Text Analysis for Delta Airlines
- Business Insight Report –

by Fabian Jaskotka

HULT
INTERNATIONAL
BUSINESS SCHOOL

Hult International Business School
Thomas Kurnicki

Introduction

The airline industry is constantly evolving and does not have a lot in common with how it was 100 years ago (Ben-Yosef, 2007). Known for its growth powered by a doubling number of international and domestic travelers from 1999 to 2019, even throughout crisis such as 9/11 or the financial crisis, the industry is struggling like never before. COVID-19 and its associated travel bans have hit the airline sector hard, resulting in losses of an estimated $315 billion (Borko, Geerts, & Wang, 2020). 20 airlines with more than 10 aircraft in their fleet have already failed since the beginning of the pandemic (Ng, 2020). In order to survive, airlines shall focus on putting people first and adapting to consumer shifts. Therefore, consumers need to be understood (Terry, 2020).
This report shall uncover insights for the United States' second largest airline based on passenger numbers, Delta. Associations with the airline as well as consumer sentiment will be analyzed and comparisons with Delta's major competitors, American Airlines and United Airlines, will be drawn.

Associations with Delta

To analyze what Delta is associated with amongst consumers, 1902 Twitter posts mentioning "Delta Airlines" have been loaded. After eliminating a flight-tracking bot, 660 posts were tokenized and analyzed by their frequency. A frequency table of the most occurring words in these 660 posts is illustrated in Figure 1.

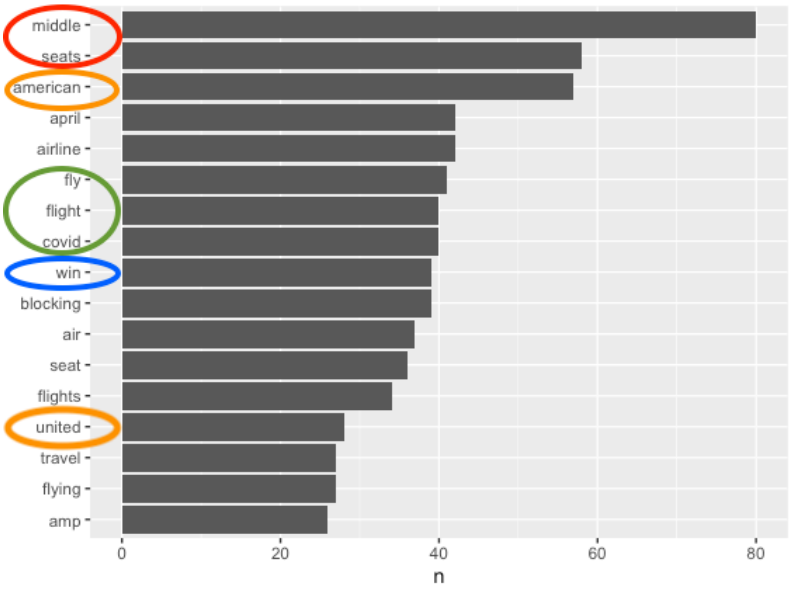


Figure 1: Histogram with frequencies of keywords

As can be derived from the figure, the most common word associated with the airline is "middle", followed by "seats". Potentially, people might be talking about Delta's announcement to be the only US airline to continue to leave middle seats blocked on their flights for the time being (Delta Airlines, 2020).

Moreover, also the terms “american” and “united” are mentioned frequently in the tweets, potentially referring to America’s largest airlines and Delta’s competitors American Airlines and United Airlines. A close examination reveals that indeed these posts regard the two main rivals of Delta.

Another insight from the histogram of frequently used words in tweets about Delta Airlines regards the ongoing pandemic. The words “fly”, “flight” as well as “covid” have been used, allowing for estimation that people are concerned about their flights due to COVID-19.

Finally, the word “win” appears 39 times throughout the posts. This could be referred to Delta performing well in the current situation or a potential lottery offered by the airline.

