Airline_Tweets

Gautham Meenakshisundaram

3/4/2020

Introduction

In this exercise, we will be analyzing the tweets of airlines such as American, Delta, Southwest, United and US Airways.

Tweets per sentiment and airline

```
library(RSQLite) # Using RSQLite library to read the sqlite database

db_con = dbConnect(SQLite(), dbname="database.sqlite") # Connecting the
database
dbListTables(db_con) # Finding the tables inside the database
```

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

```
dbListFields(db_con,"tweets") # Finding the columns inside the Tweets table
```

```
##  [1] "tweet_id"                     "airline_sentiment"
##  [3] "airline_sentiment_confidence" "negativereason"
##  [5] "negativereason_confidence"    "airline"
##  [7] "airline_sentiment_gold"       "name"
##  [9] "negativereason_gold"          "retweet_count"
## [11] "text"                         "tweet_coord"
## [13] "tweet_created"                "tweet_location"
## [15] "user_timezone"
```

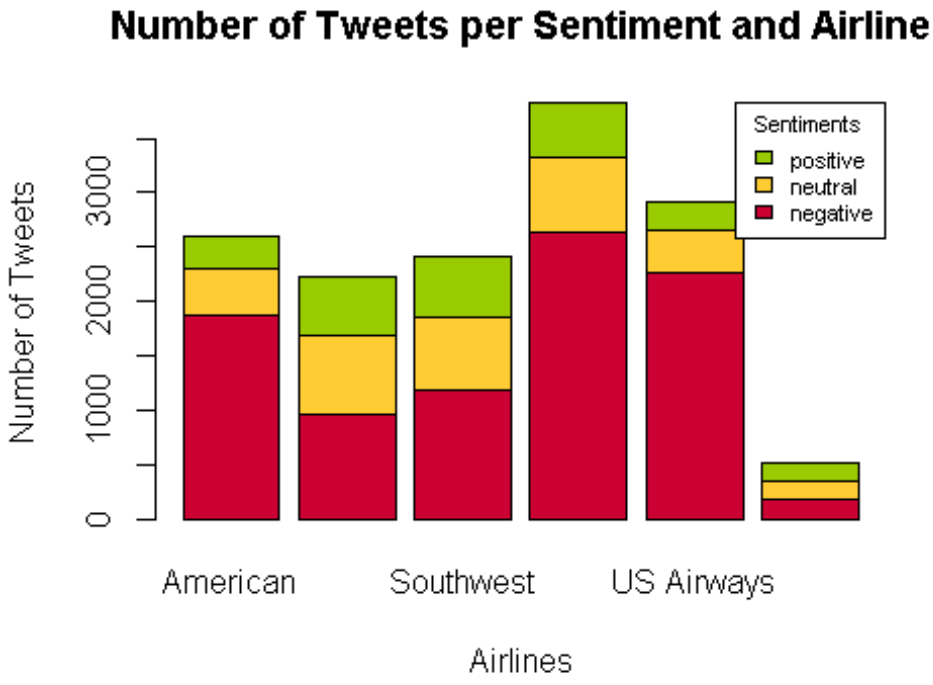
```
tweets_data = dbGetQuery(db_con,"Select * from tweets") # Creating a subset
with columns "airline_sentiment and "airline"
```

```
## Warning in result_fetch(res@ptr, n = n): Column
## `negativereason_confidence`: mixed type, first seen values of type string,
## coercing other values of type integer, real
```

```
task1 = table(tweets_data$airline_sentiment,tweets_data$airline)
task1
```

```
##
##             American Delta Southwest United US Airways Virgin America
##   negative     1864   955      1186   2633       2263            181
##   neutral       433   723       664    697        381            171
##   positive      307   544       570    492        269            152
```

```r
# Creating a bar chart to illustrate the number of tweets per sentiment and
airline

barplot(task1,  ylab = 'Number of Tweets', # adding label to Y axis
        xlab = 'Airlines', # adding label to X axis
        main = 'Number of Tweets per Sentiment and Airline', # adding title
to the chart
        col=c("#CC0033","#FFCC33","#99CC00"), # using the colours of traffic
lights for illustrating positive, neutral and negative sentiments
        legend.text = row.names(task1), # adding a legend
        args.legend = list(x = "topright", title = "Sentiments", cex = 0.70)
# adding a title to the legend box and adjusting its size and position
        )
```

