# Airline Data Sentiment Analysis

Gursheen Kaur Anand(221)

- Preparations
- Sentiment Analysis
- General Info
  - Preliminary visual inspection
  - Findings
- Most frequent terms used and a Wordcloud
- Basic Model
  - Document data
  - Seperate data
  - Randomforest

## Preparations

Loading Libraries    Loading Data

```
library(dplyr)   #Data manipulation
library(forcats) #ggplot frequency
library(ggplot2) #visualizations
library(caTools) #Data wrangling
library("tm")  # for text mining
library("SnowballC") # for text stemming
library("wordcloud") # word-cloud generator
library("RColorBrewer") # color palettes
library(randomForest) #randomforest
library(sentimentr)
```

## Sentiment Analysis

```r
df$ave_sentiment=0
df$sentiment=0
df$positivereason="Unspecified"
df$negativereason="Unspecified"
for(i in 1:15){
  sentiment=sentiment_by(df[i,4])
  df[i,5]=sentiment$ave_sentiment
  if(sentiment$ave_sentiment<0){
    df[i,6]='negative'
    df[i,9]=df[i,7]
  }
  else if(sentiment$ave_sentiment==0){
    df[i,6]='neutral'
  }
  else if(sentiment$ave_sentiment>0){
    df[i,6]='positive'
    df[i,8]=df[i,7]
  }
}
```

# General Info

```r
df
```

| tweet_id | airline | name |
|---:|---|---|
| <dbl> | <chr> | <chr> |
| 5.70e+17 | Virgin America | Rebecca Morales |
| 5.80e+17 | Virgin America | Jessica Wong |
| 5.90e+17 | Virgin America | Laura Dominguez |
| 6.70e+00 | United | Aaron Horne |
| 6.87e+00 | United | Vanessa Gonzalez |
| 2.60e+17 | United | James Guerrero |

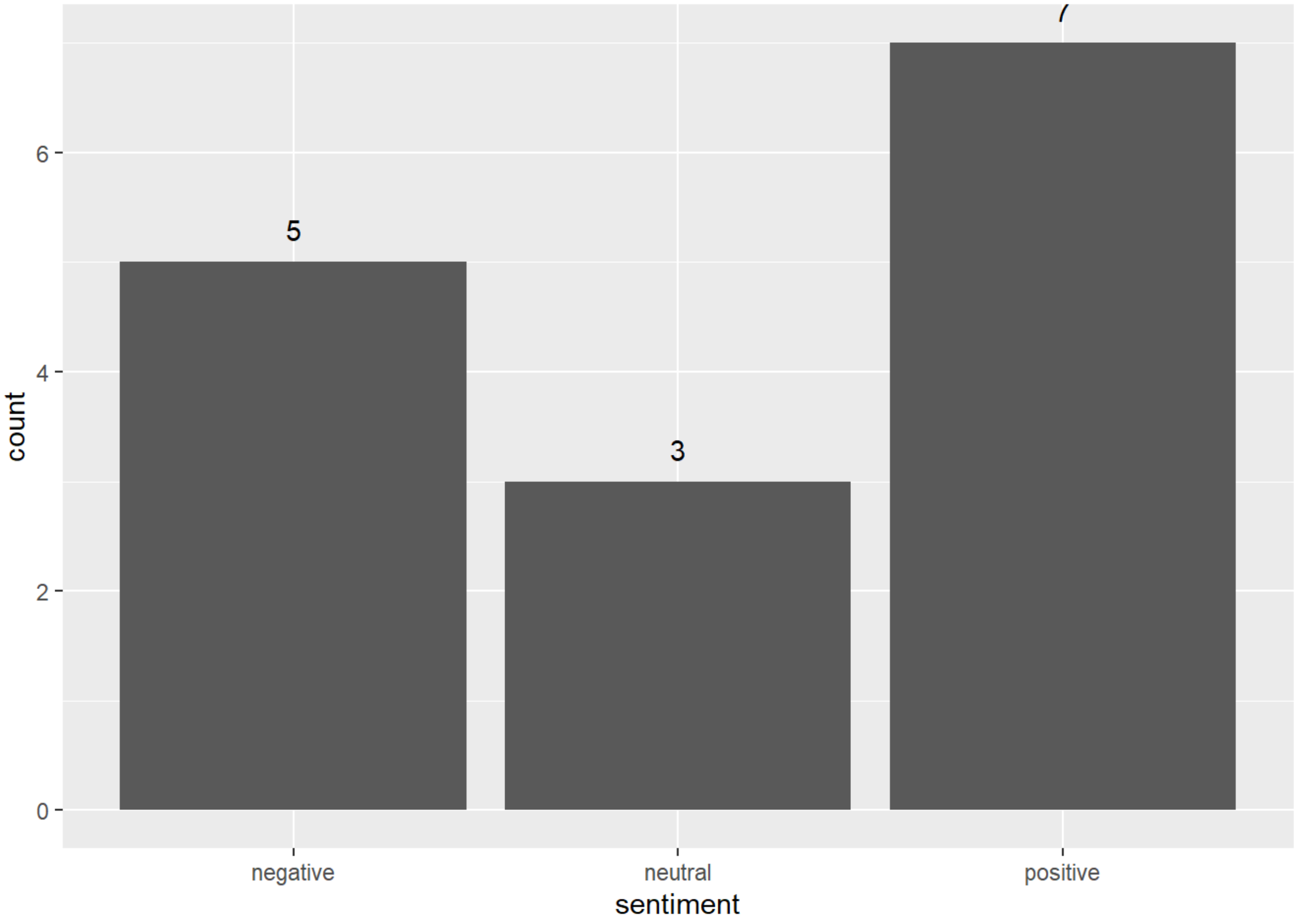
| tweet_id | airline | name | ▶ |
|---:|---|---|---|
| <dbl> | <chr> | <chr> | |
| 8.90e+17 | Delta | Daniel Diaz | |
| 5.50e+17 | Delta | Luis Stevens | |
| 4.90e+17 | Delta | James Smith | |
| 4.80e+17 | US Airways | Adam Robinson | |