Topics of Conversation Across the Airlines

To compare the topics of conversation at delta with its major competitors American Airlines and United Airlines, a correlogram has been developed (Figure 2).

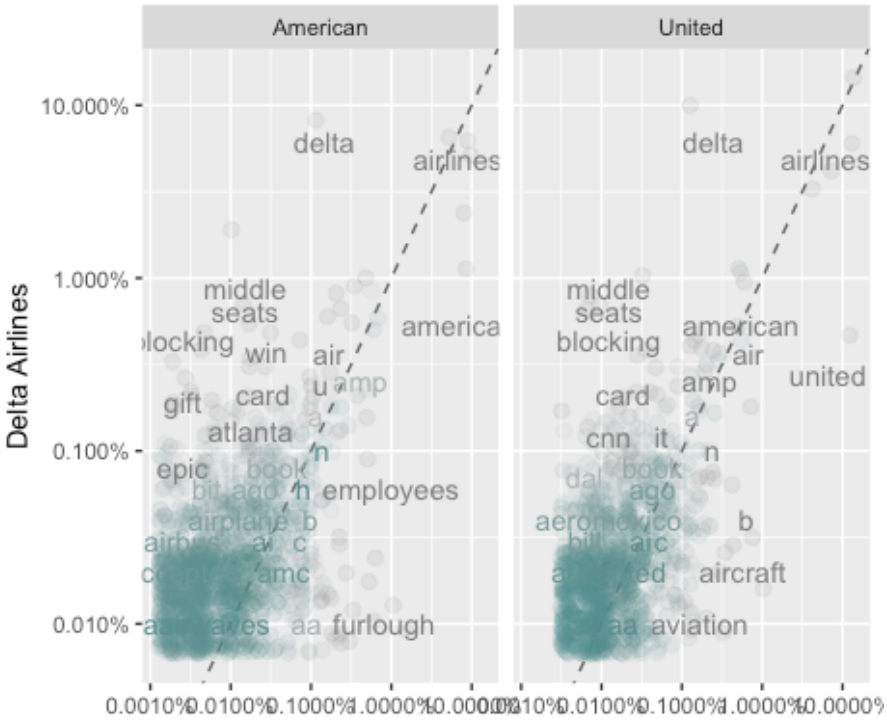


Figure 2: Correlogram comparing relative frequency of terms

The correlogram reveals that the keywords “middle”, “seat”, and “blocking” are more used in association with Delta than with both of the other airlines. Moreover, they occur 55 times as a combination throughout the tweets analyzed. This proves that Delta truly is the only airline in the trio to block middle seats on their flights.

When comparing the conversation of Delta and American in the left diagram in Figure 2, it can be noted, that also the words “win”, “gift” and “card” are mentioned more often with Delta. A

detailed analysis reveals that this regards a promotion where people could win a $100 gift card. On the other hand, tweets mentioning American much more often include the word “furlough” (= vacation). In order to be more associated with this term, Delta could launch a vacation campaign to win over consumers that currently prefer choosing American for their leisure travel.

Matching the conversation of Delta with United (Figure 2, right), it can be seen that “cnn” is more frequently mentioned in Delta tweets. Further analysis could reveal why CNN seems to be focused more on Delta than on United. Furthermore, the terms “aircraft” as well as “aviation” occur more with United related tweets. Since Delta aircraft are on average 2 years younger than United planes, the carrier could set a focus on their state-of-the-art equipment such as the Airbus A350 in their advertisement strategy (Airfleets, 2021; Airfleets, 2021).

Correlation analysis reveals that Delta Airlines tweets correlate very similarly with the other two airlines. Delta tweets correlate with topics of American for about 75% while they correlate with United 76%. This means that overall topics might be shared throughout the airlines, however, there are differences that can be analyzed in more detail and used in Delta’s diversification strategy. Figure 3 illustrates that the most powerful words specifically for Delta appear to be “win”, “smart”, “empty” as well as “nightmare”.

