VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Social Network Analysis of Worldwide Airline Routes & Airports.

Prepared For:

Project Component

(Social and Information Networks - CSE3021)

Submitted to Prof.

Meenakshi S.P

School of Computer Science and Engineering

Submitted By:

Shricharan S K – 19BCE2404

Mukesh Kanna R – 19BCE2385

Introduction of the project topic

The airplane and aviation is without a doubt the most influential invention of the 20th century, simply because it shrunk the world. It has connected nations that would have never been connected otherwise, and shown us a new, unseen and spectacular perspective of our earth. Commercial aviation has been proven to have a direct impact on our nation's economy, creating more than 10 million well paying Indian jobs and driving almost 4 percent of the nation's annual gross domestic product and nearly $1.1 trillion in annual economic activity.

Analysis of these airline networks would hence lead to valuable information about movement of people and goods across the globe, and the relationships of countries among themselves and its effect on trade, economy and other factors.

This data has further implications in vicarious industries. It can help us build better systems for transportation and logistics of people and goods. Social network analysis in general studies the behavior of the individual at the micro level, the pattern of relationships (network structure) at the macro level, and the interactions between the two.

In this particular case the analysis would involve the interaction of airline networks globally and its implications on a particular country and airline, the nodes in this instance are airports, or countries through which flights depart or arrive, and the edges attempt to trace the path that the airlines follow.

Many previous studies on air transport network examined several specific airports or regions and mainly utilized the internal indicators of airports. Conversely, this project conducts a comprehensive analysis covering 173 countries by using air route, which is an external indicator of airports.

This project attempts to present the general characteristics of major countries and regions from the perspective of Social and Information Network Analysis and tries to compare the individual networks of countries like United States, China, North Korea and some other Asian trade giants, which have the greatest influence on international air logistics within the scope of the entire network analysis. This study can also aid in better understanding the nuances of air transport networks and logistics connectivity in inter-city and inter-country transport.

In particular, major airports in the world are often seen to be characterized by the capability to connect directly with hundreds of airports without intermediate routes. The air transportation system can be represented as a network, in which nodes denote airports and an edge will be created if a direct flight exists between two airports. From this point of view, this project not only aims to analyze the characteristics of airports, airlines, countries and the correlation between them based on the air routes that are connected to major airports around the world, but also helps in inculcating and developing the general understanding of an air transport network.

The analysis of logistics connectivity in inter-city or inter
country relations has implications for the diversified application of research methodology and international logistics research, which will be further discussed in the upcoming section.

Project Abstract

In today’s competitive world, airlines are playing essential roles as major transportation for human and commercial goods. Be it via the trade or travel sector, none of us is alien to the enormous positive impact the air travel industry has bought in our lives.

Almost every single one of us gets and reaps the benefits of this advanced transportation technology. Thus, having general understanding of global airlines/airports data is the need of the hour.

The number of airports a country has is often a direct measure of the development and flourishment of its travel and trade sector, both of which are invaluable pillars of a country’s economic progress and development. Hence by analyzing the global and domestic air traffic of a particular country, one can often get a fair idea of its financial and economic conditions.

In this project, we aim to make visualizations using social network analysis techniques to analyze the worldwide airport and flight network to help our fellow peers understand and extract necessary information out of those hard to read and interpret datasets, and Hopefully the analysis in the form of this project inspires others to have and develop deeper interests in these type of datasets.

As all of us are well versed with the fact that social networks are seen to depict homophily that is, the tendency of individuals to associate and bond with similar others. We will also be taking up a case study of certain countries, particularly some communist and anti-social countries like China, North Korea, etc. and attempt to draw parallels between their ideologies and airline networks, which can give us a further insight into their trade routes and the countries with which they have good and friendly international relations.

We will be using information and data from multiple data sets and data sources, to obtain necessary information about the flights and airlines in various countries, primarily obtaining our data via web scraping techniques from online sources like the OurAirports website.

OurAirports is a free site where visitors can explore the world’s airports, read other people’s comments, and leave their own. The site is dedicated to both passengers and pilots. Users can find any airports around the world. The site started in 2007 to create a good source of global aviation data available to anyone and provides us with just the right amount of information to get started on the project.

As we progress further, we plan to be open to changes and are willing to adapt according to the needs of the project. Making the necessary changes to the data and the analysis techniques used, so that the information can be conveyed in an intuitive and easy to understand way for even naïve users.

By the means of this project, we not only aim to implement the topics and techniques taught in the classroom practically, but also make an attempt to help understand the information these networks attempt to convey at not just a superficial, but a deeper level.

The questions that our project will attempt to find answers to are:

- Which country has the most airports?
- Which country has the most airlines?
- How is a county’s development related to the number of airports a country has?
- What route is the busiest international airline route?
- Are there any flights that arrive in anti-social countries?
- Does quantity always mean quality? – Sentiment analysis of one of the biggest flight networks of the world.

Literature Review

1. Programming languages Used

The programming languages used for this project are R and Python. R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing.

The R language is widely used among statisticians and data miners for developing statistical software and data analysis. R and its libraries implement a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others. R is easily extensible through functions and extensions.[1]

R is a language and environment for statistical computing and graphics. R is also extremely flexible and easy to use when it comes to creating visualizations[3]. One of its capabilities is to produce good quality plots with minimum codes.

1. Basic graphs in R can be created quite easily. The **plot** command is the command to note.
2. It takes in many parameters from x axis data, y axis data, x axis labels, y axis labels, color and title. To create line graphs, simply use the parameter, type=l.
3. If you want a boxplot, you can use the word boxplot, and for barplot use the barplot function, etc. This way the R programming language not only offers a wide range of options, but also calls for its effective and easy implementation through the ggplot library.

We have made use of python[2] in for sentiment analysis of twitter flight data[6] later in the project to analyze the overall satisfaction of the customers with the airlines they tend to fly with.

2. Social Network Representation in Real life

Social networks appear to enrich our social life, which raises the question whether they remove cognitive constraints on human communication and improve human social capabilities. In this project, we analyze the worldwide flight network and attempt to shed light on the various structural and functional properties of the same. Compared with online social networks, our results confirm some similar features. However, the worldwide flight network also shows its own specialties, such as hierarchical structure and degree disassortativity, which all mark a deviation from other real-life social networks.[4] The wide spread of the airline network forms a broader perspective, and the two-way link relationships make it easy to spread information, but unlike the online social network that does not make too much difference in the creation of strong interpersonal relationships, flight network between two countries is often an indication of their relationship with each other.[10] Finally, we describe the mechanisms for the formation of these characteristics and discuss the implications of these structural properties for the airline social networks.

3. Social Network Analysis

Networks are defined by nodes and the edges between them. In the case of social networks, the nodes represent individual people and the edges the relationships between them. Quantifying the relationships between people is hence fundamental in characterizing social networks[10]. To estimate these relationships, most studies have tended to follow the lead of the pioneering study by Moreno, who used questionnaires to investigate friendship choices among selected children. In these studies, researchers simply ask respondents to identify their friends and use these data to define the edges of the social network. This approach to define the edges is time consuming, subjective and depends on the nature of the questions – or name generators – that are asked [11]. That is, respondents will describe networks of varied size and the characteristics of relationships will vary considerably depending on the questions asked.

Our study focusses on worldwide flight networks, whose nodes and edges are a part of a more concrete basis of formation and do not depend on characteristics

Iike nature of questions asked, which are subject to change depending on the type of questions asked and to whom these questions have been posed.

4. Deep learning ModeI

The method used in the above probIem is inspired from the officiaI UMIFiT paper[5]. To understand what UIMFiT is, one must first familiarize oneseIf with the basics of Ianguage ModeIing[8]. A Ianguage modeI is a probability distribution over sequences of words. The Ianguage modeIs can then be used as base modeIs for various naturaI Ianguage processing tasks including text classification, summarization, analysis and more.

UniversaI Ianguage ModeI Fine-Tuning (ULMFIT) is a transfer Iearning technique which can heIp in various NLP tasks. It has been state-of-the-art NLP technique for a Iong time, before it was dethroned by techniques Iike BERT and XINet. We have used the fast.ai library[7] to implement the same.

According to the officiaI UIMFiT paper, the advantage it boasts of over its counterparts is the fact that deep Iearning techniques, that often require very Iarge datasets, tend to overfit when the dataset given to us is relatively smaIler and domain specific, Iike in this case. UIMFiT heIps address this very predicament by first breaking our data into batches known as a databunch, we then fit and train our deep Iearning modeI for a few cycIes by running 1 epoch at a time, then unfreezing some Iayers and running subsequent epochs to fine tune the same, instead of the conventionaI approach of modeI training by running aII epochs at once [9]. The modeI is saved after every epoch, whiIe adjusting the hyper-parameters after every epoch to heIp our modeI Iearn various edge cases in our test set better.

Proposed Method

Analysis of flight networks would lead to valuable information about movement of people and goods across the globe, and the relationships of countries among themselves and its effect on trade, economy and other factors.

This data has further implications in vicarious industries. It can help us build better systems for transportation and logistics of people and goods. Social network analysis in general studies the behavior of the individual at the micro level, the pattern of relationships (network structure) at the macro level, and the interactions between the two.

In this particular case the analysis would involve the interaction of airline networks globally and its implications on a particular country and airline, the nodes in this instance are airports, or countries through which flights depart or arrive. Many previous studies on air transport network examined several specific airports or regions and mainly utilized the internal indicators of airports. Conversely, this study conducts a comprehensive analysis covering 173 countries by using air route, which is an external indicator of airports.

This study presents the general characteristics of major countries and regions from the perspective of SNA and compared the individual networks of the United States and China, which have the greatest influence on international air logistics within the scope of the entire network analysis. This study can also aid in the understanding of air transport networks and logistics connectivity in inter-city and inter-country transport.

Algo and approach

The first step we followed post data preprocessing and wrangling was plotting all the airports on the world map so as to form the nodes of the social network that we would be working with. Next, we iterated through our data and connected the nodes via a directed edge if they had a direct flight between them, this formed our complete visualization of the social network that we would be performing our analysis on containing both the nodes and the global airline routes.

Next we plotted the visualizations of the airport altitudes and plotted the airports located at an altitude of over 5000 feet.

The next part of our study focused on analyzing the density and concentration of airports in all the countries and then ranking them from the highest to the least number of airports and airlines.
Here we found out that the United States of America has both the highest number of airports and the highest number of airlines amidst all the other countries. After social network analysis, we conducted a case study between the countries if the most and least dense flight network, namely United States of America and North Korea.

The code and the visualizations obtained will be enclosed in the next section.

Does quantity always equate with quality?

We are currently working on a deep learning model to perform sentiment analysis on the crowdflower United States Airline Tweets dataset to get a general idea of how satisfied the flyers were with the airlines in their country.
We would be using the UIMFiT method to process the data and then we would be training a multi-layer deep neural network (RNN) to perform sentiment analysis on the twitter dataset containing over 50k tweets from different users.

This would give us an idea as to how satisfied the people actually are with the services provided to them despite the United States having the most robust and dense flight network in the world.