1-10 of 15 rows | 1-3 of 9 columns Previous **1** 2 Next

# Preliminary visual inspection

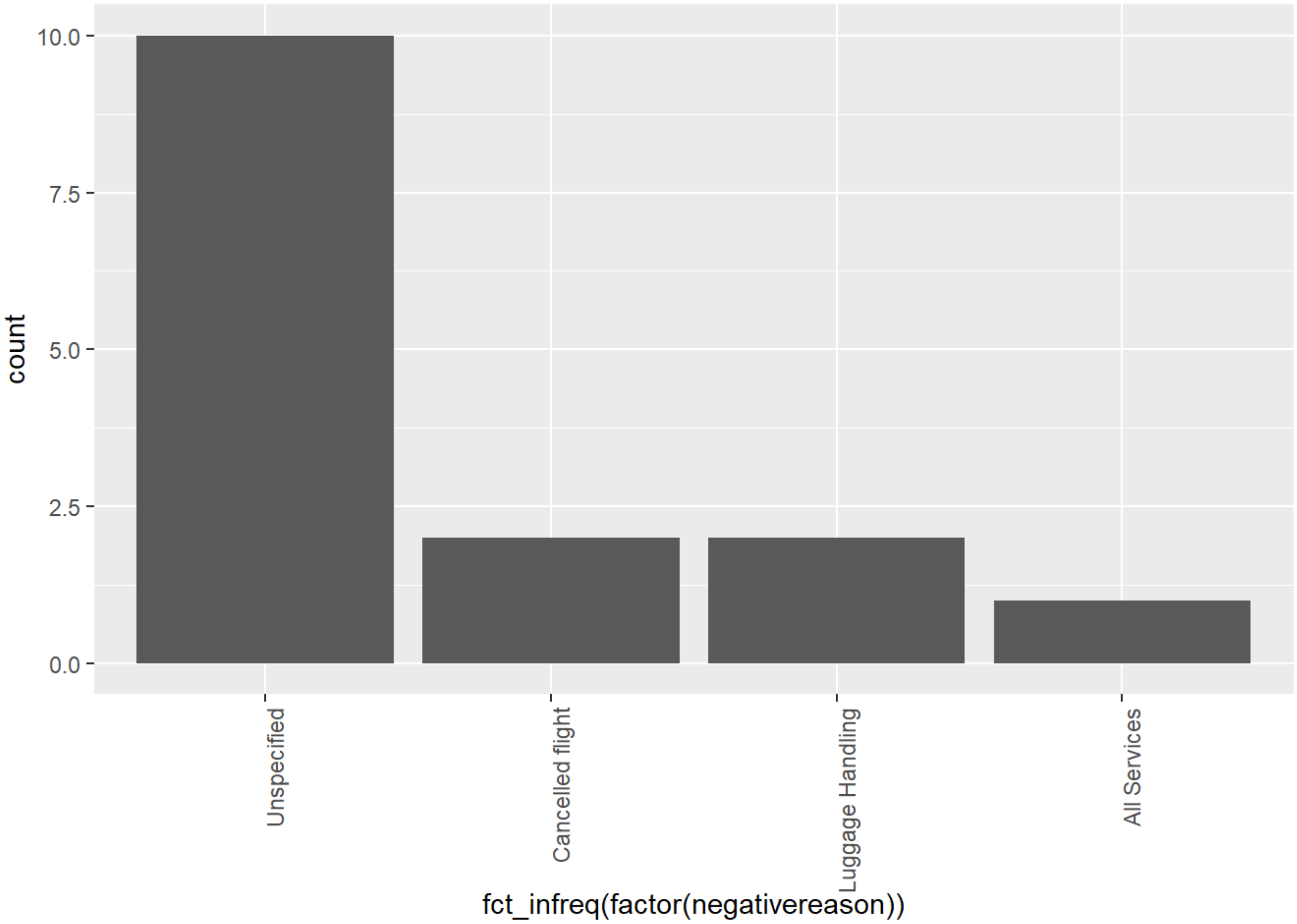
## Total distribution of tweets with sentiment

```
#Total tweets distribution as negative,positive,neutral
ggplot(df, aes(x = sentiment))+geom_bar(stat = "count")+
geom_text(stat='count', aes(label=..count..), vjust=-1)
```