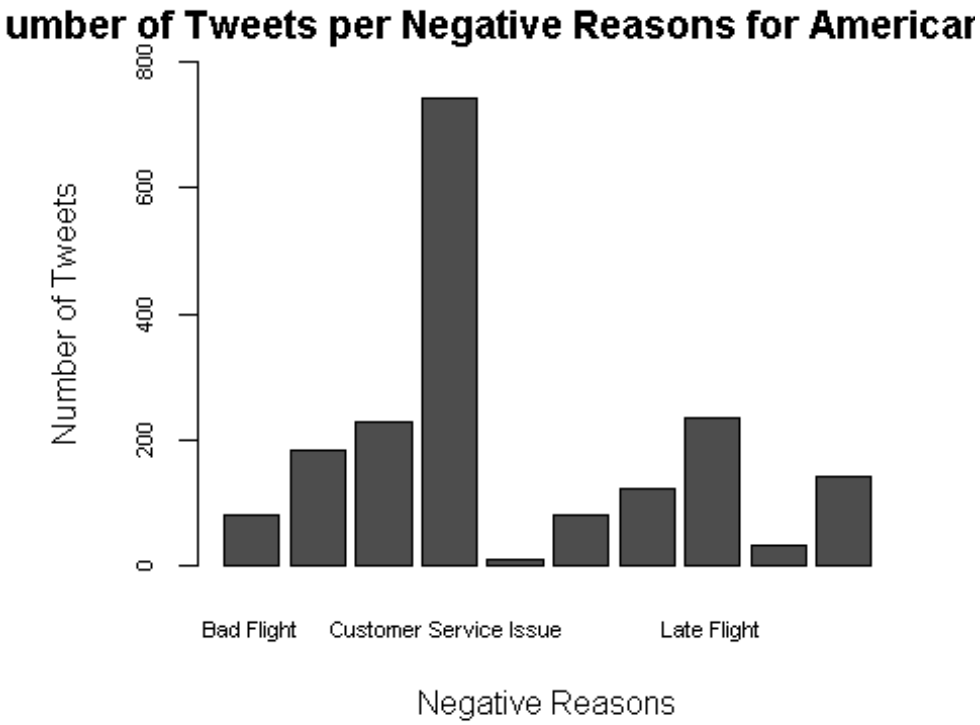


- United, US Airways, American airlines seems to have more tweets with negative sentiment compared the other two sentiments.
- Delta and Southwest seem to almost have all the three types of sentiments in equal parts, in their tweets.
- Virgin America clearly has all the three types of sentiments in equal parts, in their tweets.

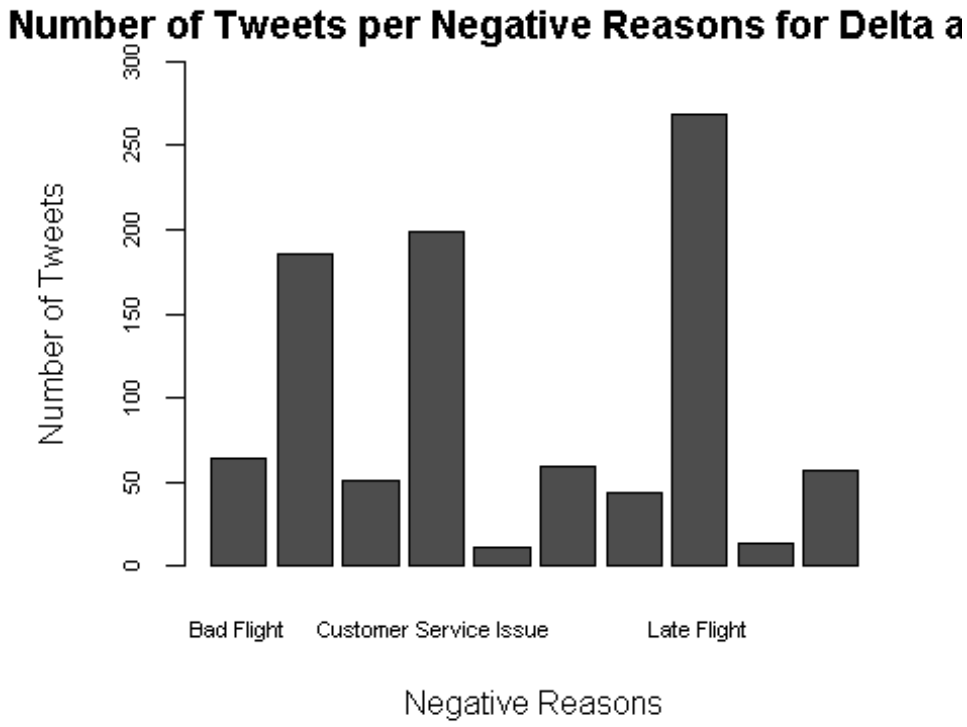
Most common causes of dissatisfaction in each company

```r
library(dplyr) # Using dplyr library to manipulate the data
```

