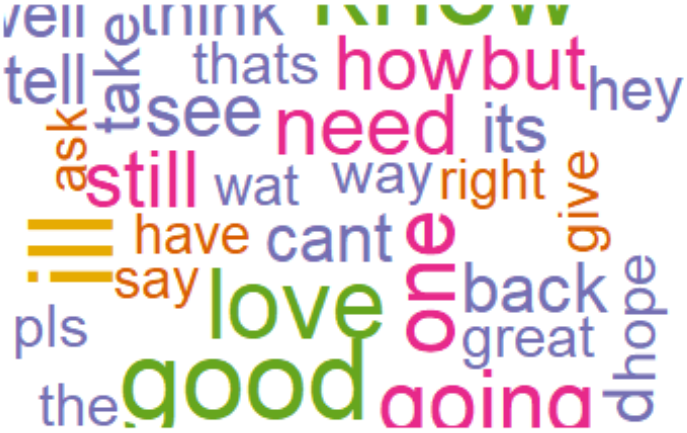
```
## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : new could not be fit on page. It will not be plotted.

## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : free could not be fit on page. It will not be plotted.

## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : message could not be fit on page. It will not be plotted.

## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : chat could not be fit on page. It will not be plotted.

## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : get could not be fit on page. It will not be plotted.

## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : your could not be fit on page. It will not be plotted.

## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : £1000 could not be fit on page. It will not be plotted.

## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : cash could not be fit on page. It will not be plotted.

## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : camera could not be fit on page. It will not be plotted.

## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : draw could not be fit on page. It will not be plotted.

## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : apply could not be fit on page. It will not be plotted.

## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : txt could not be fit on page. It will not be plotted.

## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : landline could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : shows could not be fit on page. It will not be plotted.

## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : £2000 could not be fit on page. It will not be plotted.

## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : prize could not be fit on page. It will not be plotted.

## Warning in wordcloud(spam$text, scale = c(5.5, 1.5), min.freq = 15,
max.words =
## 60, : video could not be fit on page. It will not be plotted.
```


```
ham <- subset(file, type == "ham")
wordcloud(ham$text,scale = c (5.5,0.5), min.freq = 15, max.words = 60,
rot.per = 0.2,colors = brewer.pal(8,"Dark2"))

## Warning in tm_map.SimpleCorpus(corpus, tm::removePunctuation):
transformation
## drops documents

## Warning in tm_map.SimpleCorpus(corpus, function(x) tm::removeWords(x,
## tm::stopwords())): transformation drops documents
```

```
## Warning in wordcloud(ham$text, scale = c(5.5, 0.5), min.freq = 15,
max.words =
## 60, : today could not be fit on page. It will not be plotted.

## Warning in wordcloud(ham$text, scale = c(5.5, 0.5), min.freq = 15,
max.words =
## 60, : time could not be fit on page. It will not be plotted.

## Warning in wordcloud(ham$text, scale = c(5.5, 0.5), min.freq = 15,
max.words =
## 60, : will could not be fit on page. It will not be plotted.

## Warning in wordcloud(ham$text, scale = c(5.5, 0.5), min.freq = 15,
max.words =
## 60, : already could not be fit on page. It will not be plotted.

## Warning in wordcloud(ham$text, scale = c(5.5, 0.5), min.freq = 15,
max.words =
## 60, : come could not be fit on page. It will not be plotted.

## Warning in wordcloud(ham$text, scale = c(5.5, 0.5), min.freq = 15,
max.words =
## 60, : you could not be fit on page. It will not be plotted.

## Warning in wordcloud(ham$text, scale = c(5.5, 0.5), min.freq = 15,
max.words =
## 60, : ltgt could not be fit on page. It will not be plotted.

## Warning in wordcloud(ham$text, scale = c(5.5, 0.5), min.freq = 15,
max.words =
## 60, : now could not be fit on page. It will not be plotted.

## Warning in wordcloud(ham$text, scale = c(5.5, 0.5), min.freq = 15,
max.words =
## 60, : dont could not be fit on page. It will not be plotted.

## Warning in wordcloud(ham$text, scale = c(5.5, 0.5), min.freq = 15,
max.words =
## 60, : can could not be fit on page. It will not be plotted.

## Warning in wordcloud(ham$text, scale = c(5.5, 0.5), min.freq = 15,
max.words =
## 60, : got could not be fit on page. It will not be plotted.

## Warning in wordcloud(ham$text, scale = c(5.5, 0.5), min.freq = 15,
max.words =
## 60, : sorry could not be fit on page. It will not be plotted.

## Warning in wordcloud(ham$text, scale = c(5.5, 0.5), min.freq = 15,
max.words =
## 60, : just could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(ham$text, scale = c(5.5, 0.5), min.freq = 15,
max.words =
## 60, : home could not be fit on page. It will not be plotted.

## Warning in wordcloud(ham$text, scale = c(5.5, 0.5), min.freq = 15,
max.words =
## 60, : later could not be fit on page. It will not be plotted.

## Warning in wordcloud(ham$text, scale = c(5.5, 0.5), min.freq = 15,
max.words =
## 60, : want could not be fit on page. It will not be plotted.

## Warning in wordcloud(ham$text, scale = c(5.5, 0.5), min.freq = 15,
max.words =
## 60, : call could not be fit on page. It will not be plotted.

## Warning in wordcloud(ham$text, scale = c(5.5, 0.5), min.freq = 15,
max.words =
## 60, : get could not be fit on page. It will not be plotted.
```