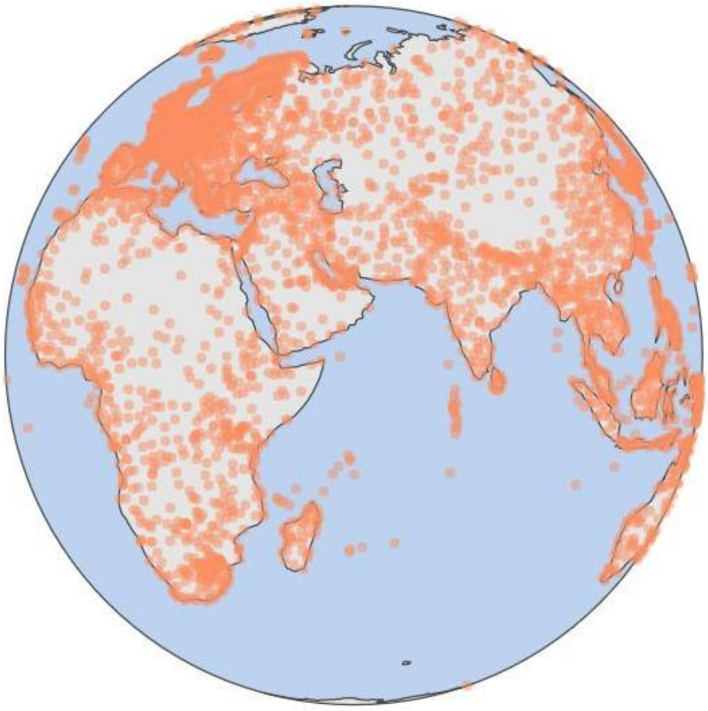
We would then continue our analysis and take up the case study of a country with a sparse international airline network, like North Korea, analyze its structure and try to infer conclusions regarding the trade routes and international relations using SNA techniques.

Sentiment analysis using Deep learning approach

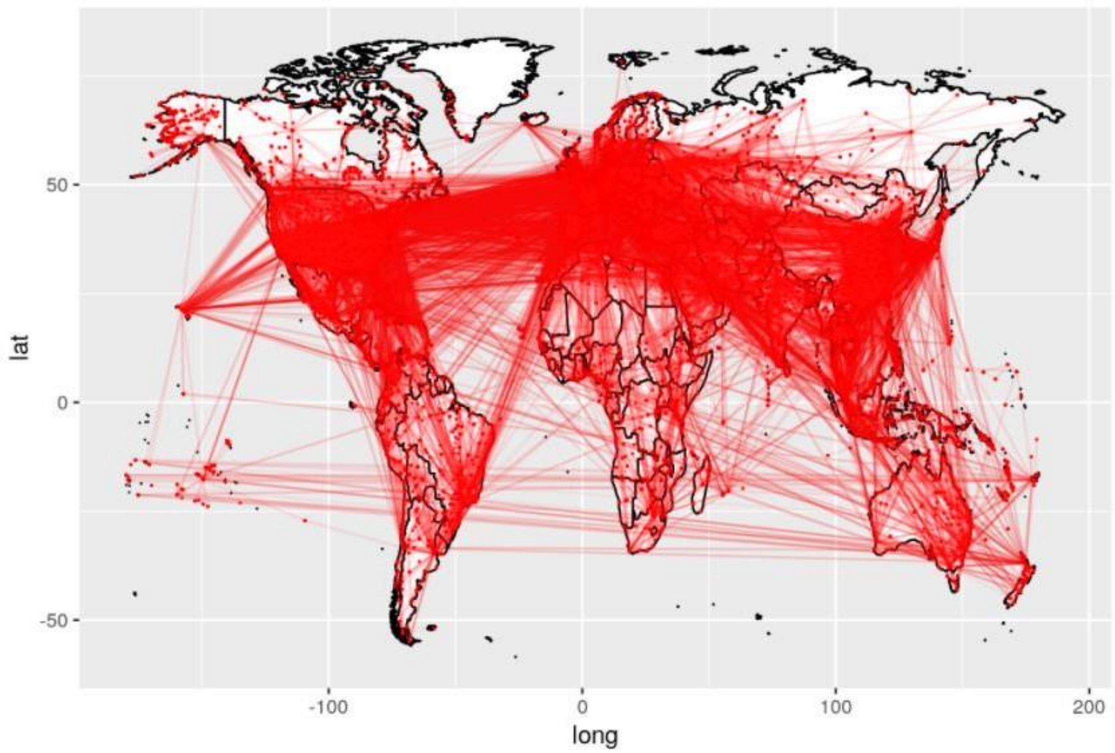
I. The approach and algorithm adopted is taken from the UIMFiT research paper which makes use of data bunches to train two language models, the first one tries to predict the word that might follow the word at hand and the second language model predicts the overall sentiment. To train these language models we **train by gradually unfreezing layers of our neural network and training our model one epoch at a time instead of performing all the epochs at once, in accordance to the suggestions in the UIMFiT paper [5]. This method of training not only ensures better accuracy, but also helps our model learn the edge cases to perfection.**

Results

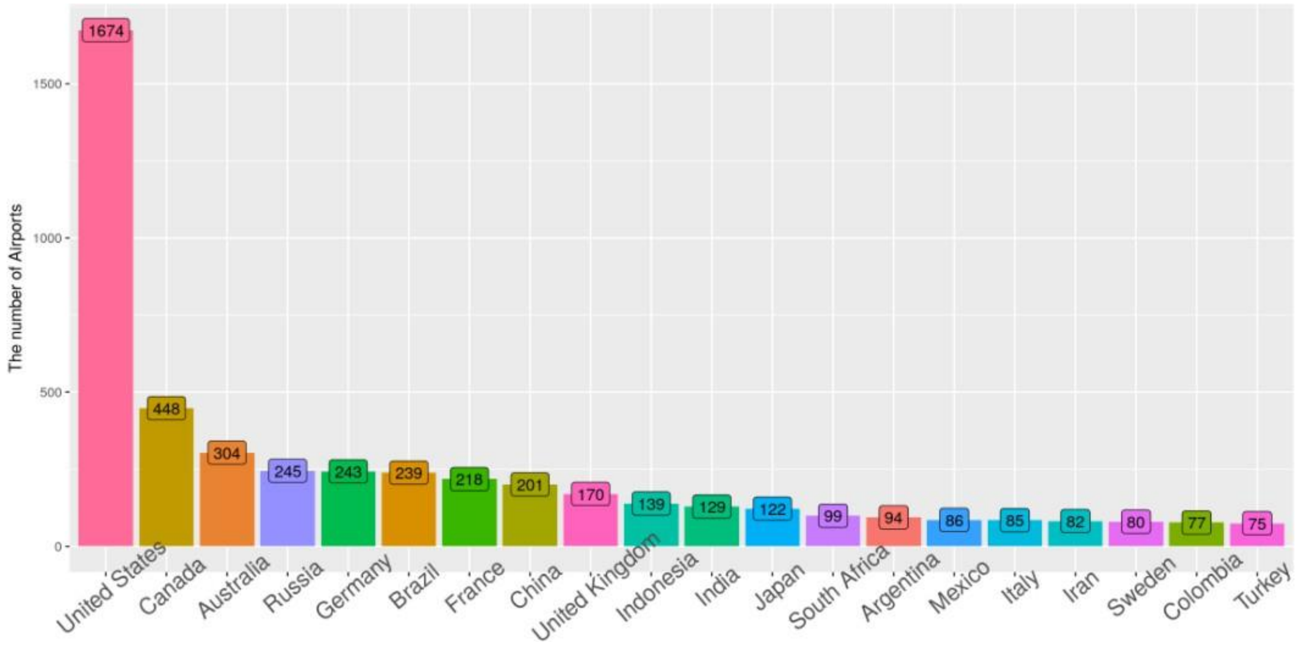
Visualizations plotted and associated code snippets.



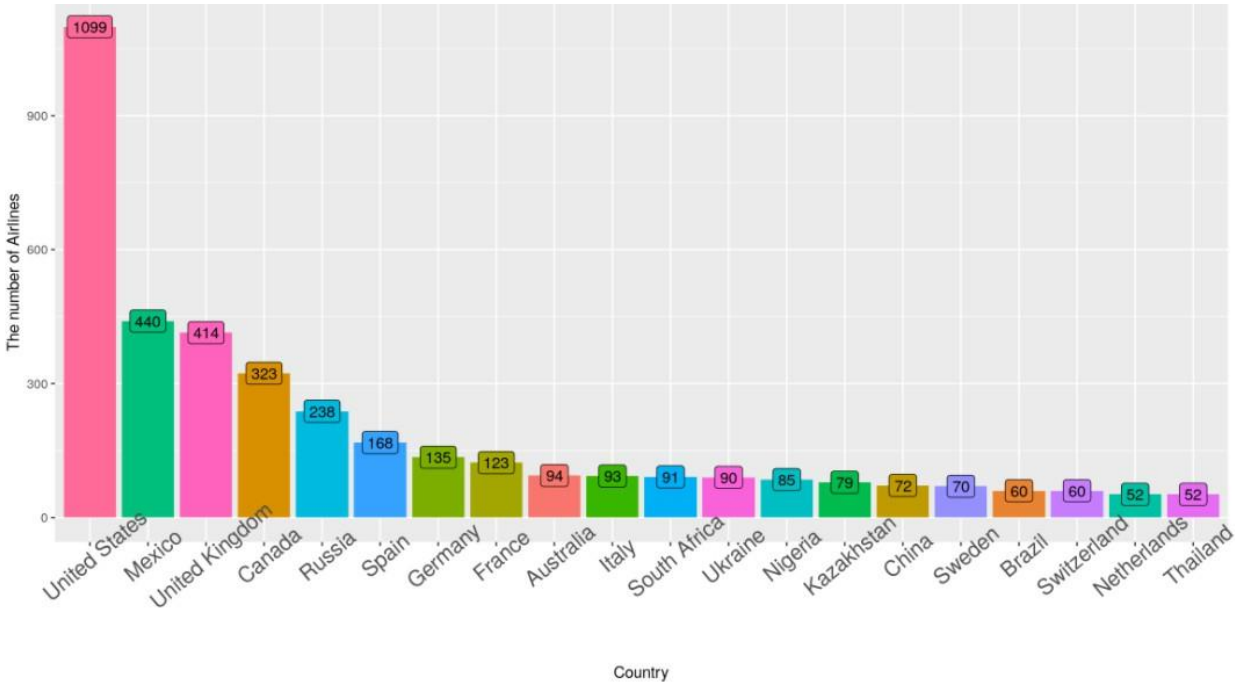
The latitude and longitude of the airport are taken as coordinates and the airport is visualized as a node on the world map. As we can see here.



