**Text Analysis of Reviews on Skytrax Website**

**Project Report**

**Fall 2018**

**Submitted by**

Pooja Aishwarya Bonthula (z1834608)

Rohith Sai Katari (z1815368)

**Submitted to**

Dr. Onkar Malgonde

Associate Professor

**In partial fulfillment of**

Social Media Analytics – OMIS 670

Department of OMIS

Northern Illinois University

**Table of Contents**

1. **Introduction**
2. **Tools Used**
3. **Data Scraping**
4. **Sentiment Analysis**
5. **Topic Modeling**
6. **Visualization**
7. **Recommendation for users**
8. **Recommendation for Airlines**
9. **Introduction:**

Skytrax is a well-known airline consultancy which is based in UK, it runs an airline and airport ranking and review site. It takes surveys from travelers in order to rate airports, airlines, cabin crew, in-flight entertainment etc, and also conducts inspection for commercial airlines. Additionally, Skytrax has an airline channel where travelers give implicit passengers views and feedback about an airline. The website also anchors flight retrospection surveys. It is responsible for giving out annual World Airline and Airport Awards.

1. **Tools Used:**

* Python for Data Scrapping, Data cleaning and analysis
* PowerBI for Data Visualization

1. **Data Scraping:**

Data scraping which is also called web scrapping is a technique of importing data from a webpage into a spreadsheet which is a saved on your workspace. It is one of the most effective technique to extract information from the internet, and in some scenarios to transmit that data to a different website.

Data Scrapping is used in:

* Examination for web content business intelligence
* Comparison and pricing sites for travelers
* Spotting sales leads/performing market research by studying public data sources
* Forwarding product information from an e-commerce site to a different online vendor

It has a varied number of applications. It can be used effectively whenever there is a need to move data from one place to another.

A screenshot of a social media post

Description automatically generated

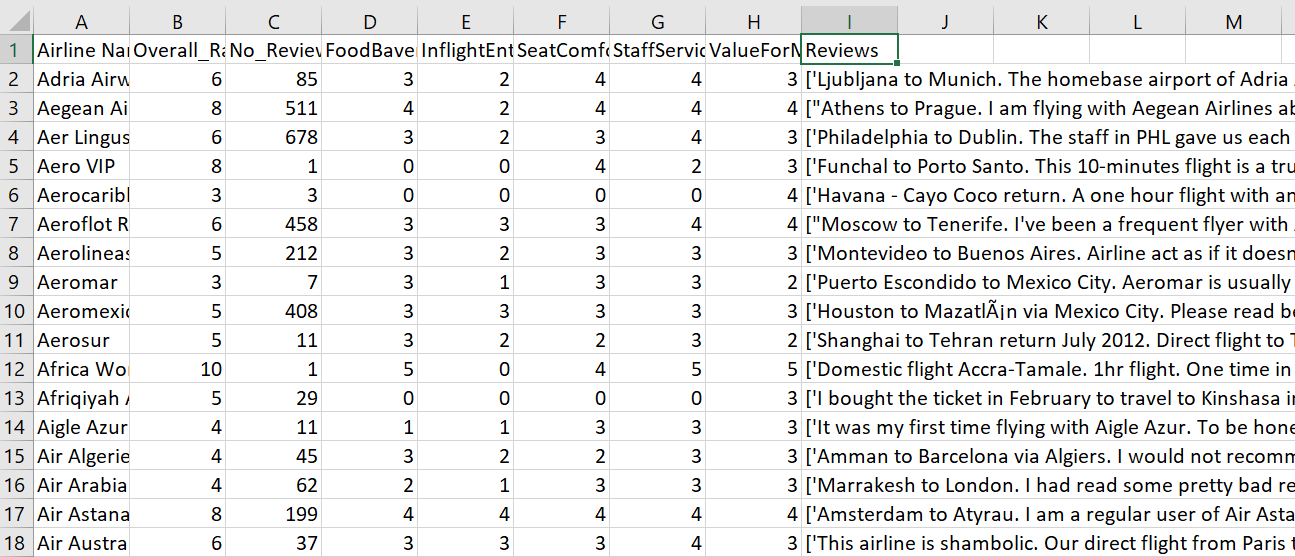
The above code is the snippet of the code we used for data scrapping for this project. We used several python libraries like urllib, requests, os, csv, re, sys, time, BeautifulSoup and Selenium to extract data from Skytrax website.

**3.1 BeautifulSoup:**

BeautifulSoup is a python library which is used for scraping data from webpages. It is used for parsing webpage files like html and XML.

**3.2 Selenium:**

Selenium is a open source tool which is used for web based automation. Selenium python api also helps in connecting with various web browsers like mozilla, chrome etc.. It also helps user to send various python packages for browsers like mozilla. In this project we used selenium tool and Geckodriver to automate the scraping process.