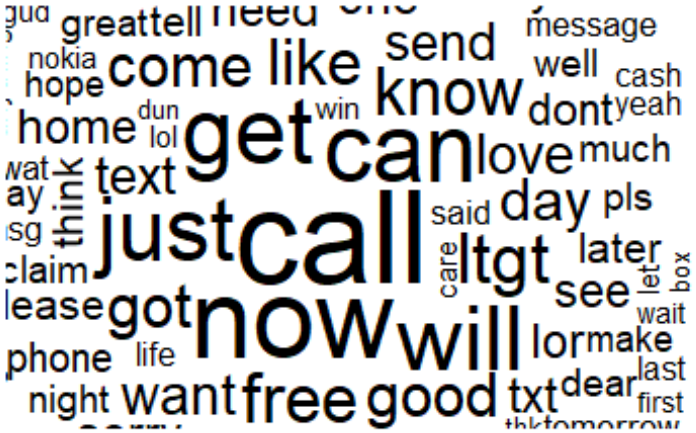



```
corpus <- VCorpus(VectorSource(file$text))


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

```r
#Imprime la nube con la mezcla de las palabras
wordcloud(clean_corpus, min.freq = 50,random.order = FALSE)
```

```
## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE): someone
## could not be fit on page. It will not be plotted.

## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE): money
## could not be fit on page. It will not be plotted.

## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE): tonight
## could not be fit on page. It will not be plotted.

## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE): many
## could not be fit on page. It will not be plotted.

## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE): service
## could not be fit on page. It will not be plotted.

## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE):
## customer could not be fit on page. It will not be plotted.

## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE): friends
## could not be fit on page. It will not be plotted.

## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE): gonna
## could not be fit on page. It will not be plotted.

## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE): always
## could not be fit on page. It will not be plotted.

## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE): leave
## could not be fit on page. It will not be plotted.

## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE): nice
## could not be fit on page. It will not be plotted.

## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE): wan
## could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE): sleep
## could not be fit on page. It will not be plotted.

## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE): things
## could not be fit on page. It will not be plotted.

## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE): waiting
## could not be fit on page. It will not be plotted.

## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE): name
## could not be fit on page. It will not be plotted.

## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE): wish
## could not be fit on page. It will not be plotted.

## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE): hello
## could not be fit on page. It will not be plotted.

## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE): coming
## could not be fit on page. It will not be plotted.

## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE): getting
## could not be fit on page. It will not be plotted.

## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE):
## guaranteed could not be fit on page. It will not be plotted.

## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE): mins
## could not be fit on page. It will not be plotted.

## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE): people
## could not be fit on page. It will not be plotted.

## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE): year
## could not be fit on page. It will not be plotted.

## Warning in wordcloud(clean_corpus, min.freq = 50, random.order =
FALSE): yet
## could not be fit on page. It will not be plotted.
```




```
dtm <- DocumentTermMatrix(clean_corpus)
#Separamos los datos con un 75% y un 25%
raw_train <- file[1:4180,]
raw_test <- file[4181:5574,]
dtm_train <- dtm[1:4180,]
dtm_test <- dtm[4181:5574,]

words <- findFreqTerms(dtm_train,5)
freq_train <- dtm_train[,words]
freq_test <- dtm_test[,words]

word <- function(x){
  x <- ifelse(x>0, "Yes","No")
}
train <- apply(freq_train,2,word)
test <- apply(freq_test,2,word)

bayes_model <- naiveBayes(train, raw_train$type,laplace = 1)


pred_bayes <- predict(bayes_model, test)
confusionMatrix(pred_bayes,raw_test$type,positive = "spam")

## Confusion Matrix and Statistics
##
##           Reference
```

```
## Prediction  ham spam
##       ham  1190   10
##       spam   22  172
##
##                Accuracy : 0.977
##                  95% CI : (0.9677, 0.9842)
##     No Information Rate : 0.8694
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.9016
##
##  Mcnemar's Test P-Value : 0.05183
##
##             Sensitivity : 0.9451
##             Specificity : 0.9818
##          Pos Pred Value : 0.8866
##          Neg Pred Value : 0.9917
##              Prevalence : 0.1306
##          Detection Rate : 0.1234
##    Detection Prevalence : 0.1392
##       Balanced Accuracy : 0.9635
##
##        'Positive' Class : spam
##
```