ASSIGNMENT 2

There are an average of 6,000 tweets produced on Twitter per second. We think Twitter provides a great value to do sentiment analysis on text. Twitter posts are mostly public and can be used for such studies extensively. Also, frequent use of hashtags makes it more interesting to draw conclusions.

Sentiment analysis on airlines intrigues us since the industry is heavily price oriented. Often prices for tickets from different airlines are in similar range, putting emphasis on the quality of travel experience for the customer. But the customer understanding of the airline is commonly based on personal experience or general news. The customer is interested to know which airlines have a better reputation since the ticket prices are alike.

I did a sentiment analysis on tweets provided by the Twitter Search API in order to find the most preferable airline. I made a selection of airlines for the analysis. The selection was made at random. The other contributing factor was my initial analysis of the data collection process to verify that a sufficient amount of data is created for each airline. The six airlines I ended up with were United, Virgin America, Southwest Air, Delta, JetBlue and American Air.

I created the following R script to collect raw data from the Twitter Search API:

```
# Install and Activate Packages
install.packages("twitteR", "RCurl", "RJSONIO", "stringr")#If this command
doesn't work install the packages separately manually
library(twitteR)
library(RCurl)
library(RJSONIO)
library(stringr)
library(plyr)

# Declare Twitter API Credentials
api_key <- "dqOmngRexwbjBQ9op25RneZW1" # From dev.twitter.com
api_secret <- "9Xg2MdGoepC4BeTm9fVu5YbI8T59e2peAKa5KYlHh7EEzJYn0W" # From
dev.twitter.com
token <- "633654027-ctjq5J0OSYtV2uvrhZTnOtFfAjW5RrnZQOKz1fqF" # From
dev.twitter.com
token_secret <- "RTPXzFHYYQEFXYIoUXqqNzKuGWI32C6JtASk31ES23rP2" # From
dev.twitter.com

# Create Twitter Connection
setup_twitter_oauth(api_key, api_secret, token, token_secret)

#tweets <- searchTwitter('@JetBlue OR @AmericanAir OR @VirginAmerica OR
@SouthwestAir OR @united OR @USAirways  OR @VirginAmerica', n=100,
lang="en", since="2016-10-11", until="2016-10-12")
tweets1 <- searchTwitter('@JetBlue', n=2500, lang="en")
tweets2 <- searchTwitter('@AmericanAir', n=2500, lang="en")
tweet3 <- searchTwitter('@VirginAmerica', n=2500, lang="en")
tweets4 <- searchTwitter('@SouthwestAir', n=2500, lang="en")
tweets5 <- searchTwitter('@united', n=2500, lang="en")
tweets6 <- searchTwitter('@FlyFrontier', n=2500, lang="en")

# Transform tweets list into a data frame
tweets1.df <- twListToDF(tweets1)
tweets2.df <- twListToDF(tweets2)
tweets3.df <- twListToDF(tweet3)
```

```
tweets4.df <- twListToDF(tweets4)
tweets5.df <- twListToDF(tweets5)
tweets6.df <- twListToDF(tweets6)

tweets1.df$airline <- "Delta"
tweets2.df$airline <- "AmericanAir"
tweets3.df$airline <- "Virgin America"
tweets4.df$airline <- "Southwest"
tweets5.df$airline <- "United"
tweets6.df$airline <- "US Airways"

tweets1.df$TweetLength <- nchar(tweets1.df$text)
tweets2.df$TweetLength <- nchar(tweets2.df$text)
tweets3.df$TweetLength <- nchar(tweets3.df$text)
tweets4.df$TweetLength <- nchar(tweets4.df$text)
tweets5.df$TweetLength <- nchar(tweets5.df$text)
tweets6.df$TweetLength <- nchar(tweets6.df$text)

tweets1.df$NoOfWords <- sapply(gregexpr("\\W+", tweets1.df$text), length) +
1
tweets2.df$NoOfWords <- sapply(gregexpr("\\W+", tweets2.df$text), length) +
1
tweets3.df$NoOfWords <- sapply(gregexpr("\\W+", tweets3.df$text), length) +
1
tweets4.df$NoOfWords <- sapply(gregexpr("\\W+", tweets4.df$text), length) +
1
tweets5.df$NoOfWords <- sapply(gregexpr("\\W+", tweets5.df$text), length) +
1
tweets6.df$NoOfWords <- sapply(gregexpr("\\W+", tweets6.df$text), length) +
1

tweets1.df = subset(tweets1.df, select = -
c(favorited,favoriteCount,truncated,replyToSID,replyToUID,statusSource,isRe
tweet,retweeted))
tweets2.df = subset(tweets2.df, select = -
c(favorited,favoriteCount,truncated,replyToSID,replyToUID,statusSource,isRe
tweet,retweeted))
tweets3.df = subset(tweets3.df, select = -
c(favorited,favoriteCount,truncated,replyToSID,replyToUID,statusSource,isRe
tweet,retweeted))
tweets4.df = subset(tweets4.df, select = -
c(favorited,favoriteCount,truncated,replyToSID,replyToUID,statusSource,isRe
tweet,retweeted))
tweets5.df = subset(tweets5.df, select = -
c(favorited,favoriteCount,truncated,replyToSID,replyToUID,statusSource,isRe
tweet,retweeted))
tweets6.df = subset(tweets6.df, select = -
c(favorited,favoriteCount,truncated,replyToSID,replyToUID,statusSource,isRe
tweet,retweeted))


tweets1.df <- rename(tweets1.df, c("text"="tweettext",
"TweetLength"="numberofcharacters", "id"="TweetID", "screenName"="userID",
"NoOfWords"="numberofwords", "created"="timestamp"))
tweets2.df <- rename(tweets2.df, c("text"="tweettext",
"TweetLength"="numberofcharacters", "id"="TweetID", "screenName"="userID",
"NoOfWords"="numberofwords", "created"="timestamp"))
tweets3.df <- rename(tweets3.df, c("text"="tweettext",
"TweetLength"="numberofcharacters", "id"="TweetID", "screenName"="userID",
"NoOfWords"="numberofwords", "created"="timestamp"))
```

```
tweets4.df <- rename(tweets4.df, c("text"="tweettext",
"TweetLength"="numberofcharacters", "id"="TweetID", "screenName"="userID",
"NoOfWords"="numberofwords", "created"="timestamp"))
tweets5.df <- rename(tweets5.df, c("text"="tweettext",
"TweetLength"="numberofcharacters", "id"="TweetID", "screenName"="userID",
"NoOfWords"="numberofwords", "created"="timestamp"))
tweets6.df <- rename(tweets6.df, c("text"="tweettext",
"TweetLength"="numberofcharacters", "id"="TweetID", "screenName"="userID",
"NoOfWords"="numberofwords", "created"="timestamp"))

Final_data <-
rbind(tweets1.df,tweets2.df,tweets3.df,tweets4.df,tweets5.df,tweets6.df)
write.csv(Final_data, file = "E:/documents/CourseWork/FODS/Tweet.csv")
```

