Twitter Airline Sentiment Analysis Insight Report

JJ Goh

10/08/2020

```r
getwd()
```

```
## [1] "C:/Users/JJ/Documents/twitter"
```

```r
library("scales")
library('tidyr')
library("tidytext")
library("tm")
library("slam")
library("wordcloud")
library("RColorBrewer")
library('stringr')
library('dplyr')
library('tibble')
library('data.table')
library("reshape2")
library("knitr")
library("ggplot2")
library('tidyverse')
```

Loading Dataset

```r
tweetdata <- fread(paste0("Tweets.csv"))
```

Inspecting tweetdata

```r
head(tweetdata)
##        tweet_id airline_sentiment airline_sentiment_confidence
## 1: 5.70306e+17           neutral                       1.0000
## 2: 5.70301e+17          positive                       0.3486
## 3: 5.70301e+17           neutral                       0.6837
## 4: 5.70301e+17          negative                       1.0000
## 5: 5.70301e+17          negative                       1.0000
## 6: 5.70301e+17          negative                       1.0000
##    negativereason negativereason_confidence        airline
```

```
## 1:                                        NA Virgin America
## 2:                                    0.0000 Virgin America
## 3:                                        NA Virgin America
## 4:     Bad Flight                     0.7033 Virgin America
## 5:     Can't Tell                     1.0000 Virgin America
## 6:     Can't Tell                     0.6842 Virgin America
##    airline_sentiment_gold       name negativereason_gold
## 1:                           cairdin
## 2:                          jnardino
## 3:                        yvonnalynn
## 4:                          jnardino
## 5:                          jnardino
## 6:                          jnardino
##    retweet_count
## 1:             0
## 2:             0
## 3:             0
## 4:             0
## 5:             0
## 6:             0
##
## 1:
## 2:                                                                  @VirginAmerica plus you've added
## 3:                                                                   @VirginAmerica I didn't today...
## 4:        @VirginAmerica it's really aggressive to blast obnoxious ""entertainment"" in your guests
## 5:                                                                                 @VirginAmerica a
## 6: @VirginAmerica seriously would pay $30 a flight for seats that didn't have this playing.\nit's re
##    tweet_coord tweet_created tweet_location
## 1:                2015-02-24
## 2:                2015-02-24
## 3:                2015-02-24      Lets Play
## 4:                2015-02-24
## 5:                2015-02-24
## 6:                2015-02-24
##                user_timezone V16 V17 V18 V19 V20 V21 V22 V23 V24
## 1: Eastern Time (US & Canada)  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 2: Pacific Time (US & Canada)  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 3: Central Time (US & Canada)  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 4: Pacific Time (US & Canada)  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 5: Pacific Time (US & Canada)  NA  NA  NA  NA  NA  NA  NA  NA  NA
## 6: Pacific Time (US & Canada)  NA  NA  NA  NA  NA  NA  NA  NA  NA
```

Frequency of airlines mentioned

```
tweetdata[, .N, airline]
##           airline    N
## 1: Virgin America  504
## 2:         United 3822
## 3:      Southwest 2420
## 4:          Delta 2222
```

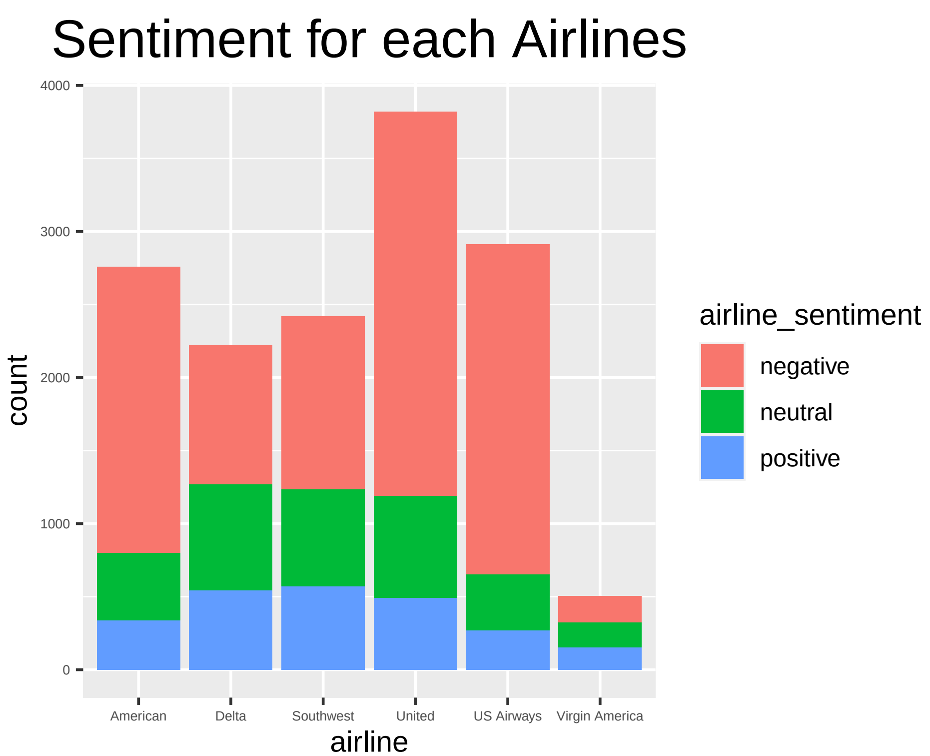
```
## 5:      US Airways 2913
## 6:        American 2759
```