The above fig shows the sample data we got after data scraping. We have got Airline name, Overall rating of the airline, No of Reviews, Food Beverage rating, Inflight Entertainment rating, Seat Comfort rating, Staff Service rating, Value for money rating and reviews.

1. **Sentiment Analysis:**

Sentiment analysis which is mainly called as opinion mining is related to the use of natural language processing, text analysis, biometrics and computational linguistics to methodically extract, identify, quantify and study affective states and subjective data. Opinion mining is vastly applicable to voice of the customer materials such as feedback and survey responses, healthcare materials, online and social media for applications that vary from marketing to customer service to clinical medicine.

A screenshot of a social media post

Description automatically generated

We used TextBlob package in python for Sentiment analysis. TextBlob is a python [package which is used mainly for text processing. It is used in NlP for many tasks like part of speech tagging, Sentiment analysis, classification etc. After performing sentiment analysis we got polarity and subjectivity values of each review of the airline.

A close up of text on a white background

Description automatically generated

The above fig is the sample output after sentiment analysis. By looking into the value of polarity we can decide whether statement is positive, negative or neutral. If the polarity ranges from [-1,0] then the statement is considered as negative statement. If the polarity value ranges between [0,1] then it is positive statement. If the value of polarity value is equal to 0 then it is a neutral statement.

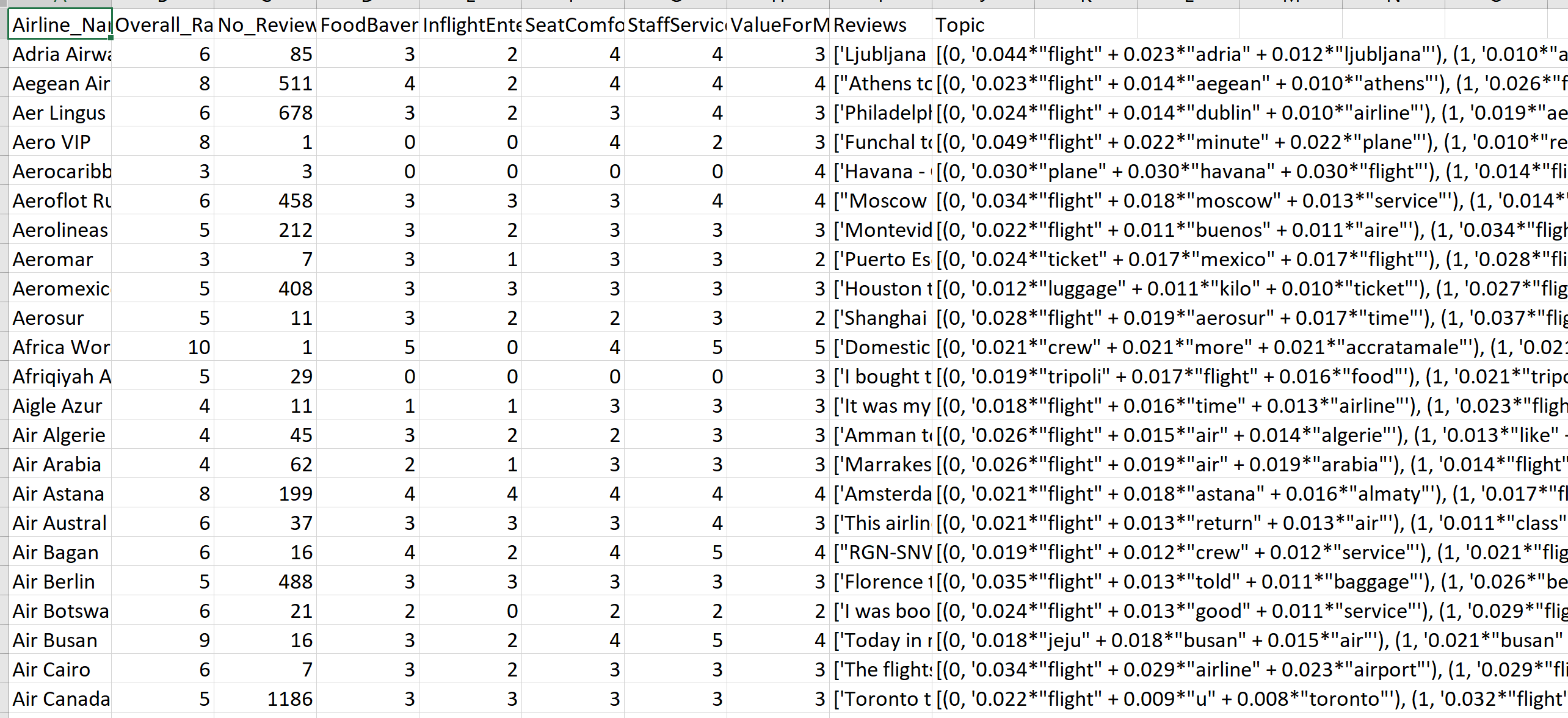
1. **Topic Modeling:**

Topic Modeling is a technique in machine learning and natural inspired processing which is used to discover the abstract “topics” that occur in set of text documents. It is one of the frequently used statistical model used for discovering hidden semantic structures of the text body. It is also referred to a probabilistic topic model which is used to discover latent semantic structures of the given text body.