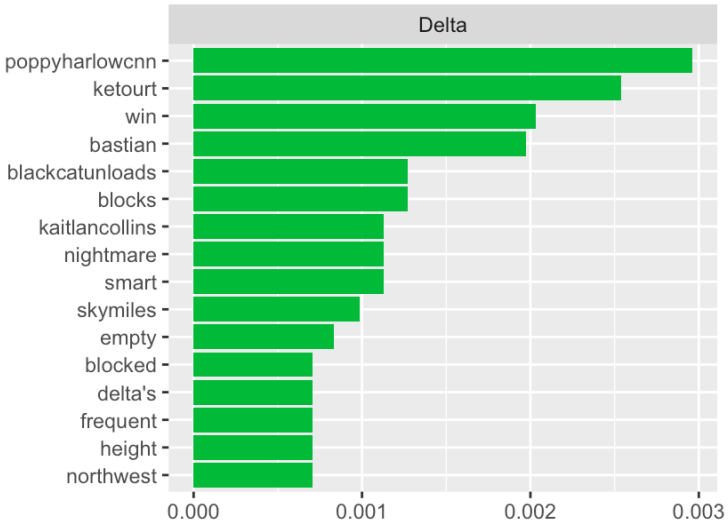


Figure 3: TF-IDF chart with relatively most unique keywords

Sentiment Analysis

The sentiment of consumer’s comments about an airline is closely linked to its current reputation and recent happenings. United Airlines experienced that in 2017 after their April 9 passenger incident (Freeberg, 2017). In order to be able to react quickly, the sentiment of consumers shall be monitored closely.

A sentiment cloud for Delta as shown in Figure 4 reveals that by the time of writing, the overall sentiment is leaning towards “disgust” and “surprise”, followed by “joy”. Most impactful for Delta’s reputation is the disgust sentiment as negative memories appear to be easier to

remember than positive ones (Warner, 2007). Therefore, Delta should focus on the elimination of the disgust sentiment amongst its consumers, potentially by taking extra care of cleanliness and appearance of staff and equipment.

Figure 4: Comparison Cloud with Sentiment Flavors

As the surprise element sentiment can be seen as both, positive and negative, a sentiment cloud with a binary analysis is generated to more easily examine development potential (Figure 5). It can be noted, that tweets including the terms “horrible”, “refusal”, and “downturn” need to be further investigated to draw learnings from the specific situations.

Figure 5: Binary Sentiment Cloud

Conclusion

For Delta Airlines to be able to focus on people in order to not only survive, but thrive in these difficult times, it is important to understand what consumers are saying. Keyword (token) examination as well as sentiment analysis have been conducted and the following insights could be derived:

1. Keep making great impact with diversification announcements as the currently most discussed topic related to Delta's news is their decision to keep middle seats blocked.
2. Emphasize safety of flying in times of a pandemic as consumers appear to be worried and uncertain.
3. Launch a campaign with emphasize on vacation to take over holiday travel conversation from American Airlines.
4. Focus on aircraft age and new Airbus A350s to win over United's aircraft conversation.
5. Analyze details of diversification between the three airlines to separate more clearly with positively related keywords
6. Take extra care of cleanliness and appearance of staff and equipment in order to shift overall sentiment away from disgusted.
7. Analyze tweets including the keywords "horrible", "refusal" or "downturn" to understand the origin of negatively related sentiment.

With a successful execution of these insights, Delta might be able to continue being America's most successful airline. However, it should be noted that this report is a snapshot analysis and should therefore be renewed regularly to keep insights up to date and track progress.

Bibliography

Airfleets. (2021, February 10). *Airline Fleet Age*. Retrieved from Airfleets: https://www.airfleets.net/ageflotte/Delta%20Air%20Lines.htm

Airfleets. (2021, February 10). *United Airlines fleet details*. Retrieved from Airfleets: https://www.airfleets.net/ageflotte/United%20Airlines.htm

Ben-Yosef, E. (2007). The Evolution of the US Airline Industry: Technology, Entry, and Market Structure - Three Revolutions. *Journal of Air Law and Commerce, 72*(2), 305-349.