Frequency of sentiment (negative/positive/Neutral)

```
tweetdata[, .N, airline_sentiment]
##    airline_sentiment    N
## 1:           neutral 3099
## 2:          positive 2363
## 3:          negative 9178
```

Sentiment for each Airlines

```
plot2 <- ggplot(tweetdata, aes(airline, fill = airline_sentiment)) + geom_bar() + ggtitle("Sentiment fo

plot2
```

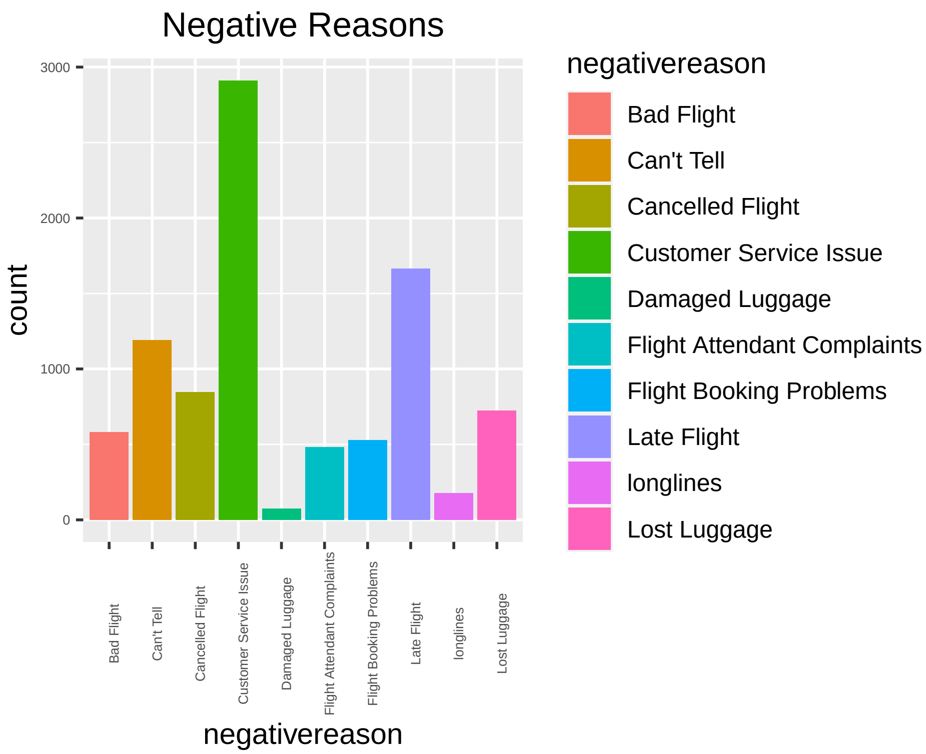


United Airlines was the most tweeted about, followed by US Airways and American Airlines. About 62% of the total tweets are negative.

Diving down into the negative reason

```
tweetdata[, .N, negativereason]
##                  negativereason    N
##  1:                             5462
##  2:                  Bad Flight  580
##  3:                  Can't Tell 1190
##  4:                 Late Flight 1665
##  5:      Customer Service Issue 2910
##  6:     Flight Booking Problems  529
##  7:                Lost Luggage  724
##  8: Flight Attendant Complaints  481
##  9:            Cancelled Flight  847
## 10:             Damaged Luggage   74
## 11:                   longlines  178
```

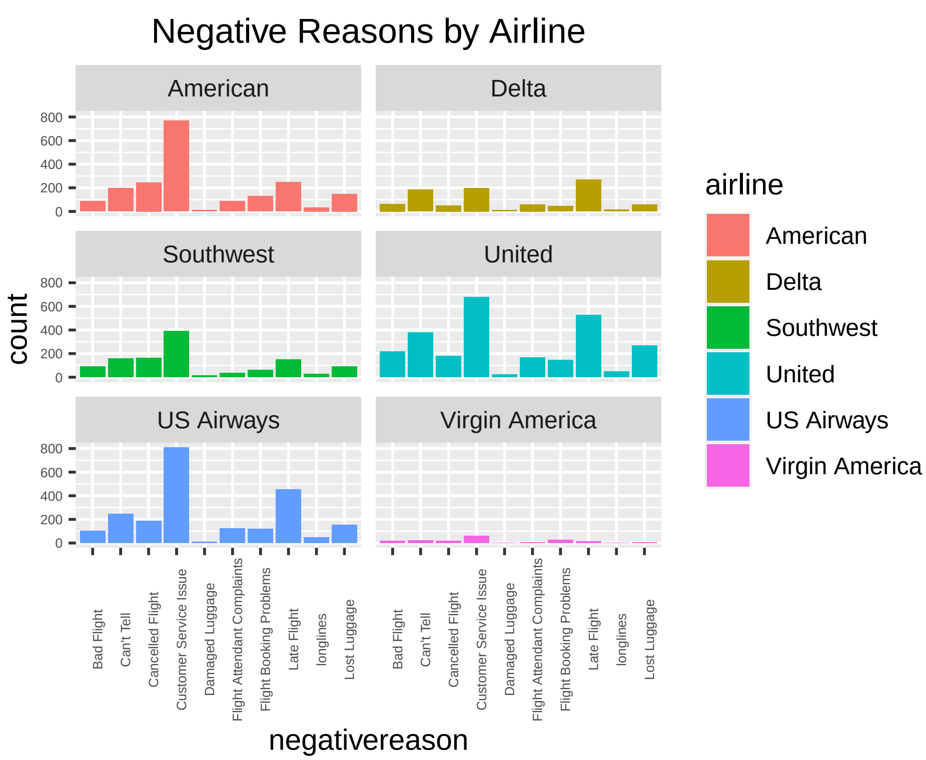
```
plot3 <- tweetdata %>%
  filter(negativereason != "") %>%
  ggplot(aes(negativereason, fill = negativereason)) + geom_bar() + theme(axis.text.x = element_text(ang

plot3
```



