lab6.R

sathvik

2021-02-27

```
# 171EC146
# Sathvik S Prabhu

# Dataset: Twitter Airline Sentiment

library(wordcloud)
```

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

```
library(tm)
```

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

```
library(e1071)
library(gmodels)

tweets_raw<-read.csv("/home/sathvik/EC8/ML/Lab/Lab6/Tweets.csv")
tweets<-data.frame("type"=tweets_raw$airline_sentiment, "text"=tweets_raw$text)
str(tweets)
```

```
## 'data.frame':    14640 obs. of  2 variables:
##  $ type: Factor w/ 3 levels "negative","neutral",..: 2 3 2 1 1 1 3 2 3 3 ...
##  $ text: Factor w/ 14427 levels ",@USAirways 2nd time this occurred in 3 weeks. I'm not patient. I ha
```

```
tweets[1,]
```

```
##      type                                text
## 1 neutral @VirginAmerica What @dhepburn said.
```

```
table(tweets$type)
```

```
##
## negative  neutral positive
##     9178     3099     2363
```

```
tweets[1:3,]
```

```
##       type
## 1  neutral
## 2 positive
## 3  neutral
##                                                                text
## 1                                 @VirginAmerica What @dhepburn said.
## 2 @VirginAmerica plus you've added commercials to the experience... tacky.
## 3  @VirginAmerica I didn't today... Must mean I need to take another trip!
```

```r
tweets_corpus <- Corpus(VectorSource(tweets$text))
print(tweets_corpus)
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 14640
```

```r
inspect(tweets_corpus[1:3])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 3
##
## [1] @VirginAmerica What @dhepburn said.
## [2] @VirginAmerica plus you've added commercials to the experience... tacky.
## [3] @VirginAmerica I didn't today... Must mean I need to take another trip!
```

```r
tweets[1:3,]
```

```
##       type
## 1  neutral
## 2 positive
## 3  neutral
##                                                                 text
## 1                                  @VirginAmerica What @dhepburn said.
## 2 @VirginAmerica plus you've added commercials to the experience... tacky.
## 3  @VirginAmerica I didn't today... Must mean I need to take another trip!
```

```r
corpus_clean <- tm_map(tweets_corpus, tolower)
```

```
## Warning in tm_map.SimpleCorpus(tweets_corpus, tolower): transformation drops
## documents
```

```r
corpus_clean <- tm_map(corpus_clean, removeNumbers)
```

```
## Warning in tm_map.SimpleCorpus(corpus_clean, removeNumbers): transformation
## drops documents
```

```r
inspect(corpus_clean[1:3])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 3
##
## [1] @virginamerica what @dhepburn said.
## [2] @virginamerica plus you've added commercials to the experience... tacky.
## [3] @virginamerica i didn't today... must mean i need to take another trip!
```

```r
corpus_clean <- tm_map(corpus_clean, removeWords, stopwords())
```

```
## Warning in tm_map.SimpleCorpus(corpus_clean, removeWords, stopwords()):
## transformation drops documents
```

```r
corpus_clean <- tm_map(corpus_clean, removePunctuation)
```

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

```r
corpus_clean <- tm_map(corpus_clean, stripWhitespace)
```

```
## Warning in tm_map.SimpleCorpus(corpus_clean, stripWhitespace): transformation
## drops documents
```

```r
inspect(corpus_clean[1:3])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 3
##
## [1] virginamerica dhepburn said
## [2] virginamerica plus added commercials experience tacky
## [3] virginamerica today must mean need take another trip
```

```r
tweets_dtm <- DocumentTermMatrix(corpus_clean)
tweets_dtm
```

```
## <<DocumentTermMatrix (documents: 14640, terms: 14147)>>
## Non-/sparse entries: 138074/206974006
## Sparsity           : 100%
## Maximal term length: 46
## Weighting          : term frequency (tf)
```

```r
# Data preparation: Training and testing sets
# size=14640
# about 80:20 ratio
tweets_raw_train <- tweets[1:12000, ]
tweets_raw_test<- tweets[12001:14640, ]
tweets_dtm_train <- tweets_dtm[1:12000, ]
tweets_dtm_test<- tweets_dtm[12001:14640, ]
tweets_corpus_train <- corpus_clean[1:12000]
tweets_corpus_test<- corpus_clean[12001:14640]

prop.table(table(tweets_raw_train$type))
```

```
##
##  negative   neutral  positive
## 0.6070833 0.2220000 0.1709167
```

```r
prop.table(table(tweets_raw_test$type))
```

```
##
##  negative   neutral  positive
## 0.7170455 0.1647727 0.1181818
```

```r
# Visualizing text data - word clouds
wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE)
```

```
## Warning in strwidth(words[i], cex = size[i], ...): conversion failure on
## '"jetblue' in 'mbcsToSbcs': dot substituted for <e2>
```

```
## Warning in strwidth(words[i], cex = size[i], ...): conversion failure on
## '"jetblue' in 'mbcsToSbcs': dot substituted for <80>
```

```
## Warning in strwidth(words[i], cex = size[i], ...): conversion failure on
## '"jetblue' in 'mbcsToSbcs': dot substituted for <9c>
```