Borko, S., Geerts, W., & Wang, H. (2020). *The Travel Industry Turned Upside Down.* McKinsey & Company.

Delta Airlines. (2020, November 18). *Delta the only U.S. airline to block middle seats, limit onboard capacity through March 30, 2021*. Retrieved from Delta News Hub: https://news.delta.com/delta-only-us-airline-block-middle-seats-limit-onboard-capacity-through-march-30-2021

Freeberg, J. (2017, May 1). *Harris Poll: United Airlines' Corporate Reputation Takes A Nose Dive*. Retrieved from Cision: https://www.prnewswire.com/news-releases/harris-poll-united-airlines-corporate-reputation-takes-a-nose-dive-300448557.html

Ng, A. (2020, October 8). *Over 40 airlines have failed so far this year — and more are set to come*. Retrieved from CNBC: https://www.cnbc.com/2020/10/08/over-40-airlines-have-failed-in-2020-so-far-and-more-are-set-to-come.html

Terry, B. (2020). *COVID-19: rising to thechallenge with resilience.* Deloitte.

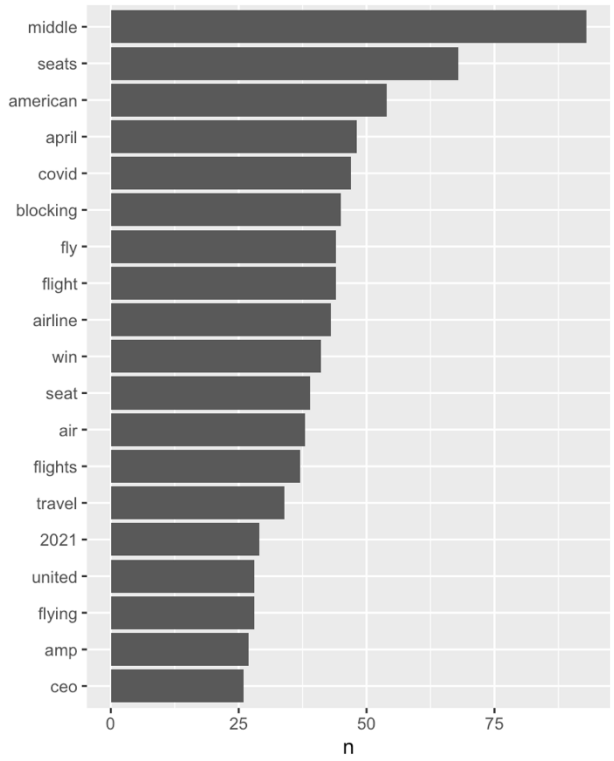
Warner, J. (2007, August 29). *Bad Memories Easier to Remember*. Retrieved from WebMD: https://www.webmd.com/brain/news/20070829/bad-memories-easier-to-remember