Majority of the negative reasons is due to Customer service issue followed by late flights. For this code, I used the filter function to filter out all the NA's within the 'negativereason' variable.

Negative reasons for each airline

```
plot4 <-  tweetdata %>%
  filter(negativereason != "") %>%
  ggplot( aes(negativereason, fill = airline)) + geom_bar() +
  facet_wrap(~airline, ncol = 2) +
  theme( axis.text=element_text(size=5),
         axis.text.x = element_text(angle = 90, margin = margin(1), vjust = 1)) +
  ggtitle("Negative Reasons by Airline") + theme(axis.text=element_text(size=5),
  plot.title = element_text(hjust = 0.5))

plot4
```



American, United, Southwest,US Airways and Virgin America's main reason for negative reason is due to customer service issue with US Airway being the highest (over 800 counts) Delta's main negative reason is due to late flights (about 200 counts) Virgin America has the least negative reasons among all the airlines.

Data cleaning

Remove airline_sentiment_gold, negativereason_gold, and tweet_coord since has most number of NAs

```r
clean_df <- subset(tweetdata, select = -c(airline_sentiment_gold, negativereason_gold, tweet_coord))
head(clean_df)
##       tweet_id airline_sentiment airline_sentiment_confidence
## 1: 5.70306e+17           neutral                       1.0000
## 2: 5.70301e+17          positive                       0.3486
## 3: 5.70301e+17           neutral                       0.6837
## 4: 5.70301e+17          negative                       1.0000
## 5: 5.70301e+17          negative                       1.0000
## 6: 5.70301e+17          negative                       1.0000
##    negativereason negativereason_confidence        airline       name
## 1:                                       NA Virgin America    cairdin
## 2:                                   0.0000 Virgin America   jnardino
## 3:                                       NA Virgin America yvonnalynn
## 4:     Bad Flight                    0.7033 Virgin America   jnardino
## 5:     Can't Tell                    1.0000 Virgin America   jnardino
## 6:     Can't Tell                    0.6842 Virgin America   jnardino
##    retweet_count
## 1:             0
## 2:             0
## 3:             0
## 4:             0
## 5:             0
## 6:             0
##
## 1:
## 2:                                                                 @VirginAmerica plus you've added
## 3:                                                                  @VirginAmerica I didn't today...
## 4:         @VirginAmerica it's really aggressive to blast obnoxious ""entertainment"" in your guests
## 5:                                                                                  @VirginAmerica a
## 6: @VirginAmerica seriously would pay $30 a flight for seats that didn't have this playing.\nit's re
##    tweet_created tweet_location               user_timezone V16 V17
## 1:    2015-02-24                Eastern Time (US & Canada)  NA  NA
## 2:    2015-02-24                Pacific Time (US & Canada)  NA  NA
## 3:    2015-02-24      Lets Play Central Time (US & Canada)  NA  NA
## 4:    2015-02-24                Pacific Time (US & Canada)  NA  NA
## 5:    2015-02-24                Pacific Time (US & Canada)  NA  NA
## 6:    2015-02-24                Pacific Time (US & Canada)  NA  NA
##    V18 V19 V20 V21 V22 V23 V24
## 1:  NA  NA  NA  NA  NA  NA  NA
## 2:  NA  NA  NA  NA  NA  NA  NA
## 3:  NA  NA  NA  NA  NA  NA  NA
## 4:  NA  NA  NA  NA  NA  NA  NA
## 5:  NA  NA  NA  NA  NA  NA  NA
## 6:  NA  NA  NA  NA  NA  NA  NA
```

Remove timestamp and count number of tweets by date

```r
clean_df$tweet_created <- as.Date(clean_df$tweet_created)
```

```
clean_df[, .N, by = tweet_created]
##    tweet_created    N
## 1:    2015-02-24 1344
## 2:    2015-02-23 3028
## 3:    2015-02-22 3079
## 4:    2015-02-21 1557
## 5:    2015-02-20 1500
## 6:    2015-02-19 1376
## 7:    2015-02-18 1344
## 8:    2015-02-17 1408
## 9:    2015-02-16    4
```