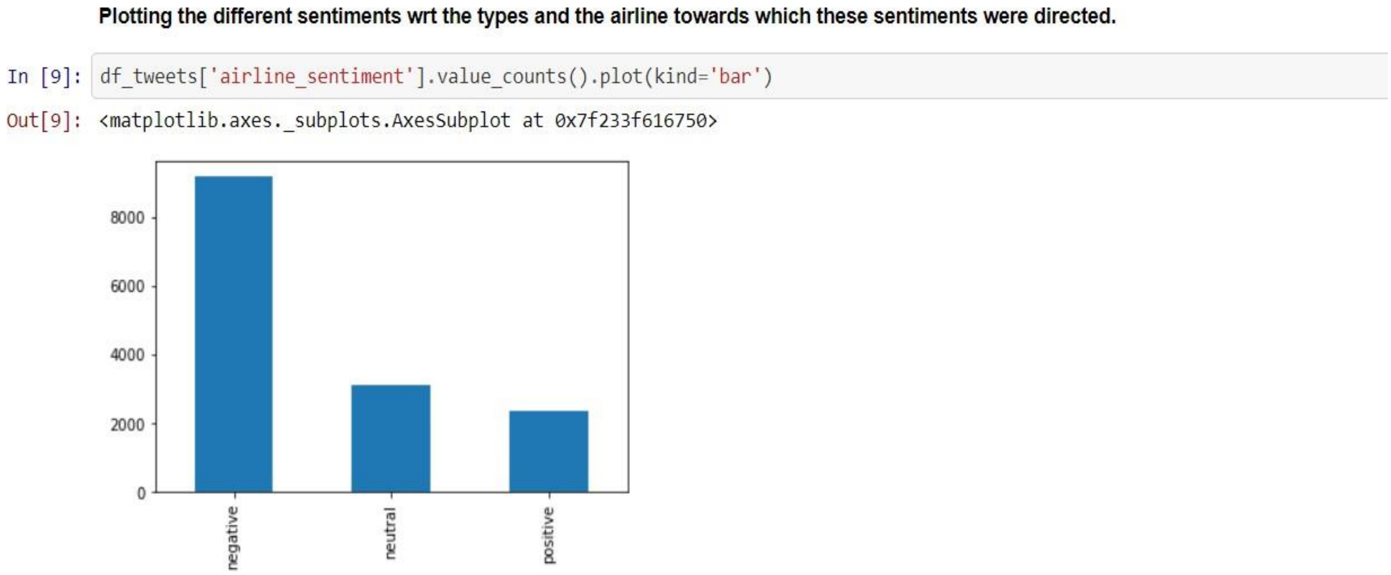
The global airports are visualized as nodes on the map and adding edges where a flight exists, forms the overall flight social network.



Number of airports in different countries. – We see that USA has the most number of airports all over the world.



Number of airlines in different countries. – We see that USA has the most number of airlines as well, alongwith other leaders in the number of airports.

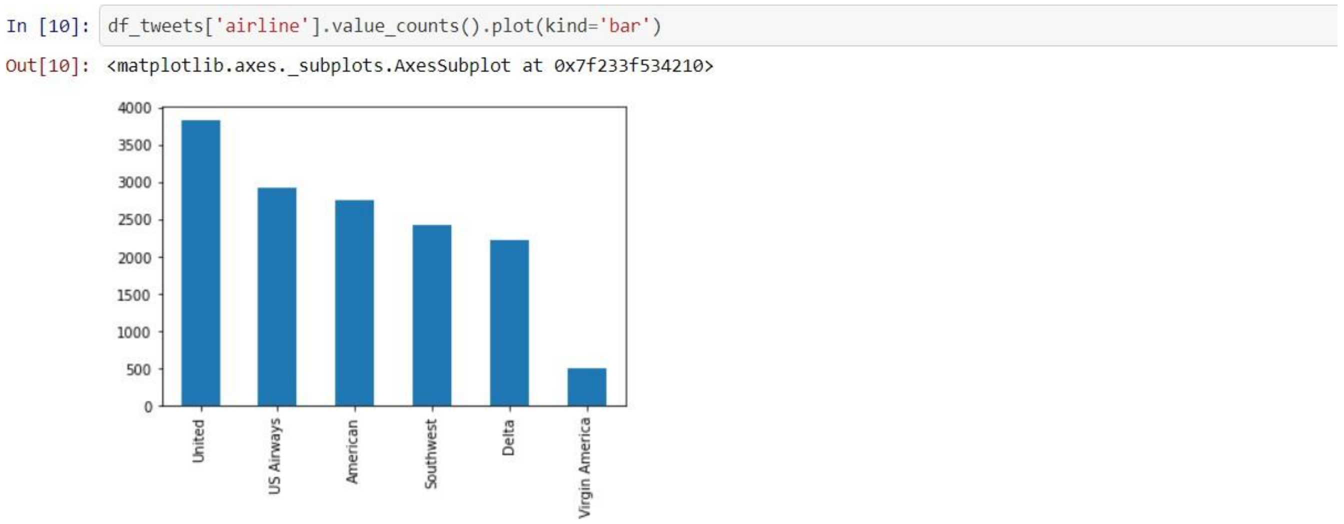


Plotting the different sentiments wrt the types and the airline towards which these sentiments were directed.

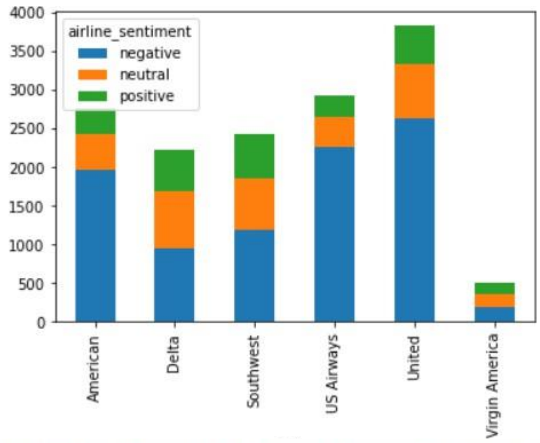
```
In [9]: df_tweets['airline_sentiment'].value_counts().plot(kind='bar')
```

```
Out[9]: <matplotlib.axes._subplots.AxesSubplot at 0x7f233f616750>
```

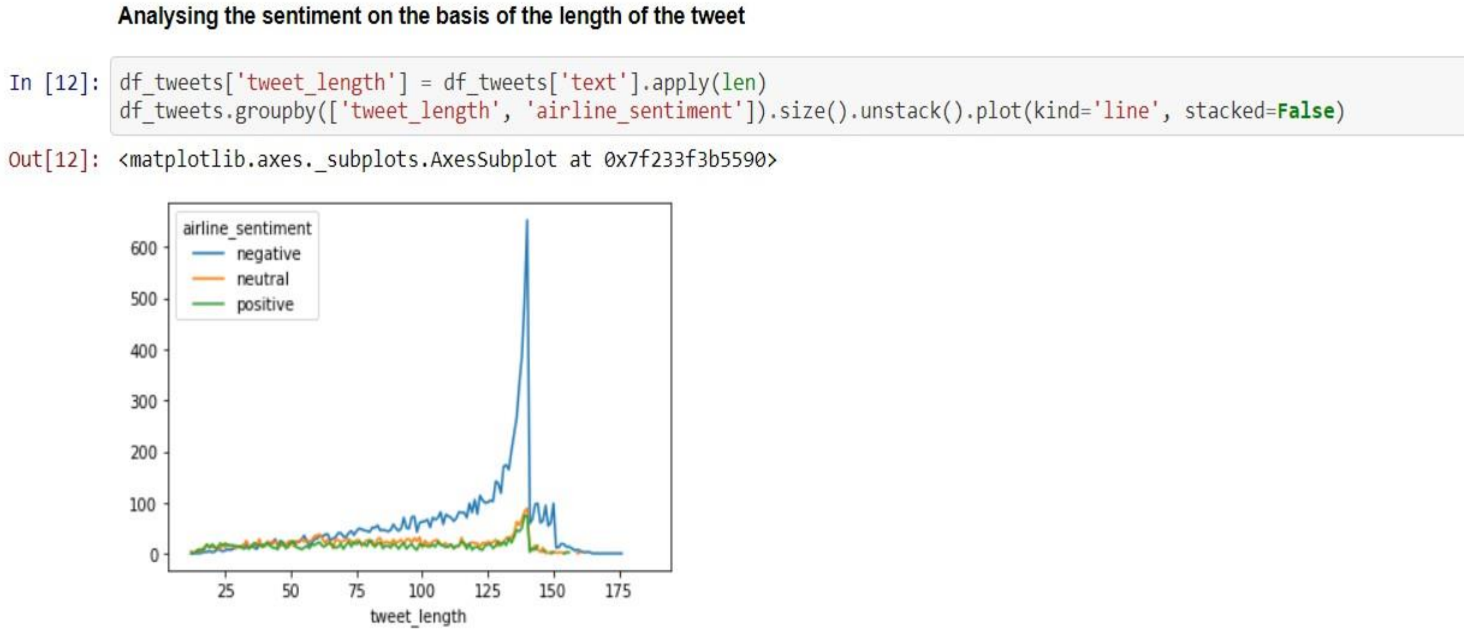
Overall sentiments in our United States flight dataset (independent of airline) – Post sentiment analysis we see that despite the biggest and the most robust flight network, the overall sentiment of the fliers is still negative.



```
In [10]: df_tweets['airline'].value_counts().plot(kind='bar')

Out[10]: <matplotlib.axes._subplots.AxesSubplot at 0x7f233f534210>
```

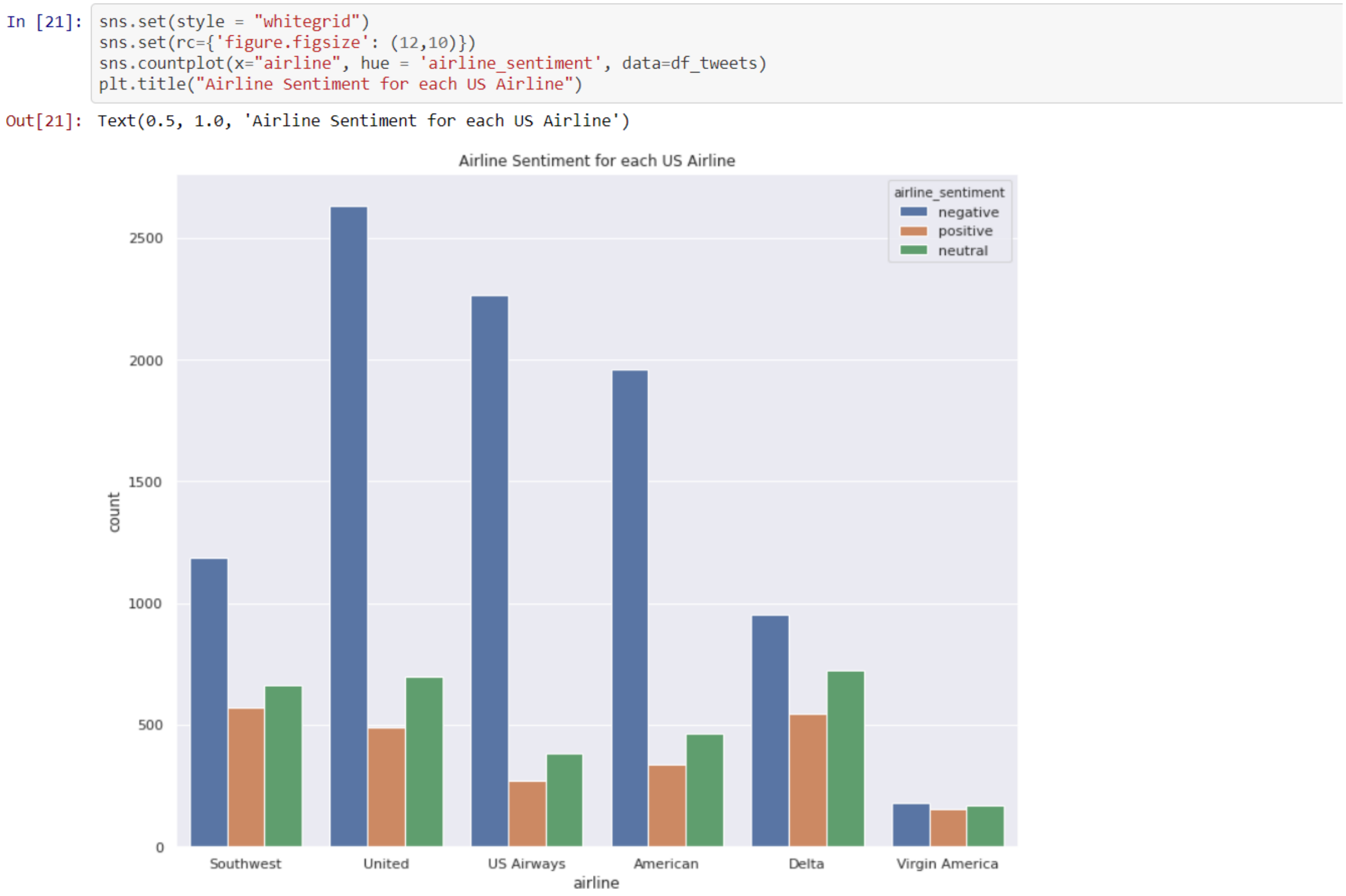


Airline vs the kinds of sentiment associated with them.- Some of the popular US airlines and the general sentiment the fliers have towards them.