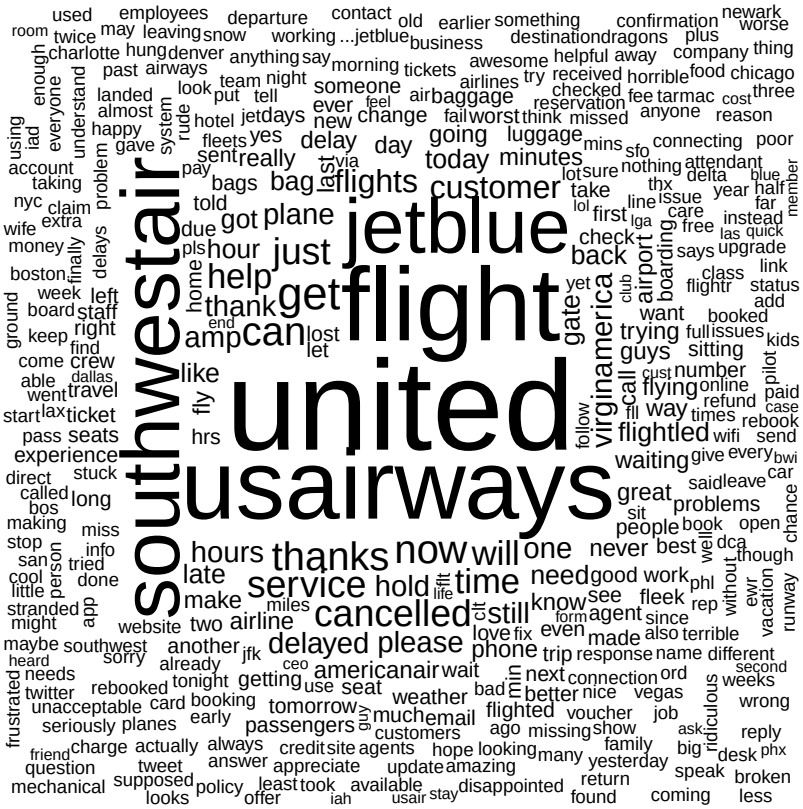
```
## Warning in text.default(x1, y1, words[i], cex = size[i], offset = 0, srt =
```

```
## rotWord * : conversion failure on '"jetblue' in 'mbcsToSbcs': dot substituted
## for <e2>

## Warning in text.default(x1, y1, words[i], cex = size[i], offset = 0, srt =
## rotWord * : conversion failure on '"jetblue' in 'mbcsToSbcs': dot substituted
## for <80>

## Warning in text.default(x1, y1, words[i], cex = size[i], offset = 0, srt =
## rotWord * : conversion failure on '"jetblue' in 'mbcsToSbcs': dot substituted
## for <9c>

## Warning in text.default(x1, y1, words[i], cex = size[i], offset = 0, srt =
## rotWord * : font metrics unknown for Unicode character U+201c

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## changed could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## talk could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## point could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## around could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## destination could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## makes could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## message could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## soon could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## option could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## houston could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## philly could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## hey could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## longer could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## given could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## asked could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## together could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## row could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## traveling could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## landing could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## swa could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## waited could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## calling could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## possible could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## minute could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## hear could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## terminal could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## arrived could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## points could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## scheduled could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## seems could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## carry could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## awful could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## real could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## idea could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## mean could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## forward could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## guess could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## things could not be fit on page. It will not be plotted.
```

```
## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## everything could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## sucks could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## international could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## attendants could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## unitedairlines could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## maintenance could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## imaginedragons could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## happened could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## glad could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## reservations could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## hard could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## years could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## hoping could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## loyal could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## else could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## frustrating could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## reflight could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## deal could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## pilots could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## saying could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## telling could not be fit on page. It will not be plotted.
```