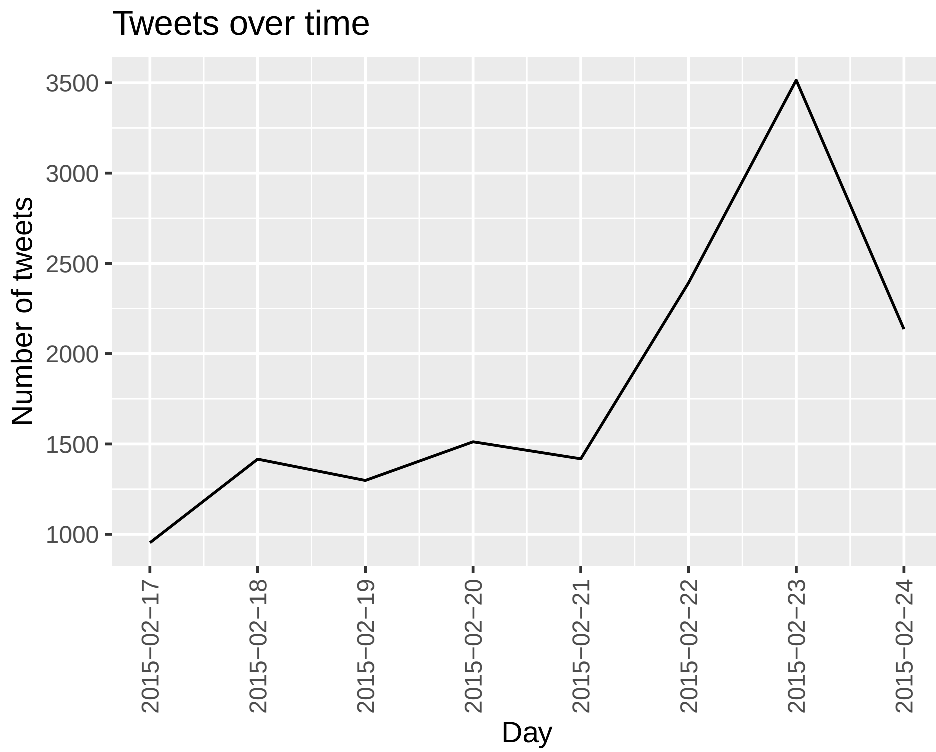
Data shows that tweets are created between 2015-02-16 to 2015-02-24

create a sequence of dates and merge with dataset

```
plot5 <- tribble(
  ~Date, ~count_of_tweets,
  "2015-02-17", 953,
  "2015-02-18", 1416,
  "2015-02-19", 1298,
  "2015-02-20", 1512,
  "2015-02-21", 1418,
  "2015-02-22", 2392,
  "2015-02-23", 3515,
  "2015-02-24", 2136
)
```

```
plot5$Date <- as.Date(plot5$Date)
```

```
ggplot(plot5, aes(x = Date, y = count_of_tweets)) +
 geom_line() +
 labs(x = "Day", y = "Number of tweets", title = "Tweets over time") +
 scale_x_date(breaks = "1 day") +
 theme(axis.text.x = element_text(angle = 90, vjust = 0.5, ))
```



It seems that the number of tweets on 2015-02-23 (Sunday) is the highest. This might be due to the fact that more people are traveling on a Sunday as they might be returning home from a short vacation. I also created this graph on Tableau as initially I was unable to figure out a way to make this on R. Let's investigate further and try to see which airlines and reason these tweets are mentioning.

Please refer to graph 'Count of Sentiment of Tweets over time' for graph (I created this graph on excel). As expected, the most number of negative tweets occurred on 2015-02-23 (Sunday).

Text Mining

Grabbing all texts and removing special characters for word cloud

```
text <- clean_df[["text"]]
text2=gsub("(RT|via)((?:\\b\\W*@\\w+)+)","",text)
text3=gsub("http[^[:blank:]]+","",text2)
text4=gsub("@\\w+","",text3)
text5=gsub("[[:punct:]]"," ",text4)
text6=gsub("[^[:alnum:]]"," ",text5)
```

Create wordcloud of the most used 150 words

```r
wordcloud(text6, min.freq=5 ,max.words=150, with=1000, height=1000,
          random.color=TRUE, random.order=FALSE, color=brewer.pal(8,"Dark2") )
```

Words such as flight, service, get, thanks appeared the most. Although words such as 'thank' might have a positive connotation, it is important to keep in mind that the tweet might be tweeted in a sarcastic way. For example, "Flight was delayed again. Thanks United."