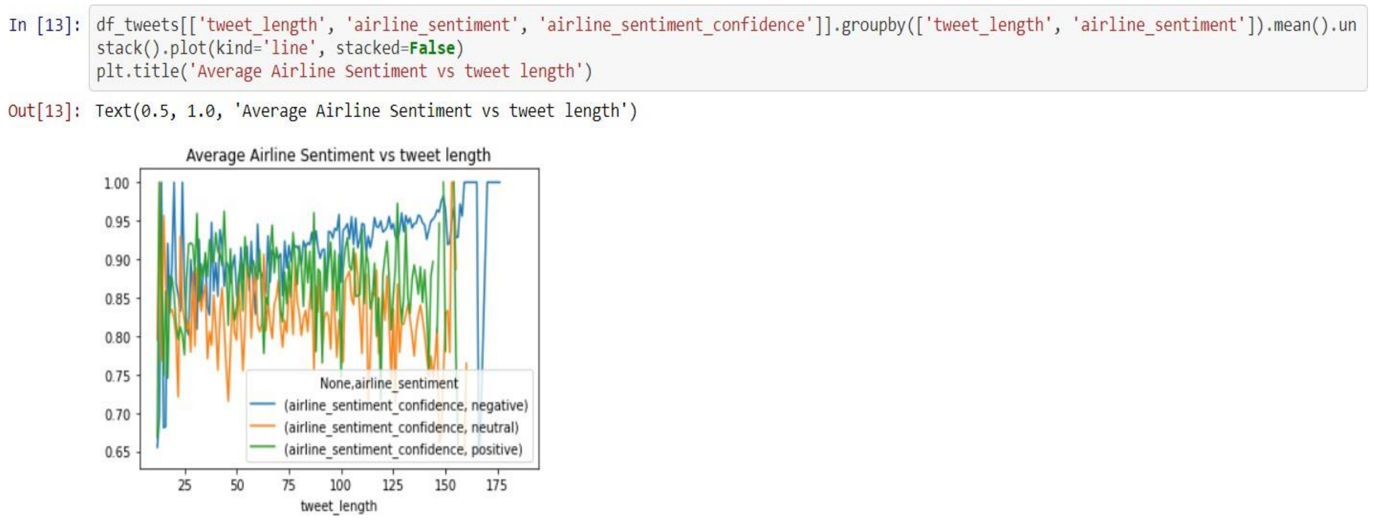
Analysing the sentiment on the basis of the length of the tweet

```
In [12]: df_tweets['tweet_length'] = df_tweets['text'].apply(len)
         df_tweets.groupby(['tweet_length', 'airline_sentiment']).size().unstack().plot(kind='line', stacked=False)

Out[12]: <matplotlib.axes._subplots.AxesSubplot at 0x7f233f3b5590>
```



```
In [21]: sns.set(style = "whitegrid")
         sns.set(rc={'figure.figsize': (12,10)})
         sns.countplot(x="airline", hue = 'airline_sentiment', data=df_tweets)
         plt.title("Airline Sentiment for each US Airline")

Out[21]: Text(0.5, 1.0, 'Airline Sentiment for each US Airline')
```

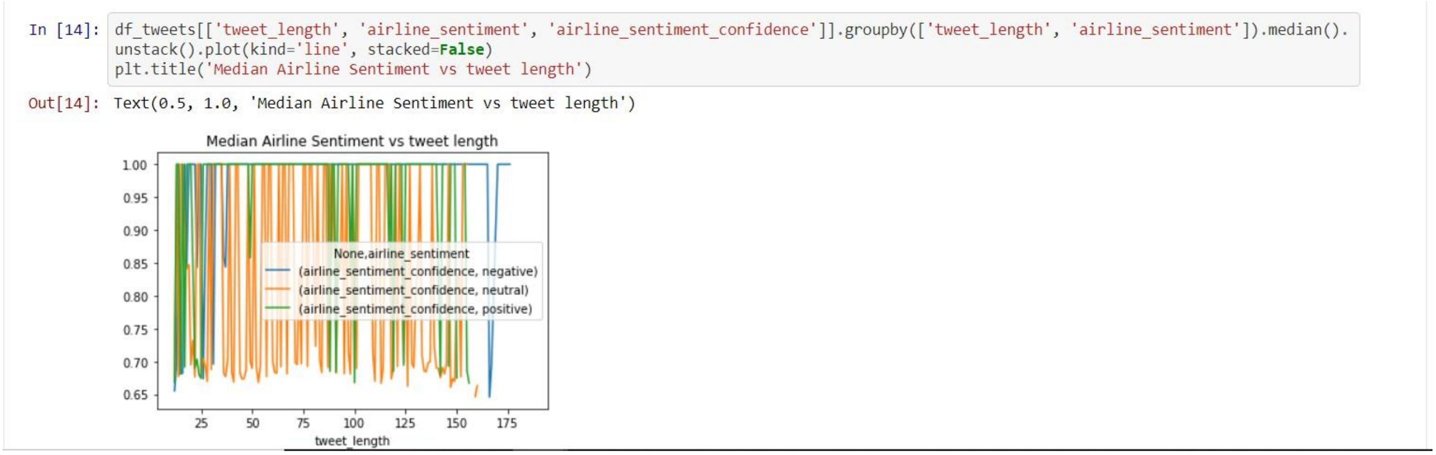
Sentiment vs Airline name



```
In [13]: df_tweets[['tweet_length', 'airline_sentiment', 'airline_sentiment_confidence']].groupby(['tweet_length', 'airline_sentiment']).mean().un
         stack().plot(kind='line', stacked=False)
         plt.title('Average Airline Sentiment vs tweet length')

Out[13]: Text(0.5, 1.0, 'Average Airline Sentiment vs tweet length')
```

Average Airline Sentiment vs tweet length
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
25
50
75
100
125
150
175
tweet_length
None,airline_sentiment
(airline_sentiment_confidence, negative)
(airline_sentiment_confidence, neutral)
(airline_sentiment_confidence, positive)

From cell no. [12] and [13] it is safe to conclude that, the longer the tweet, the more likely it is to be negative. And hence, length could be a parameter for our model.

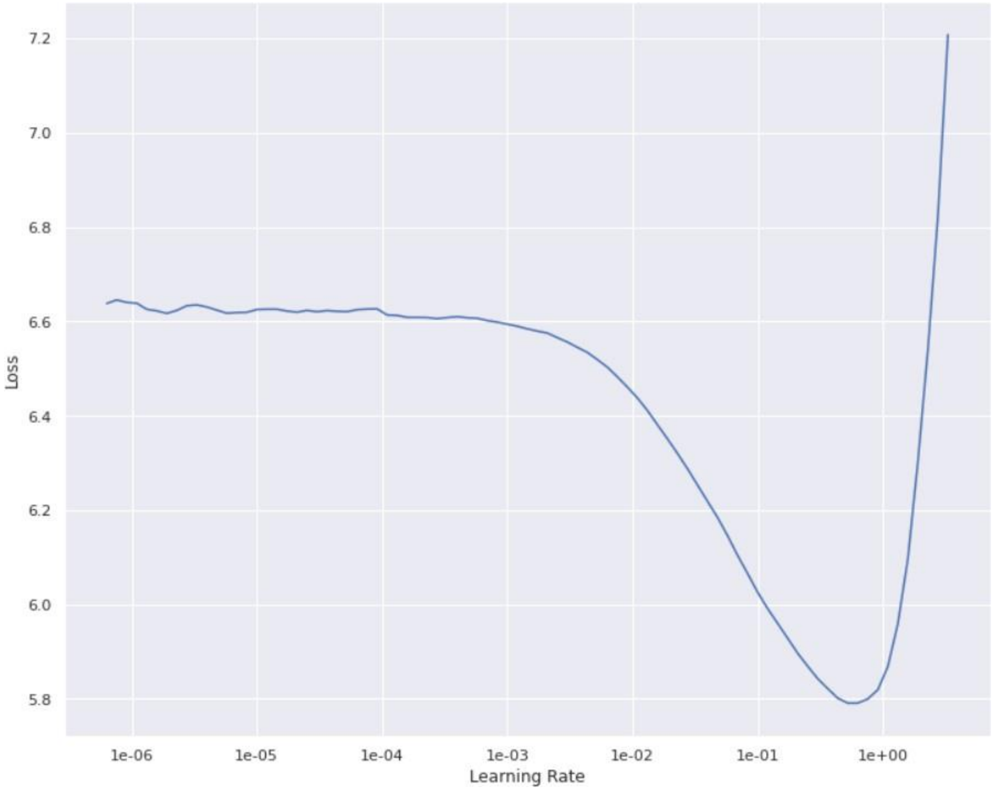


```
In [14]: df_tweets[['tweet_length', 'airline_sentiment', 'airline_sentiment_confidence']].groupby(['tweet_length', 'airline_sentiment']).median().
         unstack().plot(kind='line', stacked=False)
         plt.title('Median Airline Sentiment vs tweet length')

Out[14]: Text(0.5, 1.0, 'Median Airline Sentiment vs tweet length')
```

Median Airline Sentiment vs tweet length
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
25
50
75
100
125
150
175
tweet_length
None,airline_sentiment
(airline_sentiment_confidence, negative)
(airline_sentiment_confidence, neutral)
(airline_sentiment_confidence, positive)

