Project Part 2

Neha Parulekar

# Analysis of Text Content

install.packages("dplyr"); install.packages("tm"); install.packages("Snowball"); install.packages("wordcloud"); install.packages("cluster");

require(dplyr)

## Loading required package: dplyr

## Warning: package 'dplyr' was built under R version 3.2.5

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

require(tm)

## Loading required package: tm

## Loading required package: NLP

require(SnowballC)

## Loading required package: SnowballC

require(wordcloud)

## Loading required package: wordcloud

## Warning: package 'wordcloud' was built under R version 3.2.5

## Loading required package: RColorBrewer

require(cluster)

## Loading required package: cluster

# loading the data

getwd()

## [1] "C:/Users/Neha/Desktop"

setwd("C:/Users/Neha/Desktop")  
data = read.csv('Tweets.csv')  
data = select(data, airline\_sentiment, negativereason, airline, text)  
head(data)

## airline\_sentiment negativereason airline  
## 1 neutral Virgin America  
## 2 positive Virgin America  
## 3 neutral Virgin America  
## 4 negative Bad Flight Virgin America  
## 5 negative Can't Tell Virgin America  
## 6 negative Can't Tell Virgin America  
## 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!  
## 4 @VirginAmerica it's really aggressive to blast obnoxious "entertainment" in your guests' faces &amp; they have little recourse  
## 5 @VirginAmerica and it's a really big bad thing about it  
## 6 @VirginAmerica seriously would pay $30 a flight for seats that didn't have this playing.\nit's really the only bad thing about flying VA

# Removing the @   
   
data$text <- gsub("^@\\w+ \*", "", data$text)  
head(data)

## airline\_sentiment negativereason airline  
## 1 neutral Virgin America  
## 2 positive Virgin America  
## 3 neutral Virgin America  
## 4 negative Bad Flight Virgin America  
## 5 negative Can't Tell Virgin America  
## 6 negative Can't Tell Virgin America  
## text  
## 1 What @dhepburn said.  
## 2 plus you've added commercials to the experience... tacky.  
## 3 I didn't today... Must mean I need to take another trip!  
## 4 it's really aggressive to blast obnoxious "entertainment" in your guests' faces &amp; they have little recourse  
## 5 and it's a really big bad thing about it  
## 6 seriously would pay $30 a flight for seats that didn't have this playing.\nit's really the only bad thing about flying VA

# Dividing tweets based on positive and negative sentiment  
  
posTweets <- subset(data, airline\_sentiment == 'positive')  
dim(posTweets)

## [1] 2363 4

NegTweets <- subset(data, airline\_sentiment == 'negative')  
dim(NegTweets)

## [1] 9178 4

# Removing these words seemed to be necessary as they are repeated a lot   
wordsToRemove = c('get', 'cant', 'can', 'now', 'just', 'will', 'dont', 'ive', 'got', 'much')  
  
# analyse corpus   
analyseText = function(text\_to\_analyse){  
 CorpusTranscript = Corpus(VectorSource(text\_to\_analyse))  
 CorpusTranscript = tm\_map(CorpusTranscript, content\_transformer(tolower), lazy = T)  
 CorpusTranscript = tm\_map(CorpusTranscript, PlainTextDocument, lazy = T)  
 CorpusTranscript = tm\_map(CorpusTranscript, removePunctuation)  
 CorpusTranscript = tm\_map(CorpusTranscript, removeWords, wordsToRemove)   
 CorpusTranscript = tm\_map(CorpusTranscript, removeWords, stopwords("english"))  
 CorpusTranscript = DocumentTermMatrix(CorpusTranscript)  
 CorpusTranscript = removeSparseTerms(CorpusTranscript, 0.97) # keeps a matrix 97% sparse  
 CorpusTranscript = as.data.frame(as.matrix(CorpusTranscript))  
 colnames(CorpusTranscript) = make.names(colnames(CorpusTranscript))  
   
 return(CorpusTranscript)  
}  
  
# Analysing positive and Negative tweets  
Nword <- analyseText(NegTweets$text)  
dim(Nword)

## [1] 9178 30

Pword <- analyseText(posTweets$text)  
dim(Pword)

## [1] 2363 18

# Determining the Frequency of negative words and creating a word cloud  
  
Freq\_Nword <- colSums(Nword)  
Freq\_Nword <- Freq\_Nword[order(Freq\_Nword, decreasing = T)]  
head(Freq\_Nword)

## flight cancelled service hours help hold   
## 2900 920 740 644 610 607

wordcloud(freq = as.vector(Freq\_Nword), words = names(Freq\_Nword), random.order = FALSE, random.color = FALSE, colors = brewer.pal(9, 'RdPu')[4:9] )

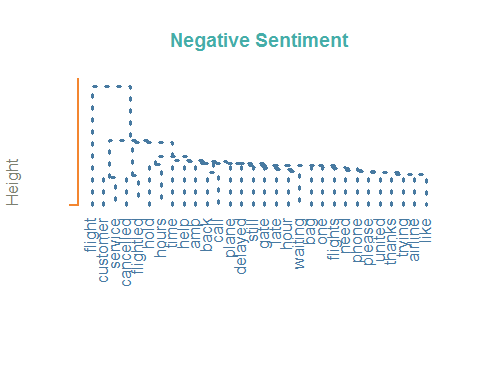


# Analysing Negative words generally mentioned in tweets   
  
analyseText2 = function(text\_to\_analyse){  
   
 CorpusTranscript = Corpus(VectorSource(text\_to\_analyse))  
 CorpusTranscript = tm\_map(CorpusTranscript, content\_transformer(tolower), lazy = T)  
 CorpusTranscript = tm\_map(CorpusTranscript, PlainTextDocument, lazy = T)  
 CorpusTranscript = tm\_map(CorpusTranscript, removePunctuation)  
 CorpusTranscript = tm\_map(CorpusTranscript, removeWords, wordsToRemove)  
 CorpusTranscript = tm\_map(CorpusTranscript, removeWords, stopwords("english"))  
 CorpusTranscript = DocumentTermMatrix(CorpusTranscript)  
 CorpusTranscript = removeSparseTerms(CorpusTranscript, 0.97) # keeps a matrix 97% sparse  
   
 return(CorpusTranscript)  
}  
  
Nword <- analyseText2(NegTweets$text)  
findAssocs(Nword, c("flight", 'customer', 'gate', 'phone'), .07)

## $flight  
## cancelled late flightled delayed   
## 0.36 0.25 0.23 0.16   
##   
## $customer  
## service   
## 0.65   
##   
## $gate  
## waiting plane   
## 0.09 0.08   
##   
## $phone  
## help   
## 0.07

# Further understanding the association with Clustering Analysis

# hierarchical clustering  
d = dist(t(as.matrix(Nword)), method = 'euclidean')  
fit = hclust(d = d, method = 'ward.D')  
  
# plotting the graph   
plot(fit, col = "#487AA1", col.main = "#45ADA8", col.lab = "#7C8071", main = 'Negative Sentiment', xlab = '', col.axis = "#F38630", lwd = 3, lty = 3, sub = "", hang = -1, axes = FALSE)  
  
# add axis   
axis(side = 2, at = seq(0, 400, 100), col = "#F38630", labels = FALSE, lwd = 2)



# k-mean clustering   
d = dist(t(as.matrix(Nword)), method="euclidean")   
kfit = kmeans(d, 3)   
clusplot(as.matrix(d), kfit$cluster, color=T, shade=T, labels=2, lines=0, cex = 0.4, main = 'Negative Sentiment')