At this point, I realised that creating a new excel file would be easier for me text mine, therefore I took the columns 'airline' and 'text' and pasted it on another file called tweetr.csv and uploaded it.I have attached the file for your reference.

```r
tweetr <- fread(paste0("tweetr.csv"))
```

Please take note that the following codes such as removing stop words, calculating word frequencies and comparing word usage was learned from: https://www.tidytextmining.com/index.html

Removing stopwords

```r
remove_reg <- "&amp;|&lt;|&gt;"
tidy_tweets <- tweetr %>%
  filter(!str_detect(text, "^RT")) %>%
  mutate(text = str_remove_all(text, remove_reg)) %>%
  unnest_tokens(word, text, token = "tweets") %>%
  filter(!word %in% stop_words$word,
         !word %in% str_remove_all(stop_words$word, "'"),
         str_detect(word, "[a-z]"))
```

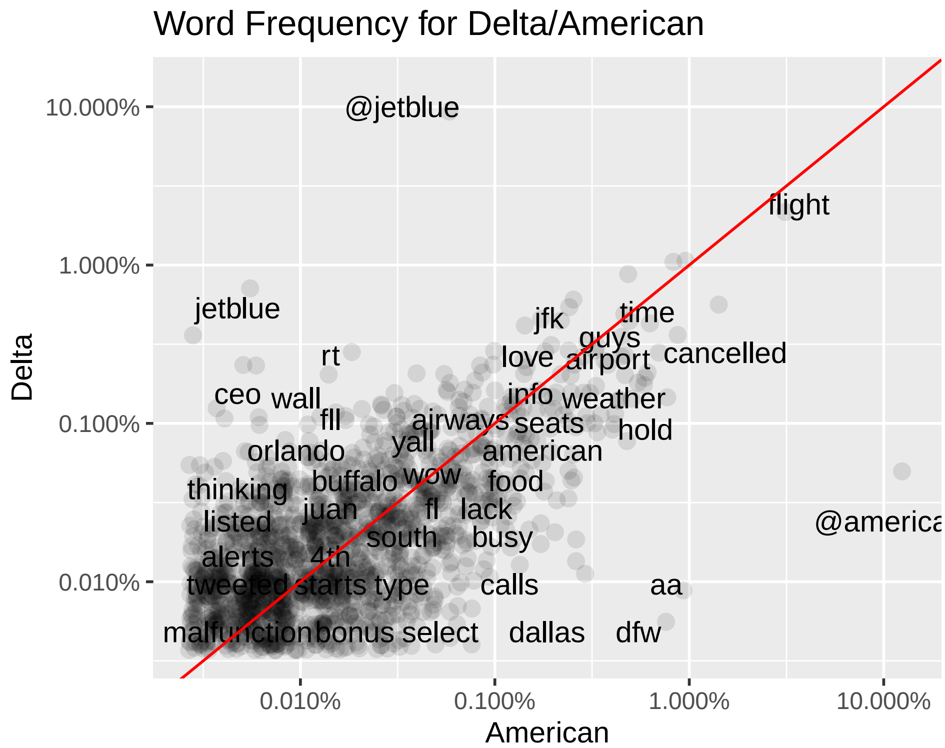
```r
frequency <- tidy_tweets %>%
  group_by(airline) %>%
  count(word, sort = TRUE) %>%
  left_join(tidy_tweets %>%
              group_by(airline) %>%
              summarise(total = n())) %>%
  mutate(freq = n/total)

frequency
## # A tibble: 26,736 x 5
## # Groups:   airline [6]
##    airline   word              n total   freq
##    <chr>     <chr>         <int> <int>  <dbl>
##  1 United    @united        3782 29725 0.127
##  2 USAirways @usairways     2904 23190 0.125
##  3 American  @americanair   2723 21022 0.130
##  4 Southwest @southwestair  2394 17537 0.137
##  5 Delta     @jetblue       2073 15218 0.136
##  6 United    flight          994 29725 0.0334
##  7 USAirways flight          877 23190 0.0378
##  8 American  flight          768 21022 0.0365
##  9 Southwest flight          610 17537 0.0348
## 10 Delta     flight          503 15218 0.0331
## # ... with 26,726 more rows
```

Converting vertical data to horizontal

```r
frequency <- frequency %>%
  select(airline, word, freq) %>%
  spread(airline, freq) %>%
  arrange(United, USAirways, American, Southwest, Delta, VirginAmerica)
```