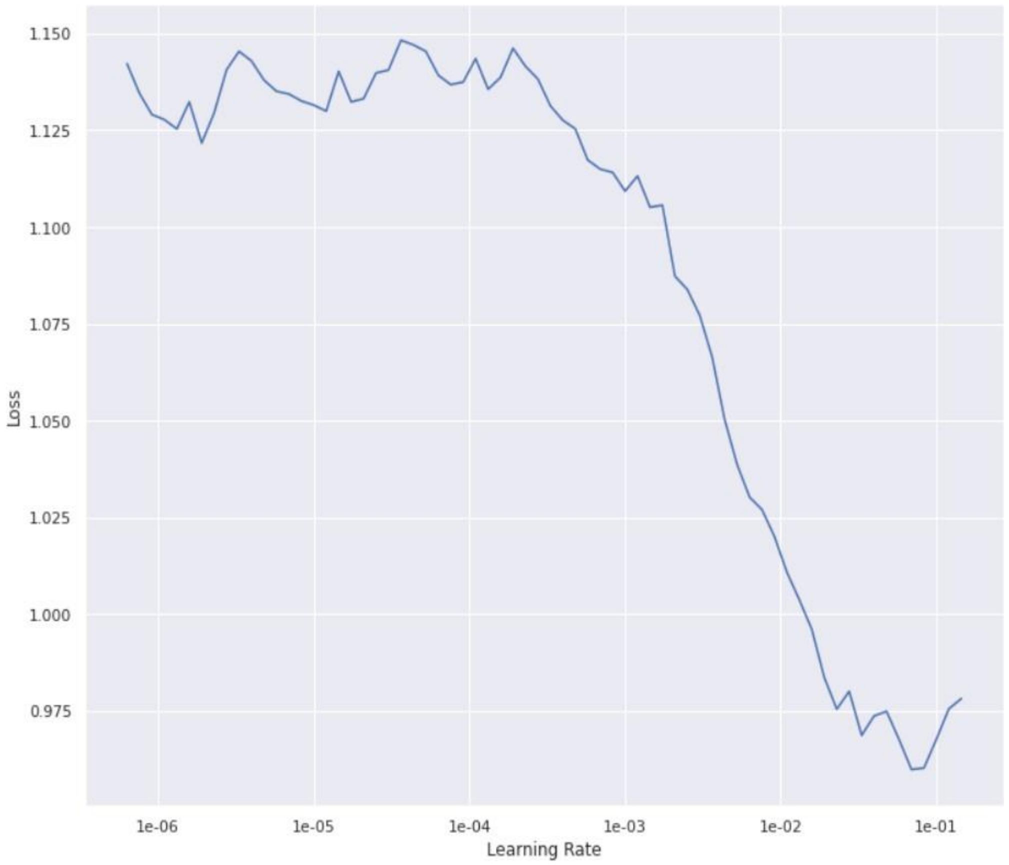
We see that we cannot form any concrete relation between sentiment confidence and length. – Sentiment confidence is hence, not correlated to the length of the tweet.

Initializing the language model, setting rates and plotting the loss function vs learning rate curve for first language model (to fine tune alpha).



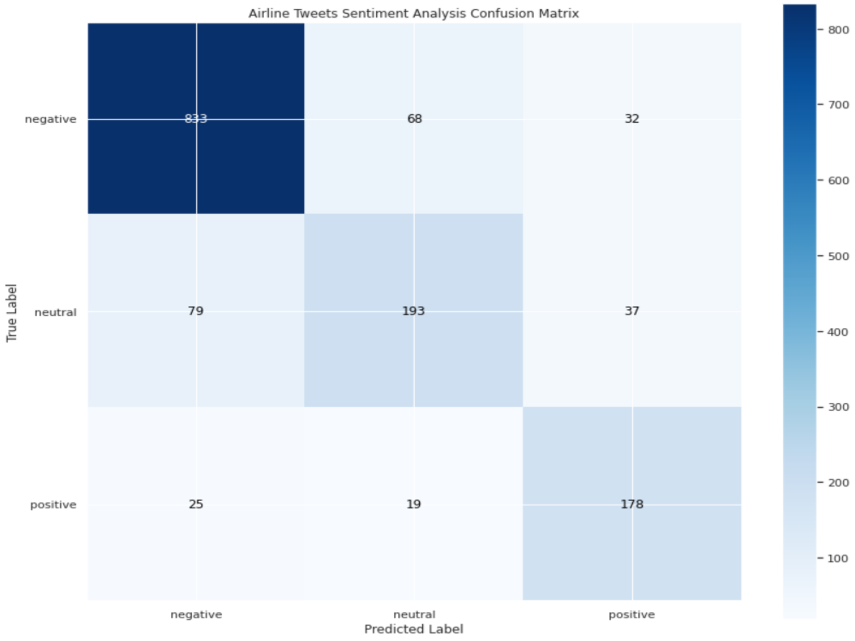
loss function vs learning rate for first language model (to fine tune alpha)

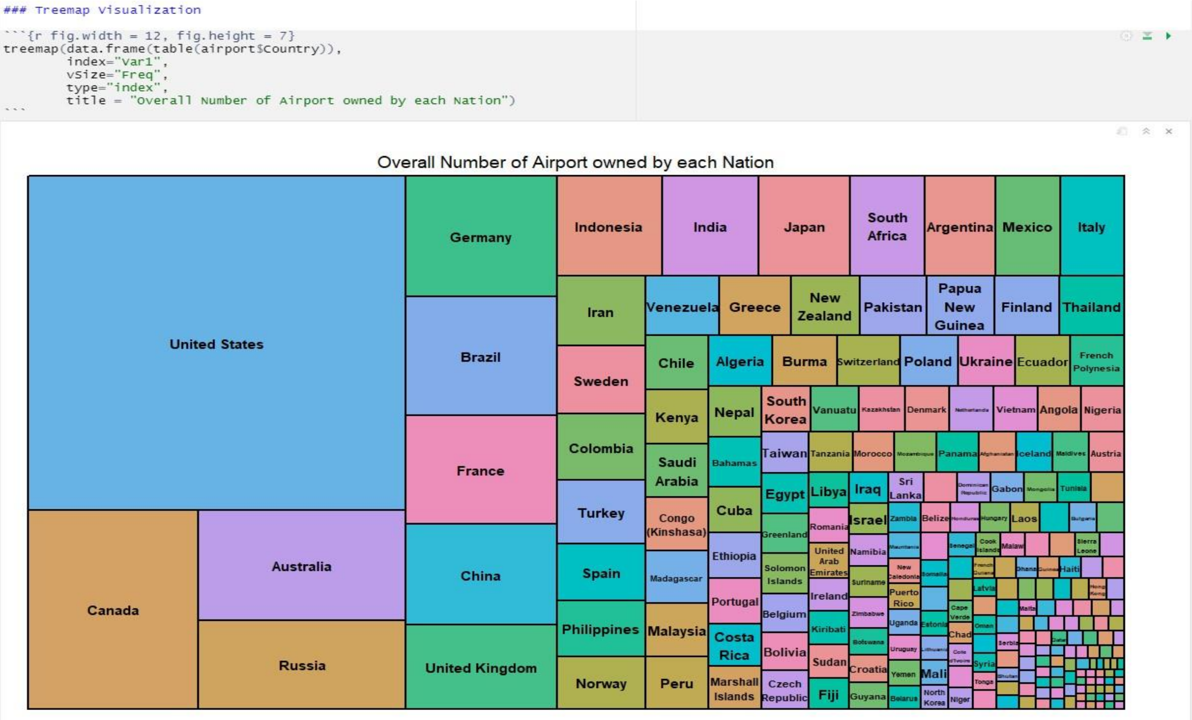
Model fine tuning.



Loss function vs learning rate curve for our final fine-tuned predictive model that detects the sentiment of the text given.

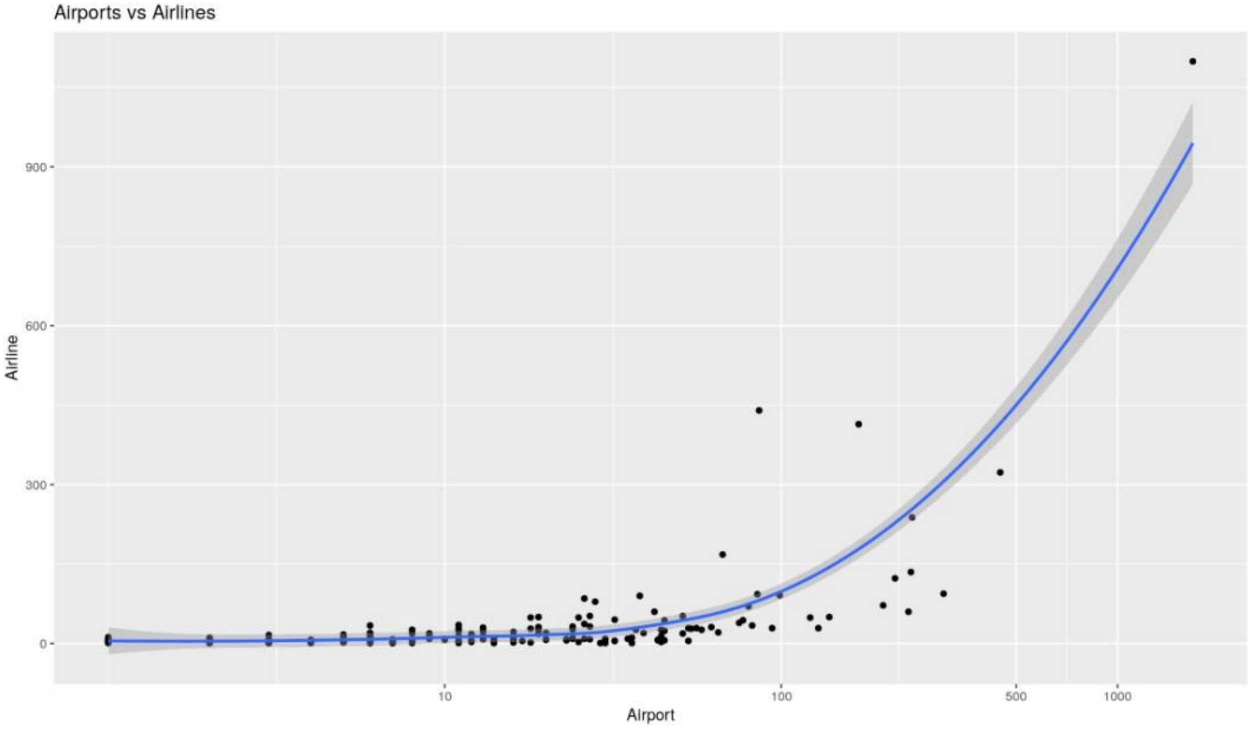
Confusion matrix obtained after evaluating on the test set – shows us that our model performs reasonably well and can generalize to unseen data having a high accuracy for true positives and true negatives.