A screenshot of a social media post

Description automatically generated

The above fig is the snippet of the code we used for the topic modeling in this project. With topic modeling we discovered hidden semantics of the user reviews. We used LDA technique for topic modeling.



The above fig shows the sample output after topic modeling. We can see Topic model is added which gives the hidden semantics (Words) for each airline review.

1. **Visualization:**

Visualization is a data visualizing technique used to represent given data in the form of graphical pictures or animations. This technique is mainly used for data communication. It plays a crucial role in analyzing data. Visualization is also used for future predictions. We used Power BI tool for visualization. Power BI is developed by Google for business analytics. Power BI tool is basically used to transform, retrieve, analyze and report the data.

* 1. **Airline vs Total number of Reviews:**

A screenshot of a cell phone

Description automatically generated

For better visualization we visualized top 5 most reviewed airlines. The above graph is a simple bar graph of airlines vs no of reviews. In the above graph x axis consists of airline names and y axis consists of count of reviews. We can see British airlines is the most reviewed airline with more than 2500 reviews. It is followed by Spirit airlines, American airlines, United airlines and Emirates respectively.

* 1. **Airline vs Overall Rating:**

A screenshot of a cell phone

Description automatically generated

The above graph depicts Airline vs Overall rating. X axis shows the different airlines and y axis represents the overall rating. From the above graph, we can see that Emirates has the highest number of ratings followed by British Airways. American, Spirit and United Airlines share the same number of overall ratings.

**6.3 Airline vs Food Beverage and Inflight Entertainment Rating:**

A screenshot of a cell phone

Description automatically generated

The above graph represents Airlines vs Food and Beverages and Inflight Entertainment Rating. X axis shows the top 5 most reviewed airlines whereas, y axis show shows Food and Beverages and Inflight Entertainment. By looking at the graph we can say that Emirates has the highest inflight entertainment. British Airwatys and Emirates share the same rating for food and beverages while American and United Airlines share the same ratings for Food and Beverages and Inflight Entertainment.

* 1. **Airline vs Staff service and Seat Comfort Rating:**

A screenshot of a cell phone

Description automatically generated

The above graph shows Airlines vs Staff Service and Seat Comfort ratings. X axis shows the top 5 airlines and y axis shows Staff Service and Seat Comfort. Brown color shows the seat comfort while blue color shows the staff service. We can see that Emirates has the highest rating for seat comfort. Emirates and British Airways share the same rating for saet comfort and staff service followed by American, Spirit and United airlines.

* 1. **Average of polarity and Subjectivity:**

A screenshot of a computer

Description automatically generated

The above graph shows average of polarity and subjectivity. X axis represents the different airlines and y axis represents average of polarity and subjectivity. Green color depicts the average of subjectivity while black color depicts the average of polarity. Vistara has the highest average of polarity which means most of the customers have given a positive feedback for this airline.

* 1. **British Airways Sentiment:**

A close up of a logo

Description automatically generated

The above pie chart represents overall sentiment of British Airline reviews.   
We can see that the pie chart has more than fifty percent of positive sentiment.

* 1. **Topic Modeling:**

**A screenshot of a cell phone

Description automatically generated**

The above fig shows the hidden semantics and their weightages of the emirates airline’s reviews. Hence it is a topic model of the emirates airline we can see flight, Dubai and emirate has more weightage. By seeing above fig we can also infer most of the passengers of the emirates are concerned about the meal time, Cabin crew and seats.

* 1. **Word Cloud:**

A picture containing text, map

Description automatically generated

Word Cloud is an illustration which is composed of various words related to a particular topic, where every word represents its significance or frequency. From the above shown word cloud we can say that flight is the most used word in most of the reviews and words like time, airline, seat etc. are next most used words. We can also see that many places names are mentioned.

A close up of a logo

Description automatically generated

The above word cloud explicitly shown for British Airways. We can say that most of the passengers travelled by British Airlines were concerned about seat selection and cabin crew

1. **Recommendations for users:**

By considering the study done on different airlines we can see that Emirates and British airlines share the highest number of ratings and reviews in all departments. Hence we can recommend these airlines for a comfortable and satisfying journey.

1. **Recommendations for Airlines:**

We can see from the above study that Spirit and United airlines did not receive much higher ratings or positive reviews in many areas like seat comfort, cabin staff etc. We recommend these airlines to improve accordingly for a better functioning and performance.