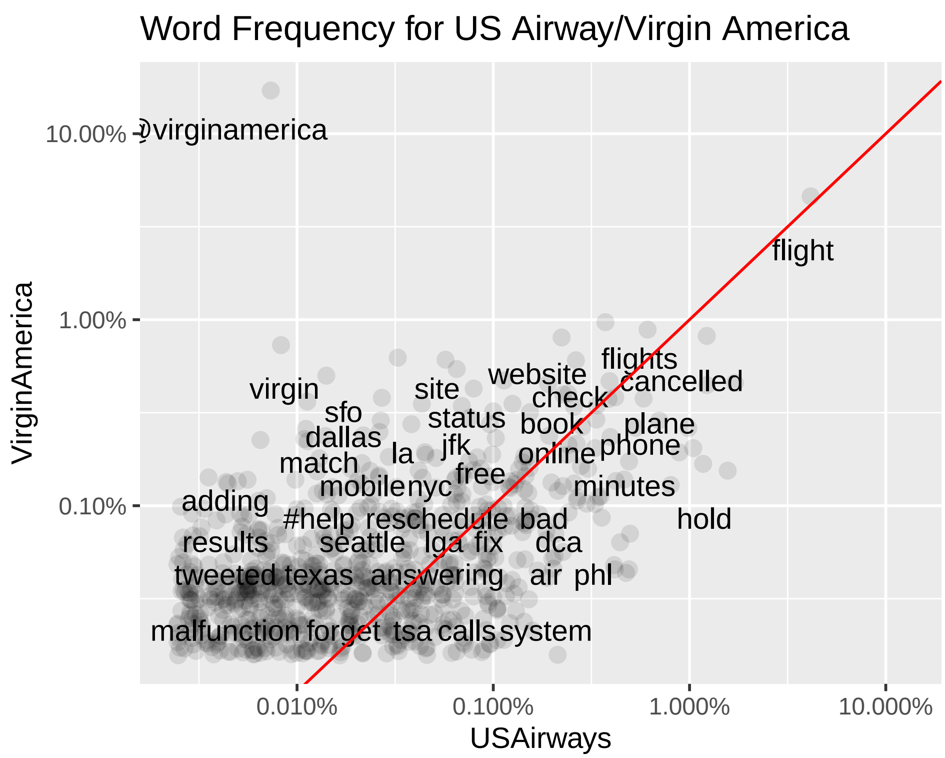
Plotting word Frequency with Geom_Jitter

```r
ggplot(frequency, aes(American, Delta)) +
  geom_jitter(alpha = 0.1, size = 2.5, width = 0.25, height = 0.25) +
  geom_text(aes(label = word), check_overlap = TRUE, vjust = 1.5) +
  scale_x_log10(labels = percent_format()) +
  scale_y_log10(labels = percent_format()) +
  geom_abline(color = "red") + labs(title = "Word Frequency for Delta/American")
```



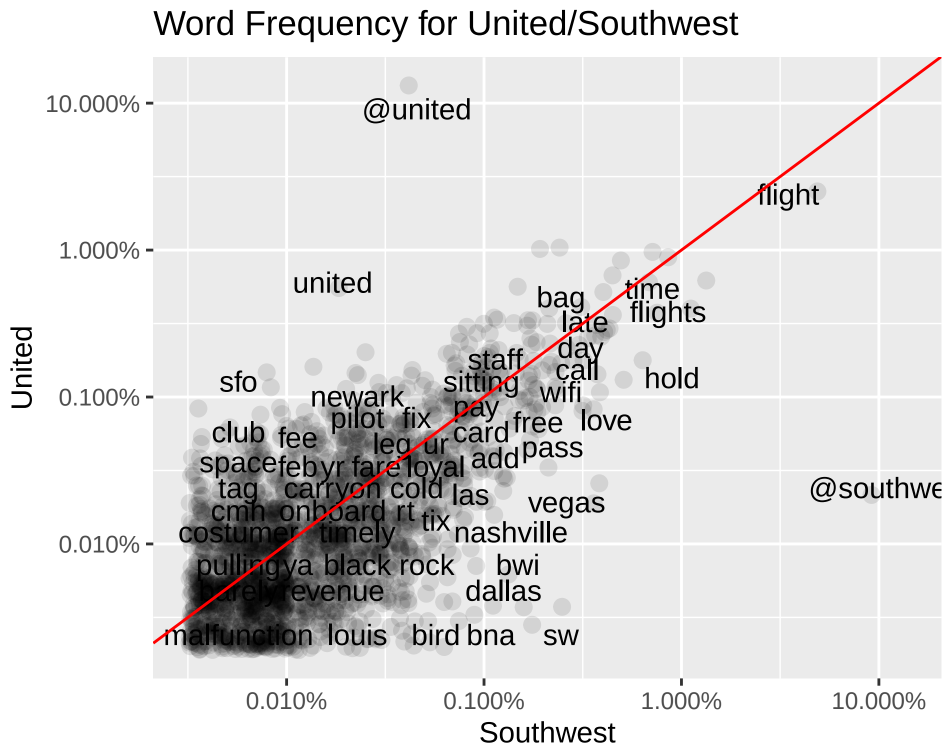
Words near the line represents the words tweeted about Delta/American with the same amount of frequency while words away from the line represents tweets that are tweeted about Delta/American more than the other. In the chart above, we note that words such as called, hour, weather, lack, flight are the most frequent words tweeted about American airlines and Delta.

```
ggplot(frequency, aes(USAirways, VirginAmerica)) +
  geom_jitter(alpha = 0.1, size = 2.5, width = 0.25, height = 0.25) +
  geom_text(aes(label = word), check_overlap = TRUE, vjust = 1.5) +
  scale_x_log10(labels = percent_format()) +
  scale_y_log10(labels = percent_format()) +
  geom_abline(color = "red") + labs(title = "Word Frequency for US Airway/Virgin America ")
```



Words frequently tweeted about Virgin America and US Airways include flights, online, book, and reschedule.

```
ggplot(frequency, aes(Southwest, United)) +
  geom_jitter(alpha = 0.1, size = 2.5, width = 0.25, height = 0.25) +
  geom_text(aes(label = word), check_overlap = TRUE, vjust = 1.5) +
  scale_x_log10(labels = percent_format()) +
  scale_y_log10(labels = percent_format()) +
  geom_abline(color = "red") + labs(title = "Word Frequency for United/Southwest ")
```



Words frequently tweeted about United and Southwest include flights, service, late, and delay.