Appendix

```
################################################
### Created by Fabian Jaskotka on 02/06/2021 ###
########### BUSINESS INSIGHT REPORT  ###########
################################################

#loading necessary libraries
library(dplyr)
library(rtweet)
library(tidytext)
library(tidyverse)
library(stringr)
library(ggplot2)
library(scales)
library(tm)
library(twitteR)

# Change consumer_key, consume_secret, access_token, and
# access_secret based on your own keys
#consumer_key <- "my_token_1"
#consumer_secret <-"my_token_2"
#access_token <- "my_token_3"
#access_secret <- "my_token_4"
setup_twitter_oauth(consumer_key, consumer_secret, access_token,
access_secret)

#############################################################################
######## Importing Data from Twitter - Aviation
#############################################################################


##################
# Delta
##################

#downloading Delta related tweets
Delta = searchTwitter('delta airlines -filter:retweets', n = 10000,
                      resultType = "recent", lang = "en") %>%
          twListToDF()

#investigating irregularly many posts from one user
Delta %>%
  count(screenName, sort = T)

#transforming into dataset
my_df_1 <- data.frame(Delta, text = Delta$text)

#adding column with search word
my_df_1$company <- "Delta"


##################
# American
##################
```

```
#downloading American related tweets
American = searchTwitter('american airlines -filter:retweets', n = 10000,
                         resultType = "recent", lang = "en") %>%
          twListToDF()

#transforming data into dataset
my_df_2 <- data.frame(American, text = American$text)

#adding column with search word
my_df_2$company <- "American"


##################
# United
##################

#downloading United related tweets
United = searchTwitter('united airlines -filter:retweets', n = 10000,
                       resultType = "recent", lang = "en") %>%
          twListToDF()

#transforming data into dataset
my_df_3 <- data.frame(United, text = United$text)

#adding column with search word
my_df_3$company <- "United"


################################################################################
######## Tokenizing dataframe
################################################################################

#tokenizing Delta
token_list_1 <- my_df_1 %>%
                  filter(screenName != "laxradar") %>%
                  unnest_tokens(word, text) %>%
                  anti_join(stop_words)

token_1_clean <- token_list_1 %>%
                  count(word,sort = T)

#tokenizing American
token_list_2 <- my_df_2 %>%
                  filter(screenName != "laxradar") %>%
                  unnest_tokens(word, text) %>%
                  anti_join(stop_words)

#tokenizing United
token_list_3 <- my_df_3 %>%
                  filter(screenName != "laxradar") %>%
                  unnest_tokens(word, text) %>%
                  anti_join(stop_words)

################################################################################
```

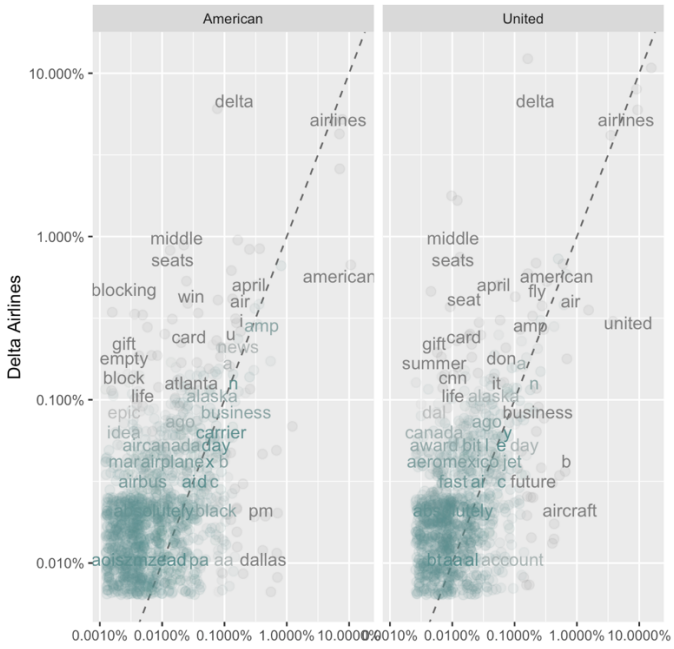
```