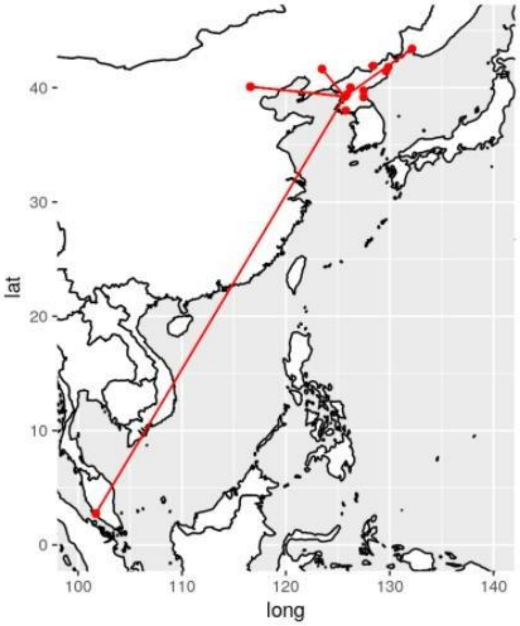
~~~
### Treemap Visualization

```{r fig.width = 12, fig.height = 7}
treemap(data.frame(table(airport$Country)),
        index="Var1",
        vSize="Freq",
        type="index",
        title = "Overall Number of Airport owned by each Nation")
```
~~~

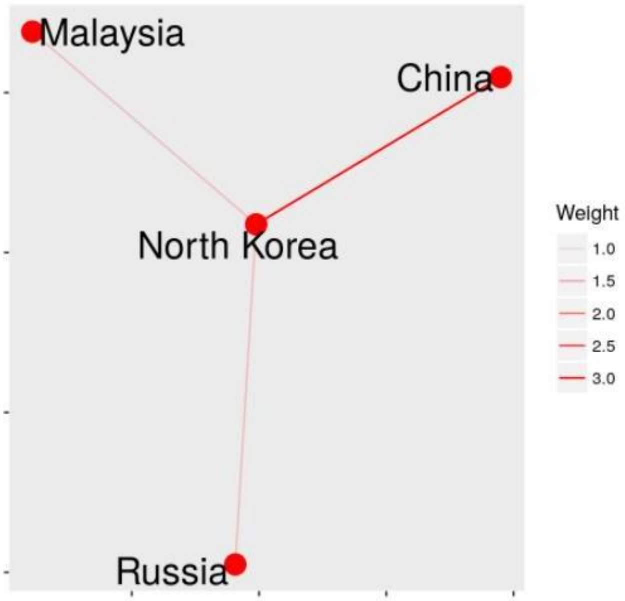
A tree-map visualization showing the number of airports owned by each country with proportion to the area of a rectangle. We see that the developed and economically progressive countries occupy the majority of this area and hold almost 60% of the world's total airports.

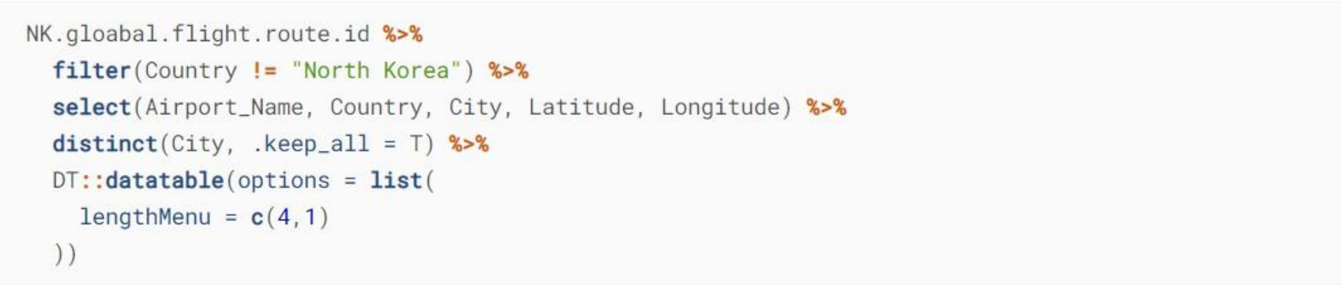


Graph showing a +ve correlation between the number of airports and airlines. As we saw from the above visualizations, countries with more number of airports have more number of airlines.

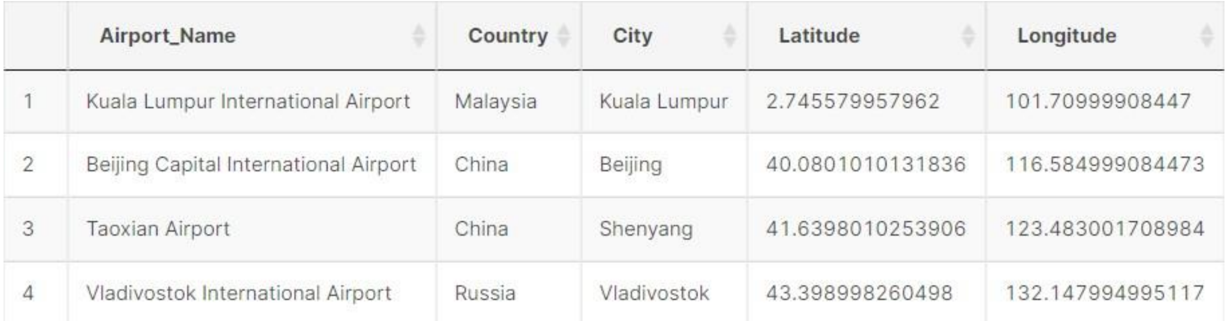
Analyzing the flight network of antisocial countries like North Korea







```
NK.gloabal.flight.route.id %>%
  filter(Country != "North Korea") %>%
  select(Airport_Name, Country, City, Latitude, Longitude) %>%
  distinct(City, .keep_all = T) %>%
  DT::datatable(options = list(
    lengthMenu = c(4,1)
  ))
```



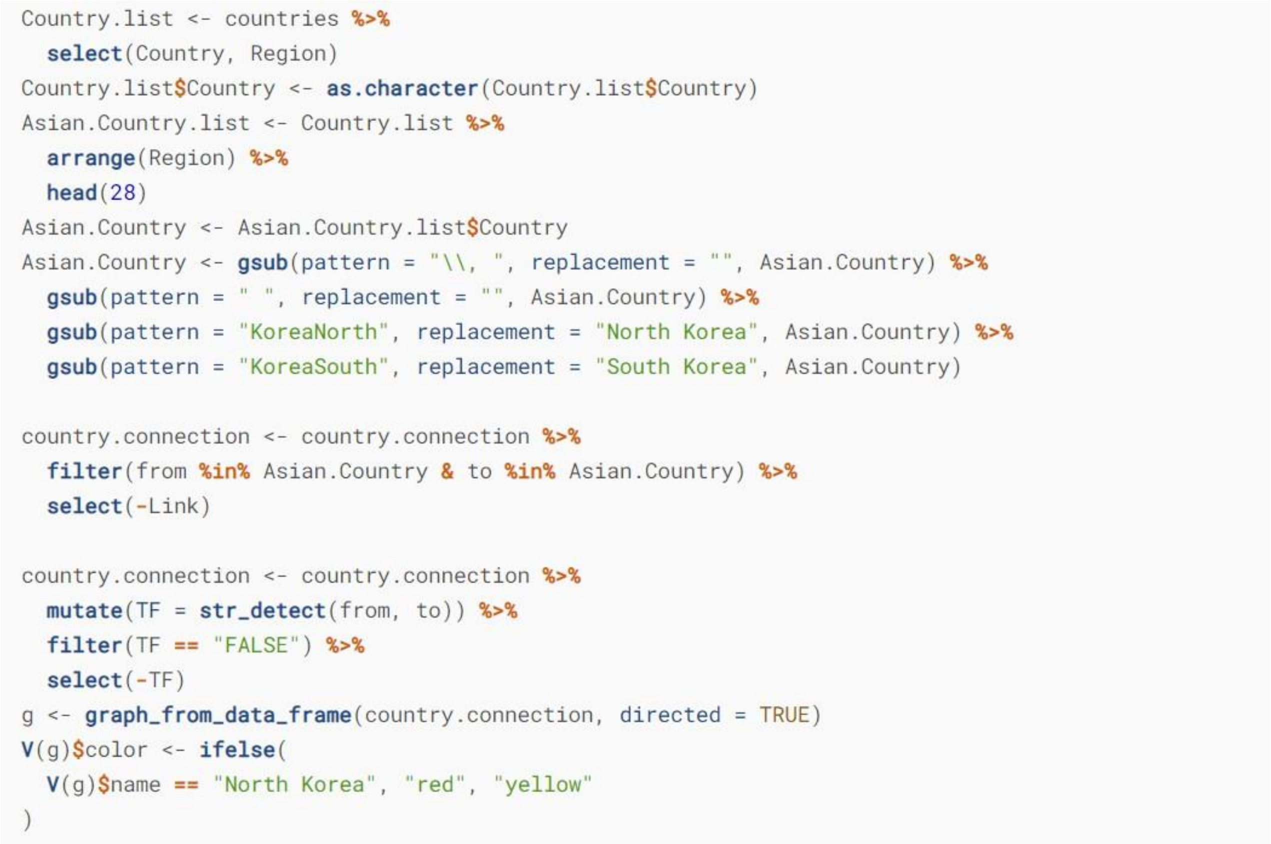
	Airport_Name	Country	City	Latitude	Longitude
1	Kuala Lumpur International Airport	Malaysia	Kuala Lumpur	2.745579957962	101.70999908447
2	Beijing Capital International Airport	China	Beijing	40.0801010131836	116.584999084473
3	Taoxian Airport	China	Shenyang	41.6398010253906	123.483001708984
4	Vladivostok International Airport	Russia	Vladivostok	43.398998260498	132.147994995117

Info of airports that have flight access to North Korea. We see that only 4 airports in 3 countries have airline access to North Korea (which is an antisocial country). These countries themselves have high centrality and act as bridge nodes to keep North Korea connected to the rest of the world.

We see that most of these countries are Asian and predominantly those Asian countries that have a large flight network themselves that keeps North Korea connected to the rest of the subcontinent and these nodes act as bridge nodes. Let us now analyze the network of these Asian countries and see which are the most prominent bridge nodes that connect North Korea to the rest of the Asian subcontinent and helps maintain its trade routes without having direct flights to many countries.

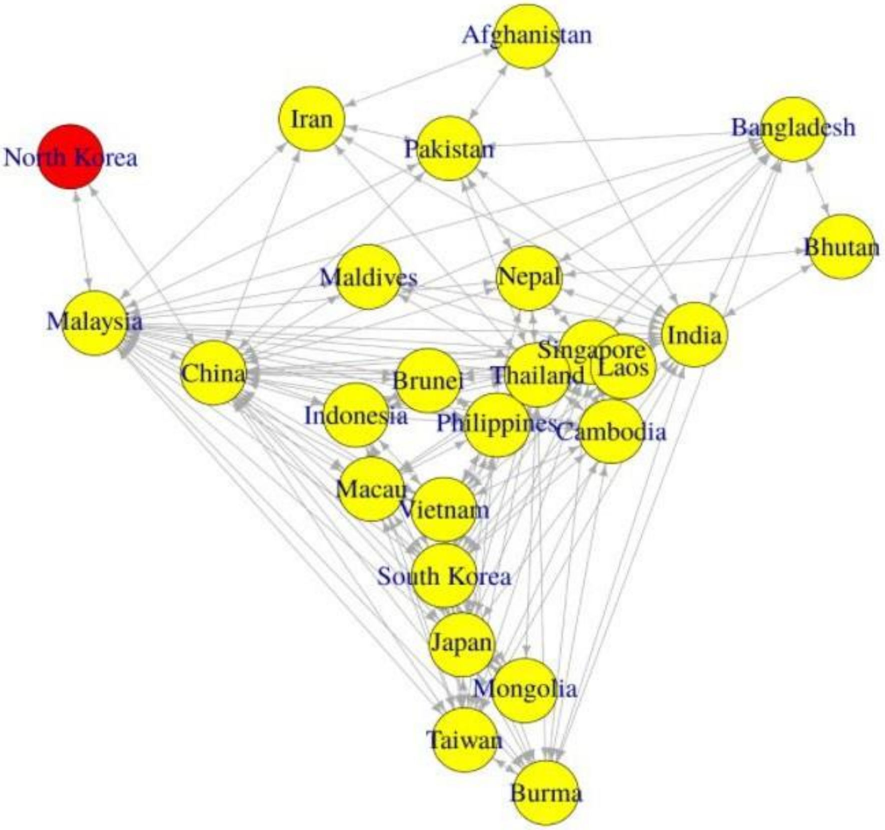
From the graph generated from the below code snippet, it is safe to infer that Malaysia and China act as prominent bridge nodes, having a very high degree centrality and hence holding a high magnitude of importance in the flight network of Asian countries. We see that almost every country has a direct flight to China, hence our social network is seen to obey typical SNA principles like Power Law, by virtue of which if a new country were to be formed, it is more likely to have a direct flight to China or Malaysia than to countries like Bhutan, Bangladesh, etc.