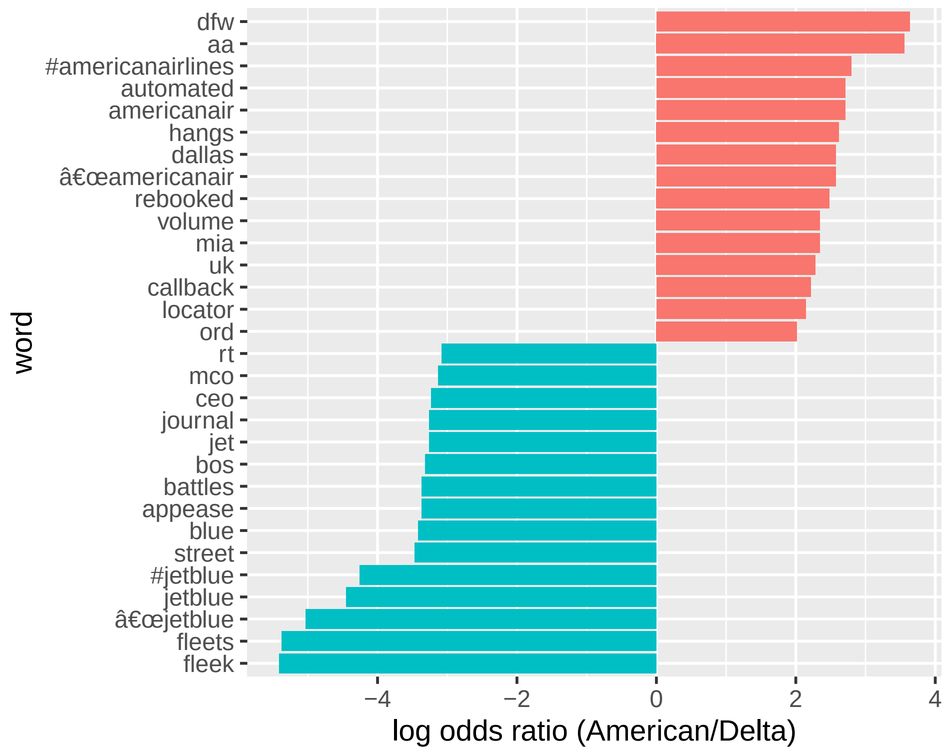
Comparing word usage

Previously I plotted which words were used frequently by each airlines, now I can calculate which words are more likely to be associated to which airlines by using the log odds ratio. In this calculation, we count the number of words that is used more than 50 times and we use that word to calculate the log odds ratio.

The reason why I thought that this is helpful is because these words could be used by top level management to look at the specific 'negativereasons' that people are tweeting about. For example, if the word refund is associated with United Airlines, people might want a refund for a certain reason and top management can make a decision based on the sentiment of passengers.

```
AmericanDelta <- tidy_tweets %>%
  filter(!str_detect(word, "^@")) %>%
  count(word, airline) %>%
  group_by(word) %>%
  filter(sum(n) >= 10) %>%
  ungroup() %>%
  spread(airline, n, fill = 0) %>%
  mutate_if(is.numeric, list(~(. + 1) / (sum(.) + 1))) %>%
  mutate(logratio = log(American / Delta)) %>%
  arrange(desc(logratio))

AmericanDelta %>%
  group_by(logratio < 0) %>%
  top_n(15, abs(logratio)) %>%
```

```
  ungroup() %>%
  mutate(word = reorder(word, logratio)) %>%
  ggplot(aes(word, logratio, fill = logratio < 0)) +
  geom_col(show.legend = FALSE) +
  coord_flip() +
  ylab("log odds ratio (American/Delta)") +
  scale_fill_discrete(name = "", labels = c("American", "Delta"))
```



In the graph above, we see the words such as rebooked, callback, automated, volume, dfw are associated with American Airlines while words such as fleek, jet, ceo and account is associated with Delta. Please take note that acronyms like dfw represents terminals e.g. dfw for Dallas Fortworth.

This shows that most people are tweeting about American Airlines with regards to the terminal DFW. This could be a sign that a lot of complaints are coming from that terminal. On the other hand, people tweeting about Delta is talking about their fleet of jets and the CEO of the company.

```
USVirgin_ratios <- tidy_tweets %>%
  filter(!str_detect(word, "^@")) %>%
  count(word, airline) %>%
  group_by(word) %>%
  filter(sum(n) >= 10) %>%
  ungroup() %>%
  spread(airline, n, fill = 0) %>%
  mutate_if(is.numeric, list(~(. + 1) / (sum(.) + 1))) %>%
  mutate(logratio = log(USAirways / VirginAmerica)) %>%
  arrange(desc(logratio))
```

```
USVirgin_ratios %>%
  group_by(logratio < 0) %>%
  top_n(15, abs(logratio)) %>%
  ungroup() %>%
  mutate(word = reorder(word, logratio)) %>%
  ggplot(aes(word, logratio, fill = logratio < 0)) +
  geom_col(show.legend = FALSE) +
  coord_flip() +
  ylab("log odds ratio (USAirways/VirginAmerica)") +
  scale_fill_discrete(name = "", labels = c("USAirways", "VirginAmerica"))
```