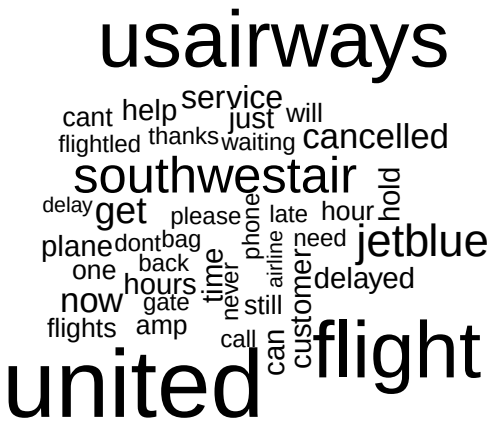
```
## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## monday could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## error could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## group could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## american could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## arrive could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## counting could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## currently could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## request could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## thought could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## believe could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## information could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## companion could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## wanted could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## standby could not be fit on page. It will not be plotted.

## Warning in wordcloud(tweets_corpus_train, min.freq = 40, random.order = FALSE):
## boarded could not be fit on page. It will not be plotted.
```




```
negative <- subset(tweets_raw_train, type == "negative")
neutral <- subset(tweets_raw_train, type == "neutral")
positive <- subset(tweets_raw_train, type == "positive")

wordcloud(negative$text, max.words = 40,scale = c(3, 0.5))
```

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

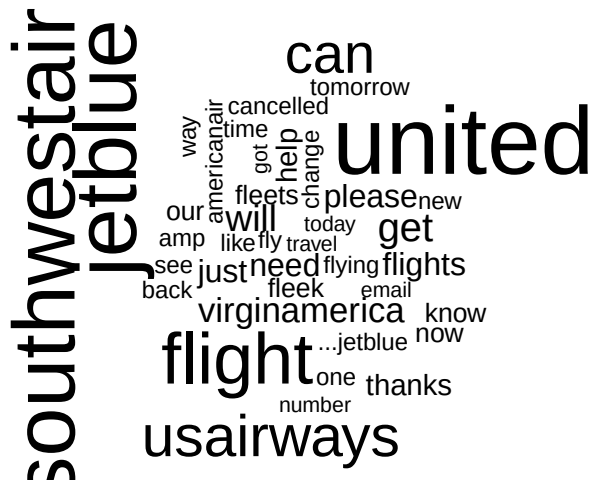
```
## Warning in tm_map.SimpleCorpus(corpus, function(x) tm::removeWords(x,
## tm::stopwords())): transformation drops documents
```




```
wordcloud(neutral$text, max.words = 40, scale = c(3, 0.5))
```

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

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

## Warning in strwidth(words[i], cex = size[i], ...): conversion failure on
## '"jetblue' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in strwidth(words[i], cex = size[i], ...): conversion failure on
## '"jetblue' in 'mbcsToSbcs': dot substituted for <80>

## Warning in strwidth(words[i], cex = size[i], ...): conversion failure on
## '"jetblue' in 'mbcsToSbcs': dot substituted for <9c>

## Warning in text.default(x1, y1, words[i], cex = size[i], offset = 0, srt =
## rotWord * : conversion failure on '"jetblue' in 'mbcsToSbcs': dot substituted
## for <e2>

## Warning in text.default(x1, y1, words[i], cex = size[i], offset = 0, srt =
## rotWord * : conversion failure on '"jetblue' in 'mbcsToSbcs': dot substituted
## for <80>

## Warning in text.default(x1, y1, words[i], cex = size[i], offset = 0, srt =
## rotWord * : conversion failure on '"jetblue' in 'mbcsToSbcs': dot substituted
## for <9c>

## Warning in text.default(x1, y1, words[i], cex = size[i], offset = 0, srt =
## rotWord * : font metrics unknown for Unicode character U+201c
```