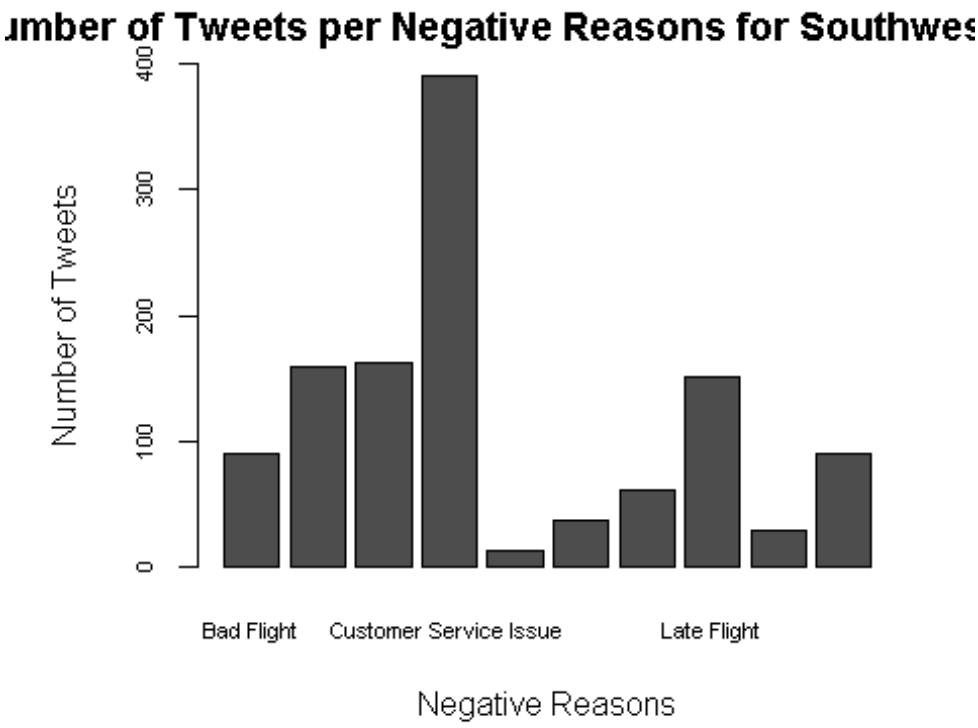
```
##
## Attaching package: 'dplyr'

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

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

american_airline = table(select(filter(tweets_data, airline == 'American',
airline_sentiment == 'negative'), c(airline, negativereason)))
# Creating a table of "airline" and "negativereason" by filtering 'American'
under "airline" and 'negative' under "airline_sentiment"
par(mar = c(5, 5, 2, 2)) # setting margins for the chart
barplot(american_airline,
        ylab = 'Number of Tweets', # adding labels for Y axis
        xlab = 'Negative Reasons', # adding labels for Xaxis
        main = 'Number of Tweets per Negative Reasons for American airline',
# adding title for chart
        ylim = c(0,800), # setting intervals in Y axis
        cex.axis = 0.7, # setting font size of intervals in Y axis
        cex.names = 0.7 # setting font size of intervals in X axis
       )
```