The graph also shows us how weakly North Korea is connected to the network of Asian Countries, having an edge cut of just two, the edge cut of a country can be used as a safe measure to judge how social a particular country is, in the case of a



```r
Country.list <- countries %>%
  select(Country, Region)
Country.list$Country <- as.character(Country.list$Country)
Asian.Country.list <- Country.list %>%
  arrange(Region) %>%
  head(28)
Asian.Country <- Asian.Country.list$Country
Asian.Country <- gsub(pattern = "\\, ", replacement = "", Asian.Country) %>%
  gsub(pattern = " ", replacement = "", Asian.Country) %>%
  gsub(pattern = "KoreaNorth", replacement = "North Korea", Asian.Country) %>%
  gsub(pattern = "KoreaSouth", replacement = "South Korea", Asian.Country)

country.connection <- country.connection %>%
  filter(from %in% Asian.Country & to %in% Asian.Country) %>%
  select(-Link)

country.connection <- country.connection %>%
  mutate(TF = str_detect(from, to)) %>%
  filter(TF == "FALSE") %>%
  select(-TF)
g <- graph_from_data_frame(country.connection, directed = TRUE)
V(g)$color <- ifelse(
  V(g)$name == "North Korea", "red", "yellow"
)
```

flight social network at least. Usually, the more flights a country has, the better international relations it has.



Social Network showing the international flight network of North Korea wrt.other Asian Countries. With an edge cut of just two, we see that it is an antisocial country which does not have international flights to many countries. We also see that Malaysia and China act as bridge nodes for North Korea.

Conclusions

The different visualizations we plotted from the data convey different information and a lot of info can still be inferred from them. We first read in all datasets into the disk and plot all the world's airports on a map, although this looks cluttered and does not convey much info, it gives us a baseline to work with. The nodes which are the airports, are connected via edges that represent whether or not they are joined by direct flights.

Then we analyze the number of airports and airlines via country and represent the same in the form of a barplot and treemap. We see that the United States of America boasts of both the maximum number of airports and airlines. Having such a robust flight network is indicative of the size of the nation's territory and economy.

But does quantity always equate with quality? To answer this question we take the US airline twitter dataset and perform sentiment analysis on the same to get an idea of the overall satisfaction of the fliers using this airline network. A deep learning neural network has been trained for the same that predicts the sentiment of a tweet with an accuracy of over 82%.

Sentiment analysis helps us conclude that the overall sentiment is a negative one, followed by neutral and positive examples.

On the contrary we have countries like North Korea that have a sparse airline network, with a limited number of flights to selective countries like China, Malaysia, etc.

This countries have a high degree centrality in and serve as bridge nodes, connecting North Korea to the rest of the world.

In the scope of the experiment we analyzed the global airline network and even saw the case study between a country having the largest flight network and an anti-social country like North Korea.

The results of social network analysis leads us to the following conclusions-

- The size of the nations' territory effects the number of Airports. The larger the territory, more the number of airports (some exceptions being India and Japan.)

- United States has the most number of airports and airlines.
- There is a positive correlation between the number of airports and airlines a country has.
- Sentiment analysis on the tweets about United States airlines, predicts a general negative trend despite USA having one of the worlds most robust and sophisticated flight networks.
- The economy also effects the number of airports. A booming economy has a greater number of airports as compared to others.
- There is a positive correlation between those 2 variables. However, there are exceptions such as Japan that has less airlines.
- There are just 3 countries that have access to an anti-social country (North Korea).
- Malasia and China act as prominent bridge nodes to keep North Korea connected from the rest of the world.

* * *

Appendix

Source code

```
library(data.table) # fast data import

library(tidyverse) # data manipulation

library(plotly) # interactive visualizations

library(janitor) # data manipulation

library(stringr) # character class data manipulation

library(treemap) # tree map visualization

library(igraph)

library(gridExtra)

library(ggraph)

airport <- read_csv("../input/airports-train-stations-and-ferry-terminals/airports-extended.csv",
col_names = F)

names(airport) <- c("Airpot_ID", "Airport_Name", "City", "Country", "IATA",

          "ICAO", "Latitude", "Longitude", "Altitude", "Timezone",

          "DST", "Tz", "Type", "Source")

airport <- airport %>%

  filter(Type == "airport")

airline <- read_csv("../input/airline-database/airlines.csv") %>%

  clean_names()

route <- read_csv("../input/flight-route-database/routes.csv") %>%

  clean_names()

names(route)[5] <- "destination_airport"

countries <- read_csv("../input/countries-of-the-world/countries of the world.csv")

airport %>%

  head(5) %>%

  DT::datatable(options = list(
```

~~~~
    lengthMenu = c(5,3,1)
  ))
## Dataset 2
```{r}
airline %>%
  head(5) %>%
  DT::datatable(options = list(
    lengthMenu = c(5,3,1)
  ))
```
## Data No.3
```{r}
route %>%
  head(5) %>%
  DT::datatable(options = list(
    lengthMenu = c(5,3,1)
  ))
```
# Analysis
## Global Airports Distribution
```{r warning = FALSE, message = FALSE}
geo <- list(
 scope = "world",
 projection = list(type = "orthographic"),
 showland = TRUE,
 resolution = 100,
 landcolor = toRGB("gray90"),
~~~~

~~~
  countrycolor = toRGB("gray80"),
  oceancolor = toRGB("lightsteelblue2"),
  showocean = TRUE
)
plot_geo(locationmode = "Greenwich") %>%
  add_markers(data = airport %>%
                 filter(Type == "airport"),
               x = ~Longitude,
               y = ~Latitude,
               text = ~paste('Airport: ', Airport_Name),
               alpha = .5, color = "red") %>%
  layout(
    title = "Global Airports",
    geo = geo,
    showlegend = FALSE
  )
```
```{r}
print(paste("There are", airport %>%
               filter(Type == "airport") %>%
               nrow(),
             "airports around the world."))
```
## Global Airline route
```{r}
route <- route %>% mutate(id = rownames(route))
~~~

~~~
route <- route %>% gather('source_airport', 'destination_airport', key = "Airport_type", value = "Airport")

gloabal.flight.route <- merge(route, airport %>% select(Airport_Name, IATA, Latitude, Longitude, Country, City),

   by.x = "Airport", by.y = "IATA")

```

```{r warning = FALSE, message = FALSE}

world.map <- map_data ("world")

world.map <- world.map %>%

filter(region != "Antarctica")

ggplot() +

 geom_map(data=world.map, map=world.map,

      aes(x=long, y=lat, group=group, map_id=region),

      fill="white", colour="black") +

 geom_point(data = gloabal.flight.route,

       aes(x = Longitude, y = Latitude),

       size = .1, alpha = .5, colour = "red") +

 geom_line(data = gloabal.flight.route,

       aes(x = Longitude, y = Latitude, group = id),

       alpha = 0.05, colour = "red") +

 labs(title = "Global Airline Routes")

```

```{r warning = FALSE, message = FALSE, fig.width = 12, fig.height = 7}

ggplot() +

 geom_map(data=world.map, map=world.map,

      aes(x=long, y=lat, group=group, map_id=region),

      fill="white", colour="grey") +

 geom_point(data = airport %>%
~~~

~~~~
             filter(Altitude >= 5000),
            aes(x = Longitude, y = Latitude, colour = Altitude),
            size = .7) +
  labs(title = "Airports located over 5,000 feet altitude") +
  ylim(-60, 90) +
  theme(legend.position = c(.1, .25))
```
```{r}
print(paste(airport %>%
             filter(Altitude >= 5000) %>%
             nrow(),
           "airports are located over 5,000 feet altitude."))
```
## Which Country has the most Airports?


```{r}
connection.route <- route %>%
  spread(key = Airport_type, value = Airport) %>%
select(destination_airport, source_airport, id)
airport.country <- airport %>%
  select(City, Country, IATA)
flight.connection <- merge(connection.route, airport.country, by.x = "source_airport", by.y =
"IATA")
names(flight.connection)[4:5] <- c("source.City", "source.Country")


flight.connection <- merge(flight.connection, airport.country, by.x = "destination_airport", by.y
= "IATA")
names(flight.connection)[6:7] <- c("destination.City", "destination.Country")
~~~~

~~~
flight.connection <- flight.connection %>%
  select(id, contains("source"), contains("destination"))
```
```{r warning = FALSE, message = FALSE, fig.width = 12, fig.height = 7}
data.frame(table(airport$Country)) %>%
  arrange(desc(Freq)) %>%
  head(20) %>%
  ggplot(aes(x = reorder(Var1, -Freq), y = Freq, fill = Var1, label = Freq)) +
  geom_bar(stat = "identity", show.legend = F) +
  labs(title = "Top 20 Countries that has most Airports",
       x = "Country", y = "The number of Airports") +
  geom_label(angle = 45, show.legend = F) +
theme(axis.text.x = element_text(angle = 40, size = 15))
### Treemap Visualization


```{r    fig.width    =    12,    fig.height    =    7}
treemap(data.frame(table(airport$Country)),
        index="Var1",
        vSize="Freq",
        type="index",
        title = "Overall Number of Airport owned by each Nation")
```
## Which Country has the most Airlines?
```{r fig.width = 12, fig.height = 7}
data.frame(table(airline$country)) %>%
arrange(desc(Freq)) %>% head(20) %>%
~~~

~~~
  ggplot(aes(x = reorder(Var1, -Freq), y = Freq,
             fill = Var1, label = Freq)) +
  geom_bar(stat = "identity", show.legend = F) +
  geom_label(show.legend = F) +
  theme(axis.text.x = element_text(angle = 40, size = 15)) +
  labs(x = "Country", y = "The number of Airlines",
       title = "Top 20 Countries that have most airlines")
```
## Airports vs Airlines
```{r warning = FALSE, message = FALSE, fig.width = 12, fig.height = 7}
country.airport <- data.frame(table(airport$Country))
names(country.airport)[2] <- "Airport"

country.airline <- data.frame(table(airline$country))
names(country.airline)[2] <- "Airline"

lineports <- merge(country.airport, country.airline, by = "Var1")
lineports %>%
  ggplot(aes(x = Airport, y = Airline)) +
  geom_point(show.legend = F) +
  geom_smooth() +
  labs(title = "Airports vs Airlines") +
  scale_x_continuous(trans = 'log10',
                     breaks = c(10, 100, 500, 1000))
```
~~~