```
wordcloud(positive$text, max.words = 40, scale = c(3, 0.5))
```

```
## 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
```


```
#Data preparation - creating indicator features for frequent words
# findFreqTerms(tweets_dtm_train, 5)

Dictionary <- function(x) {
  if( is.character(x) ) {
    return (x)
  }
  stop('x is not a character vector')
}
tweets_dict <- Dictionary(findFreqTerms(tweets_dtm_train, 5))
# 2449 fequent words

tweets_train <- DocumentTermMatrix(tweets_corpus_train, list(dictionary = tweets_dict))
tweets_test<- DocumentTermMatrix(tweets_corpus_test,list(dictionary = tweets_dict))

convert_counts <- function(x) {
  x <- ifelse(x > 0, 1, 0)
  x <- factor(x, levels = c(0, 1), labels = c("No", "Yes"))
  return(x)
}

tweets_train <- apply(tweets_train, MARGIN = 2, convert_counts)
tweets_test<- apply(tweets_test, MARGIN = 2, convert_counts)

# Step 3 - training a model on the data
tweets_classifier <- naiveBayes(tweets_train, tweets_raw_train$type)

# Step 4 - evaluating model performance
tweets_test_pred <- predict(tweets_classifier, tweets_test)
CrossTable(tweets_test_pred, tweets_raw_test$type,
           prop.chisq = FALSE, prop.t = FALSE,
           dnn = c('predicted', 'actual'))
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
```

```
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  2640
##
##
##              | actual
##    predicted |  negative |   neutral |  positive | Row Total |
## -------------|-----------|-----------|-----------|-----------|
##     negative |      1618 |       130 |        43 |      1791 |
##              |     0.903 |     0.073 |     0.024 |     0.678 |
##              |     0.855 |     0.299 |     0.138 |           |
## -------------|-----------|-----------|-----------|-----------|
##      neutral |       188 |       255 |        33 |       476 |
##              |     0.395 |     0.536 |     0.069 |     0.180 |
##              |     0.099 |     0.586 |     0.106 |           |
## -------------|-----------|-----------|-----------|-----------|
##     positive |        87 |        50 |       236 |       373 |
##              |     0.233 |     0.134 |     0.633 |     0.141 |
##              |     0.046 |     0.115 |     0.756 |           |
## -------------|-----------|-----------|-----------|-----------|
## Column Total |      1893 |       435 |       312 |      2640 |
##              |     0.717 |     0.165 |     0.118 |           |
## -------------|-----------|-----------|-----------|-----------|
##
##
```

```
# Accuracy: 79.89%

# Step 5 - improving model performance
# Using the laplace smoothing parameter.
tweets_classifier2 <- naiveBayes(tweets_train, tweets_raw_train$type, laplace = 1)
tweets_test_pred2 <- predict(tweets_classifier2, tweets_test)
CrossTable(tweets_test_pred2, tweets_raw_test$type,
           prop.chisq = FALSE, prop.t = FALSE,
           dnn = c('predicted', 'actual'))
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  2640
##
##
##              | actual
##    predicted |  negative |   neutral |  positive | Row Total |
## -------------|-----------|-----------|-----------|-----------|
##     negative |      1676 |       155 |        53 |      1884 |
```

```
##              |     0.890 |     0.082 |     0.028 |     0.714 |
##              |     0.885 |     0.356 |     0.170 |           |
## -------------|-----------|-----------|-----------|-----------|
##      neutral |       160 |       246 |        38 |       444 |
##              |     0.360 |     0.554 |     0.086 |     0.168 |
##              |     0.085 |     0.566 |     0.122 |           |
## -------------|-----------|-----------|-----------|-----------|
##     positive |        57 |        34 |       221 |       312 |
##              |     0.183 |     0.109 |     0.708 |     0.118 |
##              |     0.030 |     0.078 |     0.708 |           |
## -------------|-----------|-----------|-----------|-----------|
## Column Total |      1893 |       435 |       312 |      2640 |
##              |     0.717 |     0.165 |     0.118 |           |
## -------------|-----------|-----------|-----------|-----------|
##
##
```

```
# Accuracy: 81.17%. Up by 1.3%
```