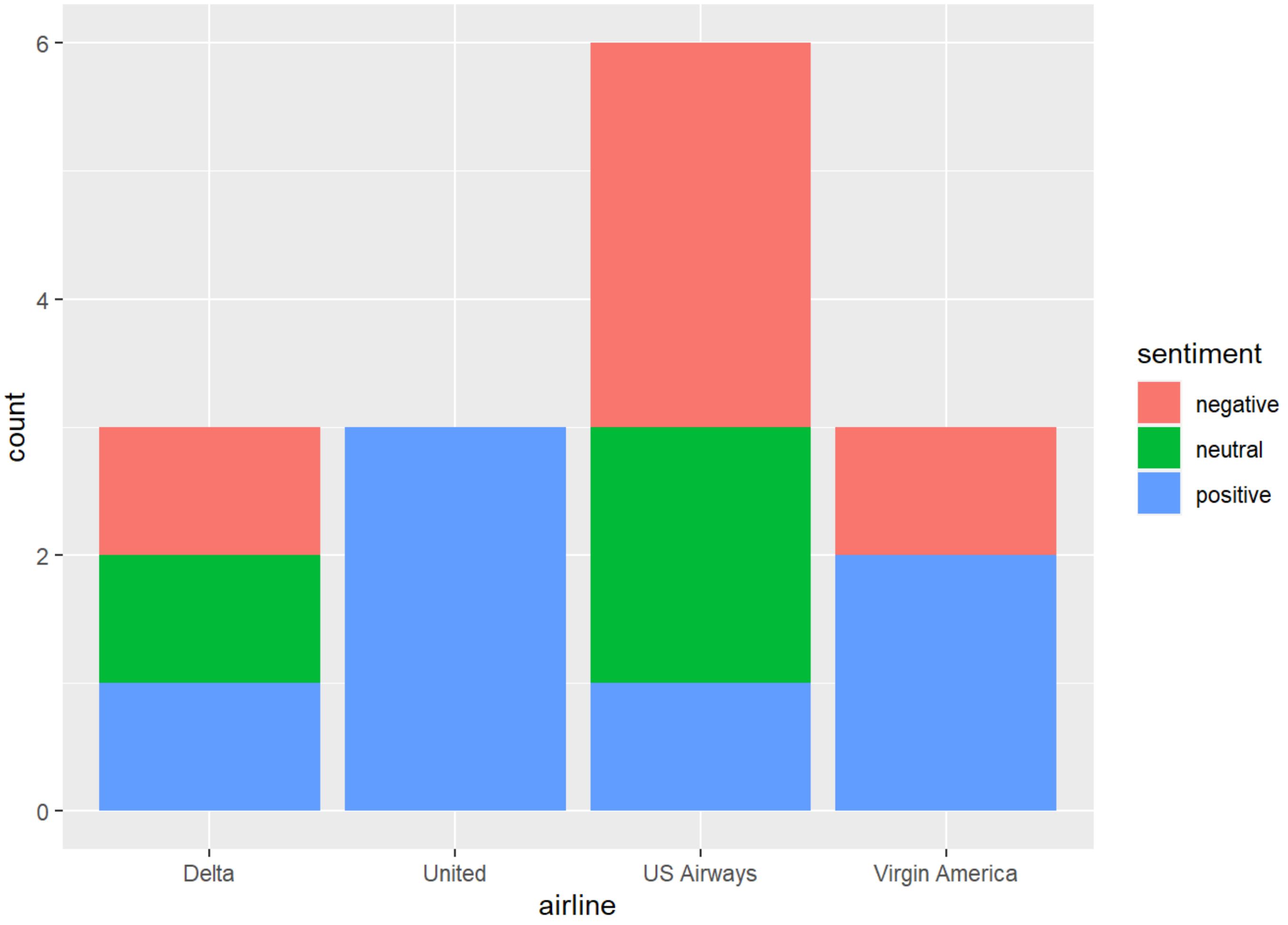
## Distribution of Negative Tweets and their Reasons

```
#fct_infreq is from package forcats (for frequency distribution)
ggplot(df, aes(x = fct_infreq(factor(negativereason))))+geom_bar(stat =
"count")+theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

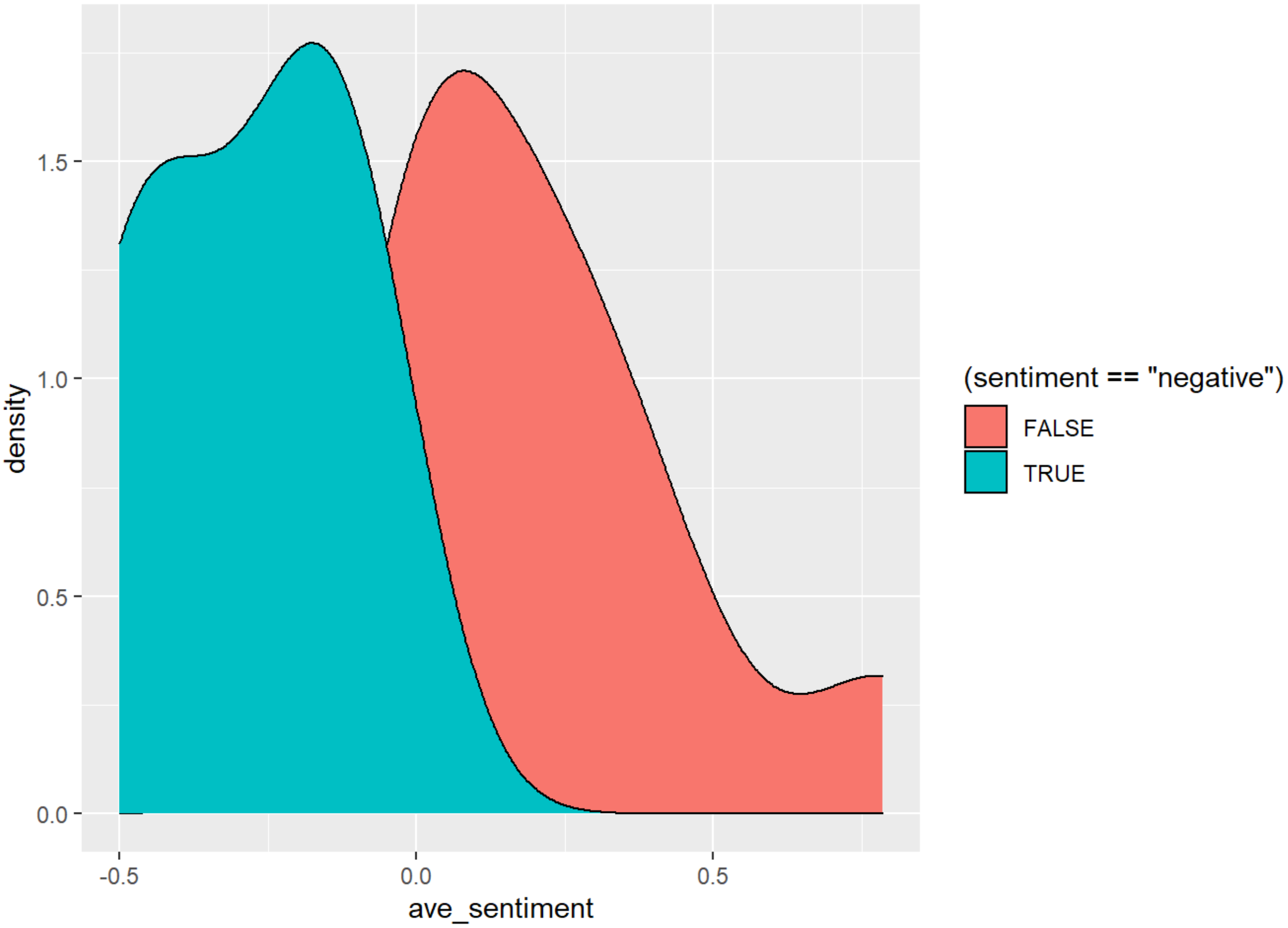


## Distribution of Sentiments for different Airlines.