####### Plotting Frequency of Delta Words
################################################################################

freq_hist <- token_list_1 %>%
              count(word, sort = TRUE) %>%
              mutate(word = reorder(word,n )) %>%
              filter(n>25) %>%
              filter(word != "delta") %>%
              filter(word != "t.co") %>%
              filter(word != "https") %>%
              filter(word != "airlines") %>%
            ggplot(aes(word, n))+
            geom_col() +
            xlab(NULL) +
            coord_flip()

print(freq_hist)
```



```
################################################################################
####### Combining Data and creating Frequency Table
################################################################################

#sorting by frequency
frequency_tokens <-bind_rows(mutate(token_list_1, company="Delta"),
                             mutate(token_list_2, company="American"),
                             mutate(token_list_3, company="United")) %>%
                            mutate(word=str_extract(word, "[a-z']+")) %>%
                            count(company, word) %>%
                            group_by(company) %>%
                            mutate(proportion = n/sum(n))%>%
                            select(-n) %>%
                            spread(company, proportion) %>%
                            gather(company, proportion, `American`,
`United`)

print(frequency_tokens)
```

```
# A tibble: 23,970 x 4
   word           Delta company  proportion
   <chr>          <dbl> <chr>         <dbl>
 1 a           0.00203  American  0.00113
 2 aa          0.000127 American  0.000979
 3 aaa        NA        American  0.0000245
 4 aaaaaaaaa   0.000127 American NA
 5 aaaand     NA        American NA
 6 aacenter   NA        American  0.0000490
 7 aadvantage NA        American  0.000294
 8 aadwydunnt NA        American  0.0000245
 9 aafrc      NA        American  0.0000245
10 aairwaves  NA        American  0.0000245
# … with 23,960 more rows
```

```
################################################################################
####### Correlogram
################################################################################

#plotting correlogram
ggplot(frequency_tokens, aes(x=proportion, y=`Delta`,
                      color = abs(`Delta`- proportion)))+
  geom_abline(color="grey40", lty=2)+
  geom_jitter(alpha=.1, size=2.5, width=0.3, height=0.3)+
  geom_text(aes(label=word), check_overlap = TRUE, vjust=1.5) +
  scale_x_log10(labels = percent_format())+
  scale_y_log10(labels= percent_format())+
  scale_color_gradient(limits = c(0,0.001), low = "darkslategray4", high =
"gray75")+
  facet_wrap(~company, ncol=2)+
  theme(legend.position = "none")+
  labs(y= "Delta Airlines", x=NULL)
```




```
#############################################################################
####### Correlation Test
#############################################################################

#correlation between Delta and American
cor.test(data = frequency_tokens[frequency_tokens$company == "American", ],
         ~proportion + `Delta`)
```