~~~
## anti-social countries
```{r}
NK.airport <- airport %>% filter(Country == "North Korea")
NK.flight.connection <- flight.connection %>%
  filter(source.Country == "North Korea" | destination.Country == "North Korea")
NK.gloabal.flight.route.id <-
  gloabal.flight.route  %>%
  filter(Country == "North Korea") %>% select(id)
NK.gloabal.flight.route.id <- NK.gloabal.flight.route.id$id %>% as.vector()
NK.gloabal.flight.route.id <-
  gloabal.flight.route %>% filter(id %in% NK.gloabal.flight.route.id)
```


```{r warning = FALSE, message = FALSE, fig.width = 10}
NorthKorea.ggmap <- ggplot() +
geom_map(data=world.map, map=world.map,
           aes(x=long, y=lat, group=group, map_id=region),
           fill="white", colour="black") +
  geom_point(data = NK.gloabal.flight.route.id,
             aes(x = Longitude, y = Latitude), colour = "red") +
  geom_point(data = NK.airport,
             aes(x = Longitude, y = Latitude), colour = "red") +
  geom_line(data = NK.gloabal.flight.route.id,
             aes(x = Longitude, y = Latitude, group = id), colour = "red") +
  xlim(100, 140) + ylim(0, 45) +
  labs(title = "Airports & International Airlines from/to/in North Korea") +
  coord_fixed(ratio = 1.1)
~~~

```
Flight.Country.Connection <- NK.flight.connection %>%
select(contains("Country"), id)
names(Flight.Country.Connection) <- c("From", "To", "id")
Flight.Country.Connection <- Flight.Country.Connection %>%
mutate(Combination = paste0(From, "-", To))
Flight.Country.Connection <- Flight.Country.Connection %>%
group_by(Combination) %>%
  mutate(Weight = NROW(Combination)) %>%
  arrange(-Weight) %>%
  ungroup()
Flight.Country.Connection <-
Flight.Country.Connection[!duplicated(Flight.Country.Connection$Combination),] %>%
  select(-id, -Combination)
Flight.Country.graph <- graph_from_data_frame(Flight.Country.Connection, directed = FALSE)

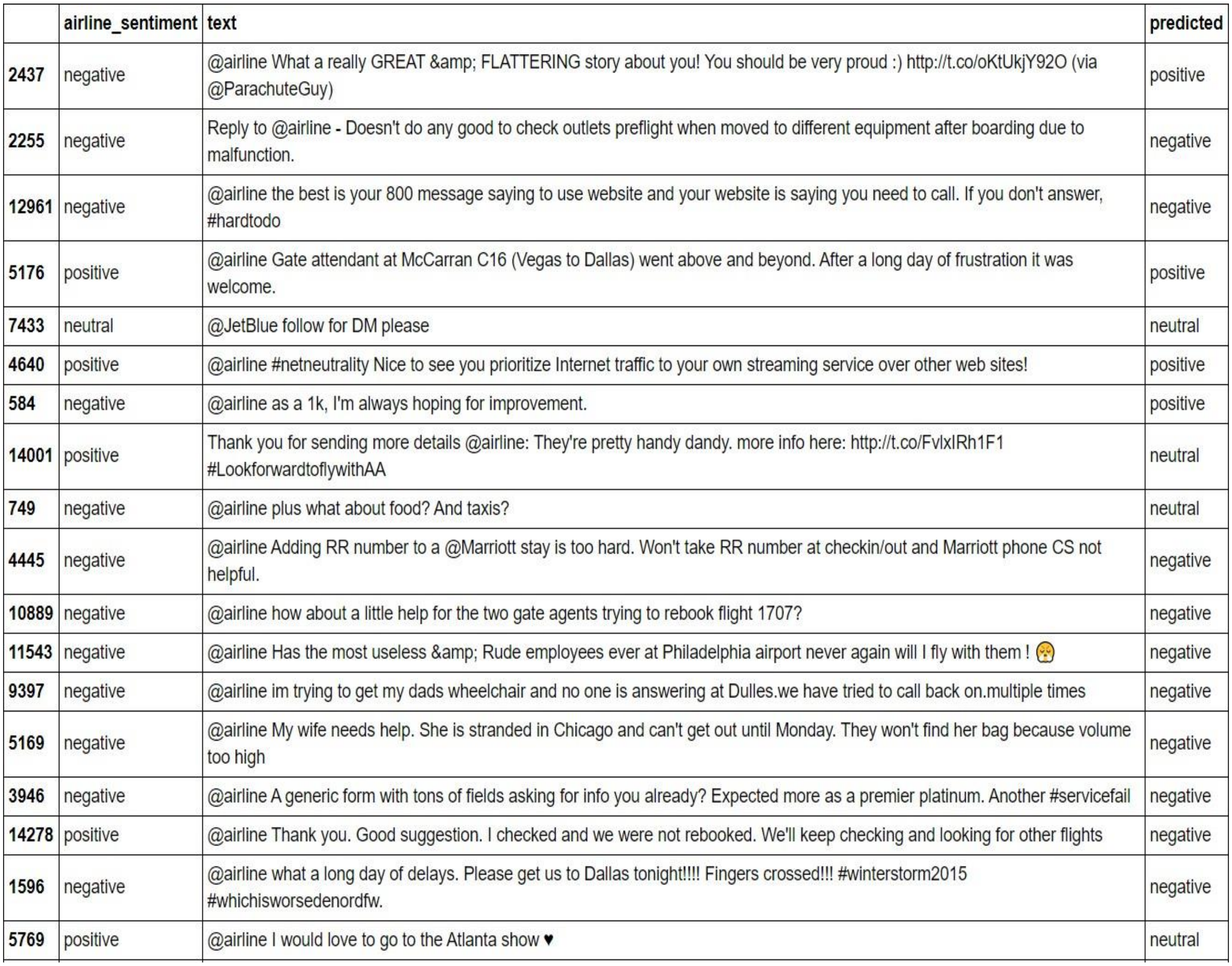
Flight.Country.graph$name <- "Flight Country Network"
V(Flight.Country.graph)$id <- 1:vcount(Flight.Country.graph)

NW.Plot <- ggraph(Flight.Country.graph, layout = "kk") +
  geom_edge_link(aes(alpha = Weight),
                 colour = "red") +
  geom_node_point(size = 5, colour = "red") +
  geom_node_text(aes(label = name), repel = TRUE, size = 7) +
  labs(title = "Flight Country Network", x = "", y = "") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
      axis.line = element_blank(),
      axis.text.x=element_blank(),  axis.text.y=element_blank())
```

```
grid.arrange(NorthKorea.ggmap, NW.Plot, ncol=2)
```
### Airport Information that have flight access to North Korea
```{r}
NK.gloabal.flight.route.id %>%
  filter(Country != "North Korea") %>%
  select(Airport_Name, Country, City, Latitude, Longitude) %>%
  distinct(City, .keep_all = T) %>%
  DT::datatable(options = list(
    lengthMenu = c(4,1)
  ))
```
### Asian Countries' Flights Network

```{r fig.width = 18, fig.height = 13}
country.connection <- flight.connection %>%
select(contains("Country")) %>%
  mutate(Link = paste0(source.Country, "-", destination.Country))
country.connection <- country.connection[!duplicated(country.connection$Link),]
names(country.connection) <- c("from", "to", "Link" )

#### Selecting Asian Countries
Country.list <- countries %>%
select(Country, Region)
Country.list$Country <- as.character(Country.list$Country)
Asian.Country.list <- Country.list %>%
  arrange(Region) %>%
```

~~~
  head(28)
Asian.Country <- Asian.Country.list$Country
Asian.Country <- gsub(pattern = "\\, ", replacement = "", Asian.Country) %>%
  gsub(pattern = " ", replacement = "", Asian.Country) %>%
  gsub(pattern = "KoreaNorth", replacement = "North Korea", Asian.Country) %>%
  gsub(pattern = "KoreaSouth", replacement = "South Korea", Asian.Country)

country.connection <- country.connection %>%
  filter(from %in% Asian.Country & to %in% Asian.Country) %>%
  select(-Link)
country.connection <- country.connection %>%
  mutate(TF = str_detect(from, to)) %>%
  filter(TF == "FALSE") %>%
  select(-TF)
g <- graph_from_data_frame(country.connection, directed = TRUE)
V(g)$color <- ifelse(
  V(g)$name == "North Korea", "red", "yellow"
)
plot(g, layout =  layout_with_dh(g),
     edge.arrow.size=0.8,
     vertex.size = 17, vertex.label.cex = 2)
```
~~~



	airline_sentiment	text	predicted
2437	negative	@airline What a really GREAT &amp; FLATTERING story about you! You should be very proud :) http://t.co/oKtUkjY92O (via @ParachuteGuy)	positive
2255	negative	Reply to @airline - Doesn't do any good to check outlets preflight when moved to different equipment after boarding due to malfunction.	negative
12961	negative	@airline the best is your 800 message saying to use website and your website is saying you need to call. If you don't answer, #hardtodo	negative
5176	positive	@airline Gate attendant at McCarran C16 (Vegas to Dallas) went above and beyond. After a long day of frustration it was welcome.	positive
7433	neutral	@JetBlue follow for DM please	neutral
4640	positive	@airline #netneutrality Nice to see you prioritize Internet traffic to your own streaming service over other web sites!	positive
584	negative	@airline as a 1k, I'm always hoping for improvement.	positive
14001	positive	Thank you for sending more details @airline: They're pretty handy dandy. more info here: http://t.co/FvlxIRh1F1 #LookforwardtoflywithAA	neutral
749	negative	@airline plus what about food? And taxis?	neutral
4445	negative	@airline Adding RR number to a @Marriott stay is too hard. Won't take RR number at checkin/out and Marriott phone CS not helpful.	negative
10889	negative	@airline how about a little help for the two gate agents trying to rebook flight 1707?	negative
11543	negative	@airline Has the most useless &amp; Rude employees ever at Philadelphia airport never again will I fly with them ! 😱	negative
9397	negative	@airline im trying to get my dads wheelchair and no one is answering at Dulles.we have tried to call back on.multiple times	negative
5169	negative	@airline My wife needs help. She is stranded in Chicago and can't get out until Monday. They won't find her bag because volume too high	negative
3946	negative	@airline A generic form with tons of fields asking for info you already? Expected more as a premier platinum. Another #servicefail	negative
14278	positive	@airline Thank you. Good suggestion. I checked and we were not rebooked. We'll keep checking and looking for other flights	negative
1596	negative	@airline what a long day of delays. Please get us to Dallas tonight!!!! Fingers crossed!!! #winterstorm2015 #whichisworsedenordfw.	negative
5769	positive	@airline I would love to go to the Atlanta show ♥	neutral

Table of results obtained post sentiment analysis of flier’s tweets.Our model predicts the sentiment accurately upto 83%.

References

1) https://www.r-project.org/other-docs.html
2) https://pandas.pydata.org/docs/
3) https://matplotlib.org/3.3.2/contents.html
4) https://www.sciencedirect.com/science/article/pii/S2405844015300566#bib0005
5) https://www.aclweb.org/anthology/P18-1031.pdf
6) https://data.world/crowdflower/airline-twitter-sentiment
7) https://www.fast.ai/
8) https://towardsdatascience.com/understanding-language-modelling-nlp-part-1-ulmfit-b557a63a672b
9) https://github.com/anmolpant/ULMFiT-Sentiment
10) https://www.bebr.ufl.edu/sites/default/files/Centrality%20in%20Social%20Networks.pdf
11) https://www.nature.com/articles/30918.
12) https://science.sciencemag.org/content/323/5916/892.abstract

* * *