```
ggplot(df, aes(x = airline,fill = sentiment ))+geom_bar(stat = "count")
```

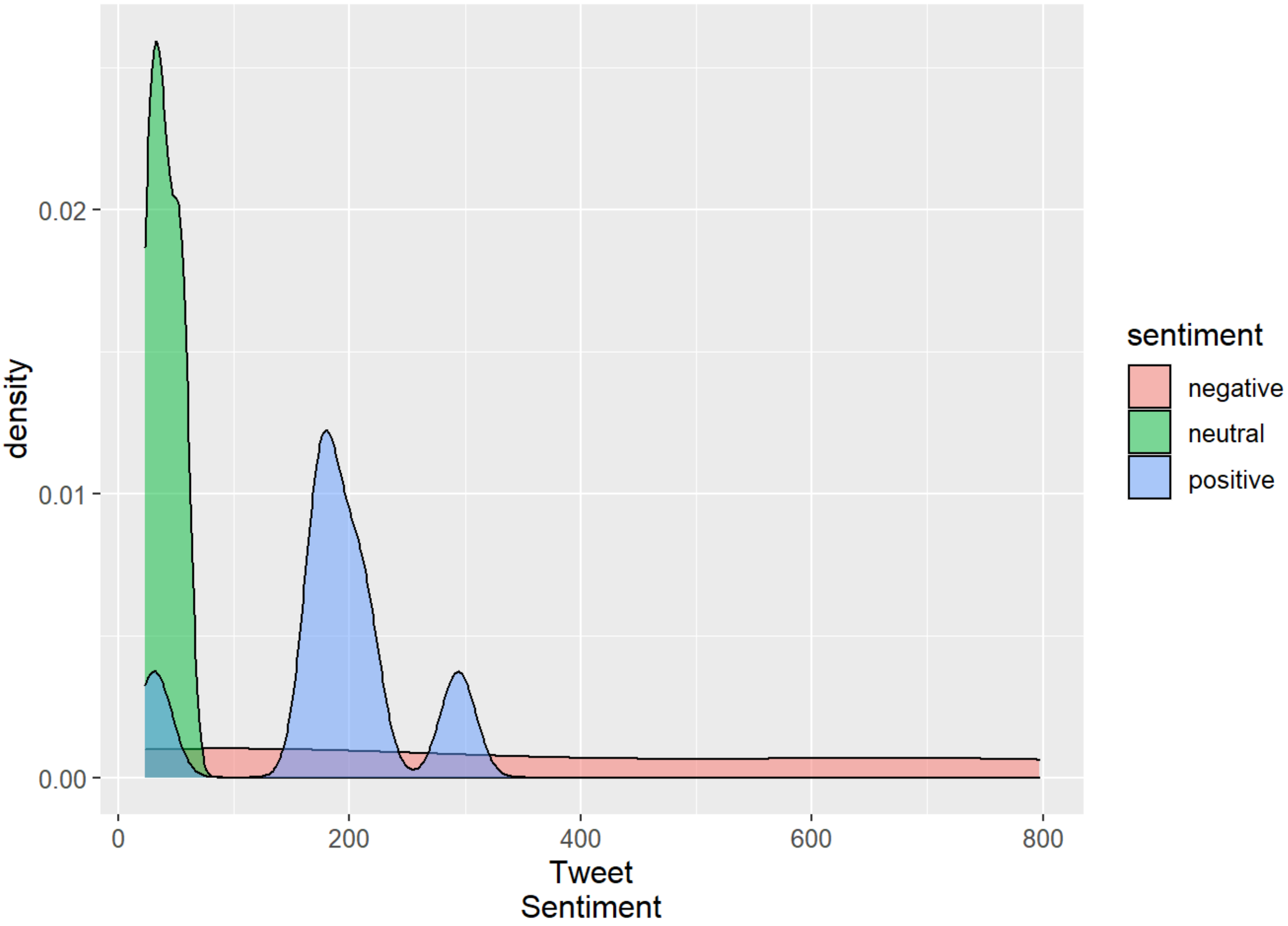