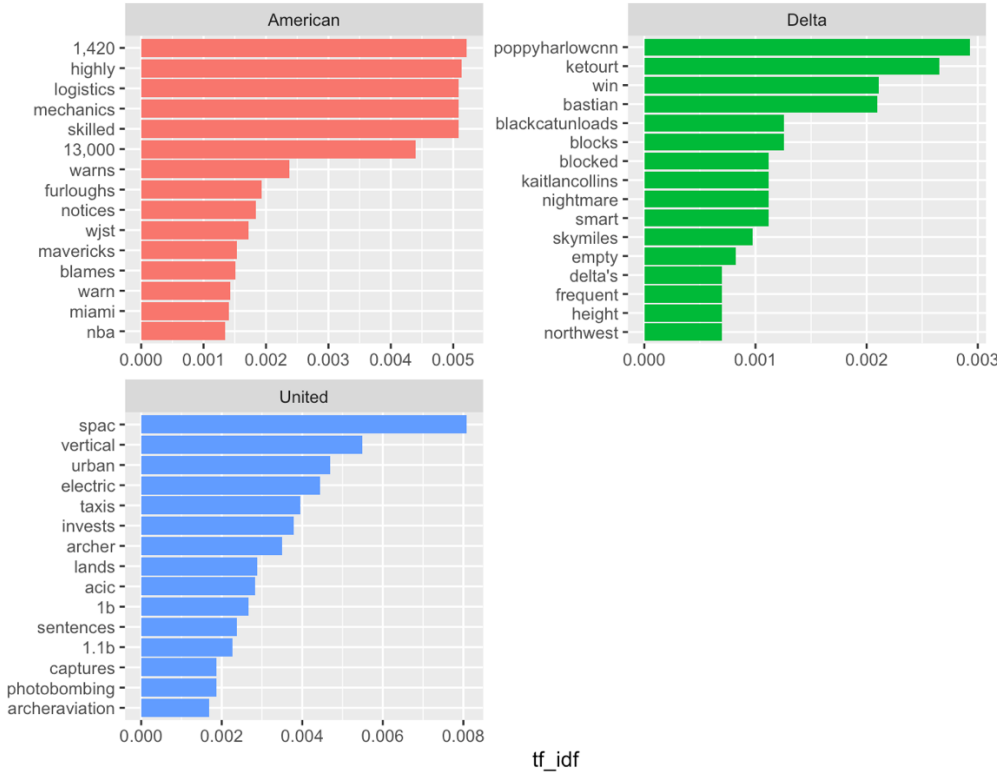
```
        Pearson's product-moment correlation

data:  proportion and Delta
t = 41.572, df = 1359, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7238371 0.7706924
sample estimates:
      cor
0.7481962
```

```
#correlation between Delta and United
cor.test(data = frequency_tokens[frequency_tokens$company == "United", ],
         ~proportion + `Delta`)
```

```
        Pearson's product-moment correlation

data:  proportion and Delta
t = 38.989, df = 1129, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7315227 0.7813040
sample estimates:
      cor
0.7575124
```

```
##############################################################################
####### Correlation Test
##############################################################################

#correlation between Delta and American
cor.test(data = frequency_tokens[frequency_tokens$company == "American", ],
         ~proportion + `Delta`)

#correlation between Delta and United
cor.test(data = frequency_tokens[frequency_tokens$company == "United", ],
         ~proportion + `Delta`)

##############################################################################
####### N-Grams
##############################################################################

all_airlines <- bind_rows(mutate(my_df_1, company = "Delta"),
                          mutate(my_df_2, company = "American"),
                          mutate(my_df_3, company = "United"))

airlines_bigrams <- all_airlines %>%
  unnest_tokens(bigram, text, token = "ngrams", n=3) %>%
  filter(!is.na(bigram)) %>%
  count(company, bigram, sort = TRUE) %>%
  filter(company =="Delta")


##############################################################################
####### TF-IDF
##############################################################################

#adding data from airlines together
combination <- bind_rows(mutate(token_list_1, company="Delta"),
                         mutate(token_list_2, company="American"),
                         mutate(token_list_3, company="United"))

combination <- combination %>%
  count(company,word, sort = T)

#sorting total words by count per company
total_words <- combination %>%
  group_by(company) %>%
  count(company, word, sort = T) %>%
  summarize(total = sum(n))

company_words <- left_join(combination, total_words)

#adding tf_idf scores to dataset
company_words <- company_words %>%
  bind_tf_idf(word, company, n) %>%
  arrange(desc(tf_idf))

#graphical output
```

```
company_words %>%
  arrange(desc(tf_idf)) %>%
  mutate(word = factor(word, levels= rev(unique(word)))) %>%
  group_by(company) %>%
  top_n(15) %>%
  ungroup %>%
  ggplot(aes(word, tf_idf, fill = company)) +
  geom_col(show.legend = F) +
  labs(x =NULL, y = "tf_idf") +
  facet_wrap(~company, ncol=2, scales="free")+
  coord_flip()
```




```
##############################################################################
####### Sentiment Analysis
##############################################################################

# flavor sentiment
token_1_clean %>%
  inner_join(get_sentiments("nrc")) %>%
  count(word, sentiment, sort = T) %>%
  acast(word ~sentiment, value.var = "n", fill=0) %>%
  comparison.cloud(color= c("grey10", "grey60"),
                   max.words = 200, scale =c(2, 0.1), random.order = T,
                   title.size = 2)
```


```
#binary sentiment
token_1_clean %>%
  inner_join(get_sentiments("bing")) %>%
  count(word, sentiment, sort = T) %>%
  acast(word ~sentiment, value.var = "n", fill=0) %>%
  comparison.cloud(color= c("grey10", "grey60"),
                   max.words = 200, scale =c(3, 0.1), random.order = T,
                   title.size = 2)
```