Survey Data Analysis Script

Paula McMahon

August 2019

```r
knitr::opts_chunk$set(echo = TRUE)
library(tm)
library(SnowballC)
library(wordcloud)
library(RColorBrewer)
library(RWeka)
library(readxl)
```

Word Cloud (single word analysis using ALL survey comments)

Text Mining

```r
# The text we wish to analyse was scraped from one field of the excel sheet and copied
# into a text file.
filePath <- "~/college/ST606_Project/scripts/sentiment_all.txt"
text <- readLines(filePath)
# Load the data as a corpus
docs <- Corpus(VectorSource(text))

## Text Transformation
# Replacing special characters /, @ and | with a space
toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
docs <- tm_map(docs, toSpace, "/")
docs <- tm_map(docs, toSpace, "@")
docs <- tm_map(docs, toSpace, "\\|")

## Cleaning the text
# Convert the text to lower case
docs <- tm_map(docs, content_transformer(tolower))
# Remove numbers
docs <- tm_map(docs, removeNumbers)
# Remove english common stopwords
docs <- tm_map(docs, removeWords, stopwords("english"))
# Remove your own stop word
# specify your stopwords as a character vector
docs <- tm_map(docs, removeWords, c("supervalu", "supervalue"))
# Remove punctuations
docs <- tm_map(docs, removePunctuation)
# Eliminate extra white spaces
docs <- tm_map(docs, stripWhitespace)
```

Build a term document matrix

```
tdm = TermDocumentMatrix(docs,
                        control = list(removePunctuation = TRUE,
                                       stopwords =  TRUE,
                                       removeNumbers = TRUE,
                                       tolower = TRUE))
m <- as.matrix(tdm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 10)
```

Generate the word cloud

```
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
          max.words=200, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```

Word Cloud (bi-word analysis using ALL survey comments)

```
## Generate a word cloud comprising of the most popular two-word phrases

# docs
docs <- Corpus(VectorSource(text))

# Convert the text to lower case
docs <- tm_map(docs, content_transformer(tolower))
# Remove numbers
docs <- tm_map(docs, removeNumbers)
# Remove english common stopwords
docs <- tm_map(docs, removeWords, stopwords("english"))
# Remove your own stop word
# specify your stopwords as a character vector
docs <- tm_map(docs, removeWords, c("supervalu", "supervalue"))
# Remove punctuations
docs <- tm_map(docs, removePunctuation)
# Eliminate extra white spaces
docs <- tm_map(docs, stripWhitespace)

minfreq_bigram <- 2
tokendelim <- " \\t\\r\\n,!?,;\"()"
bitoken <- NGramTokenizer(docs, Weka_control(min=2, max=2, delimiters=tokendelim))
two_word <- data.frame(table(bitoken))
sort_two <- two_word[order(two_word$Freq, decreasing=TRUE),]

wordcloud(sort_two$bitoken, sort_two$Freq[c(1:120)],
          random.order=FALSE, scale=c(2,0.35),
          min.freq=minfreq_bigram, colors = brewer.pal(8, "Dark2"))
```

Word Cloud (single or bi-word analysis, rating 5 or less)

```r
filePath <- "~/college/ST606_Project/scripts/sentiment_rating_5_or_less.txt"
text <- readLines(filePath)
# Load the data as a corpus
docs <- Corpus(VectorSource(text))

# Convert the text to lower case
docs <- tm_map(docs, content_transformer(tolower))
# Remove numbers
docs <- tm_map(docs, removeNumbers)
# Remove english common stopwords
docs <- tm_map(docs, removeWords, stopwords("english"))
# Remove your own stop word
# specify your stopwords as a character vector
docs <- tm_map(docs, removeWords, c("supervalu", "supervalue"))
# Remove punctuations
docs <- tm_map(docs, removePunctuation)
# Eliminate extra white spaces
docs <- tm_map(docs, stripWhitespace)

minfreq_bigram <- 1
tokendelim <- " \\t\\r\\n,!?,;\"()"
bitoken <- NGramTokenizer(docs, Weka_control(min=1, max=2, delimiters=tokendelim))
two_word <- data.frame(table(bitoken))
sort_two <- two_word[order(two_word$Freq, decreasing=TRUE),]

wordcloud(sort_two$bitoken, sort_two$Freq[c(1:120)],
          random.order=FALSE, scale=c(3,0.35),
          min.freq=minfreq_bigram, colors = brewer.pal(8, "Dark2"))
```

Generate histogram of customer ratings

```r
survey_customers <- read_excel("~/college/ST606_Project/data_files/survey-responses.xlsx")

hist(survey_customers$Q34,
     main="On a 1-10 scale, are customers likely to recommend SuperValu?",
     xlab="Rating given by surveyed customers on their online shopping experience",
     col="cornflowerblue")
```