```
#only negative tweets

ggplot(df,aes(x=ave_sentiment,fill=(sentiment=="negative" )))+geom_density()
```



## Distribution of Text length for various Sentiments

```
#creating a variable with text length of each tweet
df$length = nchar(as.character(df$text))

ggplot(df, aes(x = length, fill = sentiment))+ geom_density(alpha=0.5)+s
cale_x_continuous(name   = 'Tweet\nSentiment') +labs(x='Tweet Length') +
theme(text = element_text(size=12))
```



# Findings

On Total, there are 5 negative tweets compared to 3 neutral 7 positive tweets. Canceled Flight is the leading reason for negative tweets. United Airlines has the most no. of positive tweets. US Airways has the highest number of tweets. Neutral Tweets are generally shorter compared to other tweets.

# Most frequent terms used and a Wordcloud

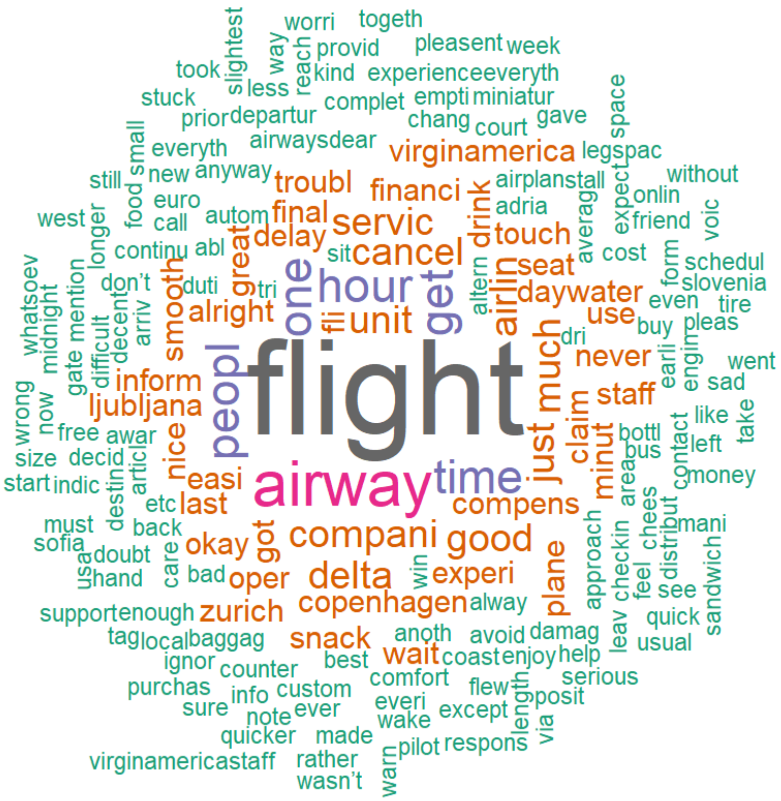
```
# Load the data as a corpus
docs <- VCorpus(VectorSource(df$text))
#To replace special characters
toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " "
, x))
docs <- tm_map(docs, toSpace, "/")
docs <- tm_map(docs, toSpace, "@")
docs <- tm_map(docs, toSpace, "\\|")
#convert to lower case
docs <- tm_map(docs,content_transformer(tolower))
# Remove numbers
docs <- tm_map(docs, removeNumbers)
# Remove english common stopwords
docs <- tm_map(docs, removeWords, stopwords("english"))
# Remove punctuations
docs <- tm_map(docs, removePunctuation)
# Eliminate extra white spaces
docs <- tm_map(docs, stripWhitespace)
# Text stemming
docs <- tm_map(docs, stemDocument)

#Text to Matrix
tdm <- TermDocumentMatrix(docs)
m <- as.matrix(tdm)
v <- sort(rowSums(m),decreasing=TRUE)
d <- data.frame(word = names(v),freq=v)
head(d, 10)
```

| | **word** | **freq** |
|---|---|---:|
| | <chr> | <dbl> |
| flight | flight | 15 |
| airway | airway | 6 |
| get | get | 4 |
| hour | hour | 4 |
| one | one | 4 |

| | word<br><chr> | freq<br><dbl> |
|---|---|---|
| peopl | peopl | 4 |
| time | time | 4 |
| airlin | airlin | 3 |
| cancel | cancel | 3 |
| compani | compani | 3 |

1-10 of 10 rows

```
#word Cloud
set.seed(1234)
wordcloud(words = d$word, freq = d$freq, min.freq = 1,
          max.words=200, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```



# Basic Model

## Document data

```
dtm = DocumentTermMatrix(docs)
dtm = removeSparseTerms(dtm,sparse = 0.99)
model_data = as.data.frame(as.matrix(dtm))
model_data$sentiment = df$sentiment
```

## Seperate data

```
model_data$sentiment <- as.factor(model_data$sentiment)
set.seed(123)
split = sample.split(model_data$sentiment,SplitRatio = 0.9)
train = subset(model_data,split = TRUE)
test = subset(model_data,split = FALSE)
```

# Randomforest

```
predict_rf = randomForest(x = train[,-159],y = train$sentiment,ntree = 1
0)
y_pred = predict(predict_rf,newdata = test[,-159])
y_pred
```

```
##        1        2        3        4        5        6        7
8
## negative positive positive positive positive positive  neutral positi
ve
##        9       10       11       12       13       14       15
## negative positive negative negative  neutral  neutral negative
## Levels: negative neutral positive
```

```
#confusion matrix
cm = table(test[,159],y_pred)
cm
```

```
##    y_pred
##     negative neutral positive
##   0        5       3        6
##   1        0       0        1
```

```
#Accuracy
sum(diag(cm))/sum(cm)
```

```
## [1] 0.3333333
```