My project focuses on analysing the text of over 10,000 tweets about airlines using sentiment analysis and other methods to determine which airline receives the most positive or negative attention on Twitter, and what topics people are happy or sad about with regards to each airline. I used the following code for text pre-processing and cleansing:

```
library(tm)
library(Rstem)
library(sentiment)

c13a <- read.csv("E:/documents/CourseWork/FODS/Tweet.csv", header = TRUE,
sep = ",")
some_tweets = c13a$tweettext# get the text
some_txt = sapply(some_tweets, function(x) x$getText())

# remove retweet entities
some_txt = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", c13a$tweettext)
# remove @ people
#some_txt = gsub("@\\w+", "", some_txt)
# remove punctuation
some_txt = gsub("[[:punct:]]", "", some_txt)
# remove numbers
some_txt = gsub("[[:digit:]]", "", some_txt)
# remove html links
some_txt = gsub("http\\w+", "", some_txt)
# remove unnecessary spaces
some_txt = gsub("[ \t]{2,}", "", some_txt)
some_txt = gsub("^\\s+|\\s+$", "", some_txt)

# define "tolower error handling" function
try.error = function(x)
{
  # create missing value
  y = NA
  # tryCatch error
  try_error = tryCatch(tolower(x), error=function(e) e)
  # if not an error
  if (!inherits(try_error, "error"))
    y = tolower(x)
  # result
  return(y)
}
# lower case using try.error with sapply
some_txt = sapply(some_txt, try.error)

# remove NAs in some_txt
```

```
some_txt = some_txt[!is.na(some_txt)]
names(some_txt) = NULL
```

For running the above code requires a bit of an effort. We need to downgrade the R version to 3.0.2. Once done, then install the `tm` and `Rstem` packages manually and download the sentiment package using the following link:

https://cran.r-project.org/src/contrib/Archive/sentiment/.