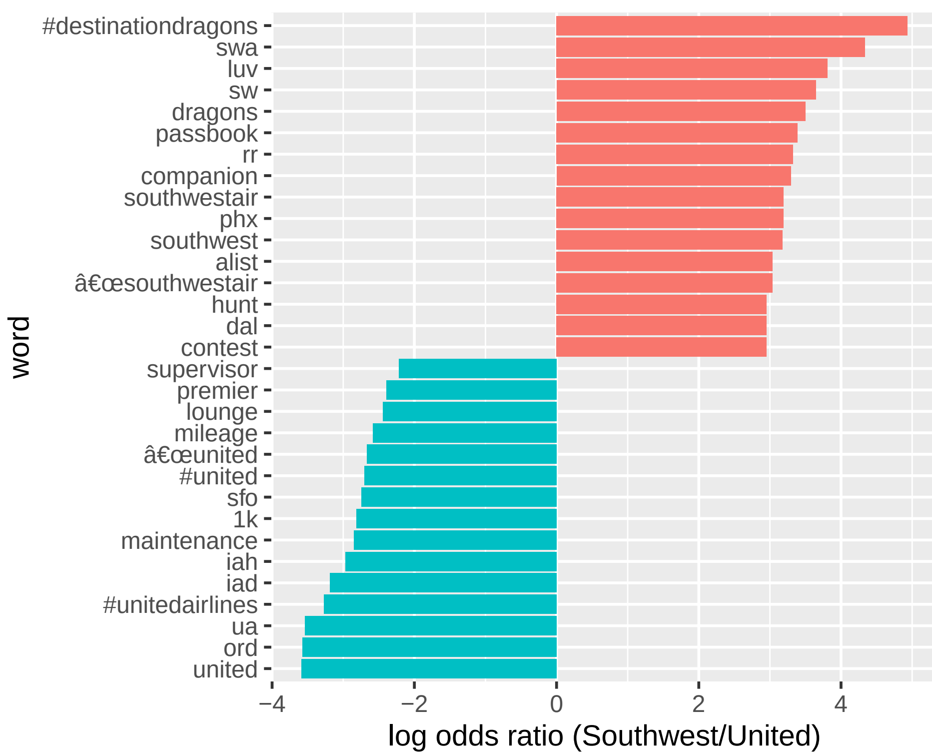
The words associated with US Airways include miles, charlotte, connection, rude. Words associated with Virgin America include incredible, luv, deals, omg.

People might think that the staff working for US Airways is rude while Virgin America has more positive words associated to them.

Southwest/United

```
SouthUnited_ratios <- tidy_tweets %>%
  filter(!str_detect(word, "^@")) %>%
  count(word, airline) %>%
  group_by(word) %>%
  filter(sum(n) >= 10) %>%
  ungroup() %>%
```

```
  spread(airline, n, fill = 0) %>%
  mutate_if(is.numeric, list(~(. + 1) / (sum(.) + 1))) %>%
  mutate(logratio = log(Southwest / United)) %>%
  arrange(desc(logratio))


SouthUnited_ratios %>%
  group_by(logratio < 0) %>%
  top_n(15, abs(logratio)) %>%
  ungroup() %>%
  mutate(word = reorder(word, logratio)) %>%
  ggplot(aes(word, logratio, fill = logratio < 0)) +
  geom_col(show.legend = FALSE) +
  coord_flip() +
  ylab("log odds ratio (Southwest/United)") +
  scale_fill_discrete(name = "", labels = c("Southwest", "United"))
```



Words associated to Southwest airlines include a hashtag destinationdragons, companion, alist and contest. People tweeting about United airline are talking about their maintenance, SFO, supervisors and IAD.

During that period, Southwest airlines had partnership with award-winning rock band, Imagine Dragons, to let people enter a contest to have the most unique concert in the world - 35000 feet in the air. This shows that their campaign has been widely talked about on twitter or the hashtag was used for a lucky draw on twitter. As for United airlines, people were talking about supervisors and the 1k, which is a premier status in their mileage program. People were also talking about the lounge and mileage which is associated with the 1k status.

Limitations

One limitation from this word frequency and word usage is that only 2 airlines can be compared to one another on a single model. I was trying to find a way to compare them and plot them on a multidimensional graph but I wasn't able to. I think if I was able to do that, the results of the word usage and frequency would come out differently since all the airlines are being compared with one another.

Key Findings

Overall, United Airlines was the most tweeted about from 16th February 2015 to 24th February 2015. Approximately 62% of the tweets are negative. For most airlines, the reason behind these negative tweets are because of customer service issue followed by late flights. This is something for top level managers to keep in mind and to improve on. Majority of the tweets are tweeted on Sunday and Monday, where people might be returning home from a short vacation and tweeting about their experiences.

Comparing word frequencies allow us to see what users are tweeting about each airlines and comparing the word usage and association will allow us to see what words are more likely to be associated with the airlines. That being said, airlines can see if their marketing campaigns are successful from text mining. This is shown by Southwest's #destinationdragons campaign where it got a lot of mentions on twitter during that time period. In addition, it also allows airlines to focus on customer's pain points and effective come up with solutions to address them.