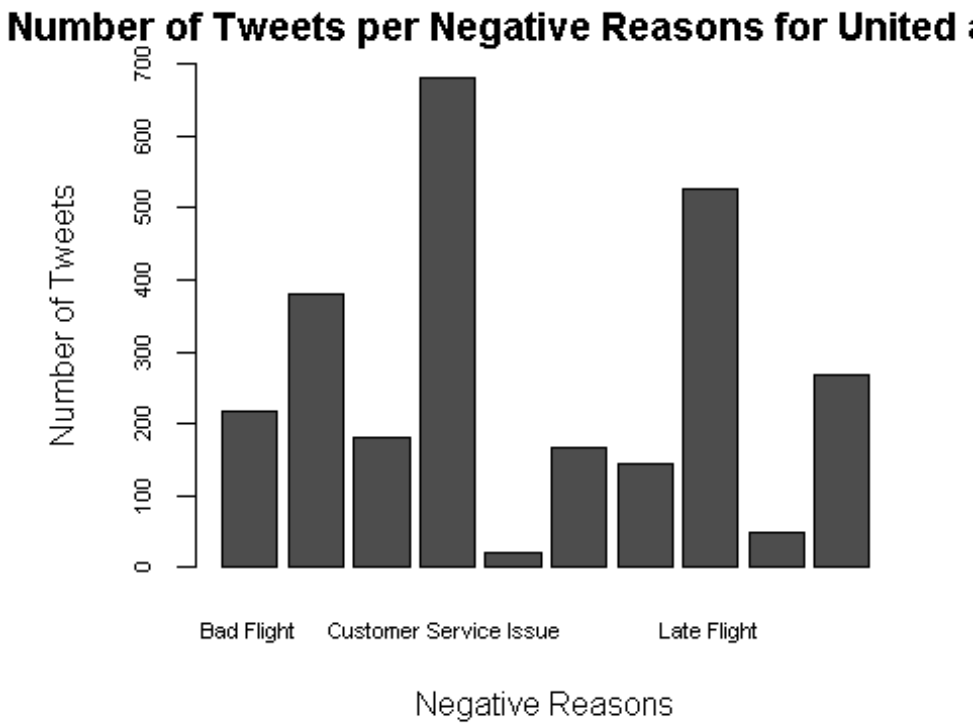
```r
delta_airline = table(select(filter(tweets_data, airline == 'Delta',
airline_sentiment == 'negative'), c(airline, negativereason)))
par(mar = c(5, 5, 2, 2))
barplot(delta_airline,
        ylab = 'Number of Tweets',
        xlab = 'Negative Reasons',
        main = 'Number of Tweets per Negative Reasons for Delta airline',
        ylim = c(0,300),
        cex.axis = 0.7,
        cex.names = 0.7
        )
```