Download the latest version of the sentiment library. Install this package using the below command:

```
install.packages("C:/Users/Abhinav/Desktop/sentiment_0.2.tar.gz", repos =
NULL, type="source")
* installing *source* package 'sentiment' ...
** package 'sentiment' successfully unpacked and MD5 sums checked
** R
** data
** preparing package for lazy loading
Warning: package 'tm' was built under R version 3.0.3
** help
*** installing help indices
** building package indices
** testing if installed package can be loaded
Warning: package 'tm' was built under R version 3.0.3
* DONE (sentiment)
```

An output like this means, the package has been installed successfully and we can go ahead with our sentiment analysis. Using the below code, I did a superficial text analysis and try to conclude what kind of expression is given by the user in his/her tweets:

```
# classify emotion
class_emo = classify_emotion(some_txt, algorithm="bayes", prior=1.0)
# get emotion best fit
emotion = class_emo[,7]
# substitute NA's by "unknown"
emotion[is.na(emotion)] = "unknown"

# classify polarity
class_pol = classify_polarity(some_txt, algorithm="bayes")
# get polarity best fit
polarity = class_pol[,4]

# data frame with results
sent_df = data.frame(text=some_txt, emotion=emotion, polarity=polarity,
stringsAsFactors=FALSE)

# sort data frame
sent_df = within(sent_df, emotion <- factor(emotion,
levels=names(sort(table(emotion), decreasing=TRUE))))

#View the data
View(sent_df)
```

The output will be like:

sent_df

Filter

	text	emotion	polarity
1	rattlesandheels mommytalkshow jetblue americanair ...	sadness	neutral
2	airbus a njb de jetblue despegando del aeropuertoci...	unknown	positive
3	jetblue utilizará kioscos de sitaonline para registro d...	unknown	positive
4	austin yall were great met these guys who flew all th...	joy	neutral
5	and sophialeonoram actonclimate by working with mi...	joy	positive
6	diversion jetblue b from san juan to chicago has dive...	unknown	negative
7	jetblue kimmeabreak skeeternutfree id love one hmm...	joy	positive
8	jetblue this seems dangerous	unknown	negative
9	yay i know you would keep it real looks like jetblue a...	joy	positive
10	inspiring humanity with huge resultscsr	unknown	positive
11	skeeternutfree jetblue these cookies r so good	joy	positive
12	prepare for snacking uufef eduaubcedubduaaskeete...	joy	positive
13	diversion jetblue b from san juan to chicago has dive...	unknown	negative
14	if you want to avoid a horrible flight experience do n...	fear	negative
15	so no courtesy forminutes im going to have to rethin...	unknown	positive
16	airplane wifi such a treat thanks jetblue youre aweso...	unknown	positive

Using some other techniques in this library, I calculated values like airline_sentiment, airline_sentiment_confidence, negativereason and negativereason_confidence. The final dataset has been compiled and has been attached with this report. Also I haven't kept columns like **latitude** and **longitude** and instead combined it into one column named "**tweet_location**" and using R code I have also got the accurate location using these co-ordinates. Following is the corresponding R-code:

```
tw1.df <- tweets1.df
tw1.df = subset(tw1.df, select = -
c(retweetCount,screenName,id,created,replyToSN,text,favorited,favoriteCount
,truncated,replyToSID,replyToUID,statusSource,isRetweet,retweeted))
#tweets1.df = subset(tweets1.df, select = -c(textAddress))

tw1.df <- data.frame(lapply(tw1.df, function(x) round(as.numeric(x), 8)))

indices <- which(complete.cases(tw1.df[,1:2]))
tw1.df$textAddress <- NA
tw1.df[indices,]$textAddress <- mapply(FUN = function(lon, lat)
  revgeocode(c(lon, lat)),
  tw1.df[indices,]$longitude,
  tw1.df[indices,]$latitude)

tweets1.df$tweet_location <- tw1.df$textAddress
```

Just don't forget to install **ggmap** and load the library, before running the above code.