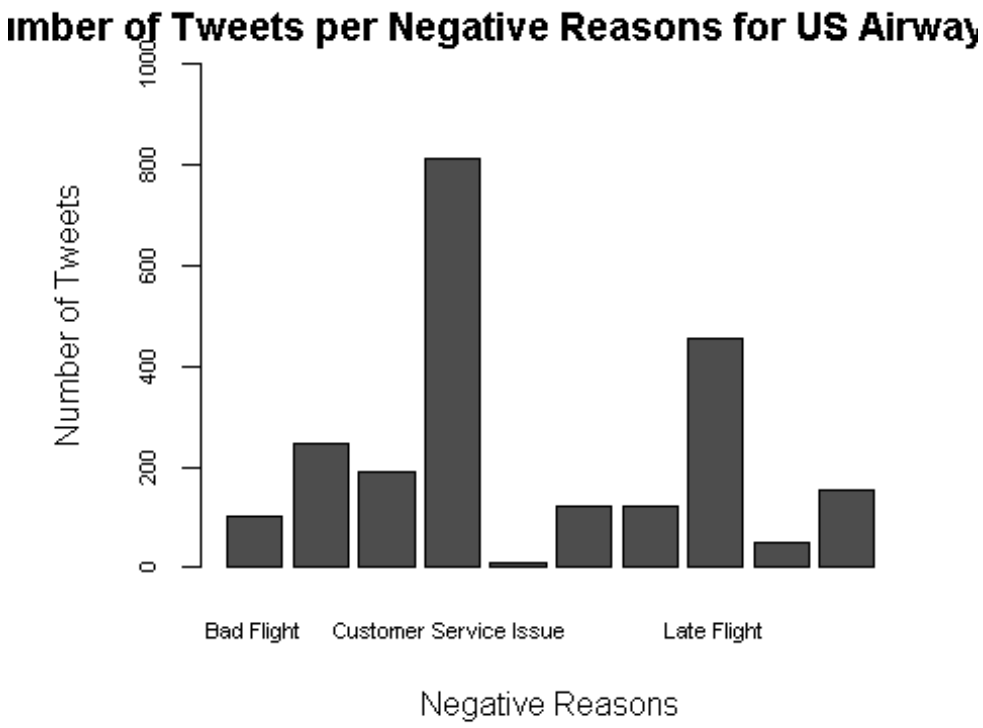

```r
southwest_airline = table(select(filter(tweets_data, airline == 'Southwest',
airline_sentiment == 'negative'), c(airline, negativereason)))
par(mar = c(5, 5, 2, 2))
barplot(southwest_airline,
        ylab = 'Number of Tweets',
        xlab = 'Negative Reasons',
        main = 'Number of Tweets per Negative Reasons for Southwest airline',
        ylim = c(0,400),
        cex.axis = 0.7,
        cex.names = 0.7
        )
```



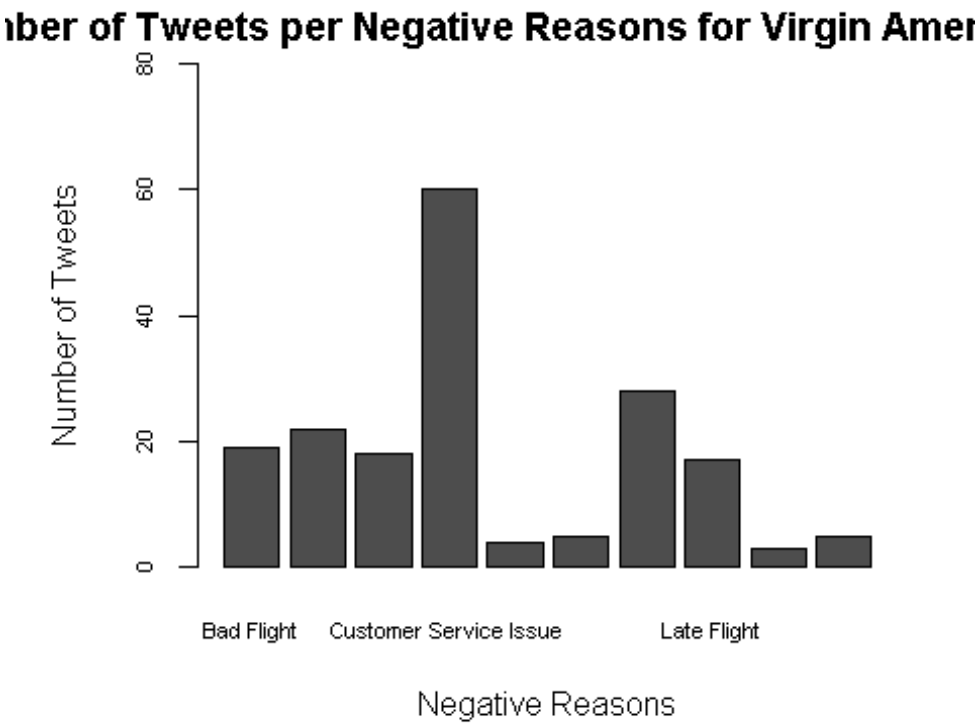

```
united_airline = table(select(filter(tweets_data, airline == 'United',
airline_sentiment == 'negative'), c(airline, negativereason)))
par(mar = c(5, 5, 2, 2))
barplot(united_airline,
        ylab = 'Number of Tweets',
        xlab = 'Negative Reasons',
        main = 'Number of Tweets per Negative Reasons for United airline',
        ylim = c(0,700),
        cex.axis = 0.7,
        cex.names = 0.7
        )
```




```
usairways_airline = table(select(filter(tweets_data, airline == 'US Airways',
airline_sentiment == 'negative'), c(airline, negativereason)))
par(mar = c(5, 5, 2, 2))
barplot(usairways_airline,
        ylab = 'Number of Tweets',
        xlab = 'Negative Reasons',
        main = 'Number of Tweets per Negative Reasons for US Airways
airline',
        ylim = c(0,1000),
        cex.axis = 0.7,
        cex.names = 0.7
        )
```




```
virginamerica_airline = table(select(filter(tweets_data, airline == 'Virgin
America', airline_sentiment == 'negative'), c(airline, negativereason)))
par(mar = c(5, 5, 2, 2))
barplot(virginamerica_airline,
        ylab = 'Number of Tweets',
        xlab = 'Negative Reasons',
        main = 'Number of Tweets per Negative Reasons for Virgin America
airline',
        ylim = c(0,80),
        cex.axis = 0.7,
        cex.names = 0.7
        )
```



- The bar chart of American airline suggests Customer Service as the most common cause of dissatisfaction.
- The bar chart of Delta airline suggests Late Flight as the most common cause of dissatisfaction with Customer Service behind.
- The bar chart of Southwest airline suggests Customer Service as the most common cause of dissatisfaction.
- The bar chart of United airline suggests Customer Service as the most common cause of dissatisfaction with Late Flight behind.
- The bar chart of US Airways suggests Customer Service as the most common cause of dissatisfaction with Late Flight behind.
- The bar chart of Virgin America suggests Customer Service as the most common cause of dissatisfaction.

Most frequent words in negative sentiments

```
library(tm)

## Loading required package: NLP

library(wordcloud)

## Loading required package: RColorBrewer

library(RColorBrewer)
# Using dplyr, tm, wordcloud, RColorBrewer for data manipulation, text
```

```
mining, word cloud generation and colourful representation

american_airline2 = select(filter(tweets_data, airline == 'American',
airline_sentiment == 'negative'), c(text))
# Creating a table of "text" by filtering 'American' under "airline" and
'negative' under "airline_sentiment"
american_corpus = Corpus(VectorSource(american_airline2))
# Creating a document called corpus which combines and collects all the texts
from the above table
american_corpus = tm_map(american_corpus, content_transformer(tolower))

## Warning in tm_map.SimpleCorpus(american_corpus,
## content_transformer(tolower)): transformation drops documents

# Converting the texts to lower case
american_corpus = tm_map(american_corpus, removeNumbers)

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

# Removing numbers from texts
american_corpus = tm_map(american_corpus, removeWords, stopwords("english"))

## Warning in tm_map.SimpleCorpus(american_corpus, removeWords,
## stopwords("english")): transformation drops documents

# Removing common english stop words from texts
american_corpus = tm_map(american_corpus, removePunctuation)

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

# Removing punctuations from texts
american_corpus = tm_map(american_corpus, stripWhitespace)

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

# Remove extra spaced between words in the texts
american_corpus = tm_map(american_corpus, stemDocument)

## Warning in tm_map.SimpleCorpus(american_corpus, stemDocument):
## transformation drops documents

# Performing Stemming process to the texts
american_corpus = tm_map(american_corpus, removeWords,
c("americanair","flight","aeroplane","airline","plane","airport","gate","agen
t"))

## Warning in tm_map.SimpleCorpus(american_corpus, removeWords,
## c("americanair", : transformation drops documents
```

```
# Removing American airline and additional common words related to airline
industry from the texts
american_tdm = TermDocumentMatrix(american_corpus)
american_matrix = as.matrix(american_tdm)
american_sorted = sort(rowSums(american_matrix), decreasing = TRUE)
# Creating a term-document matrix which will describe the frequency of words
that occur in the texts
american_df = data.frame(word = names(american_sorted), freq =
american_sorted)
# Converting the matrix into a dataframe

#wordcloud(american_df$word,american_df$freq,c(3,.5),3,,FALSE,.05,colors =
brewer.pal(6,"Dark2"),random.order=FALSE)
```