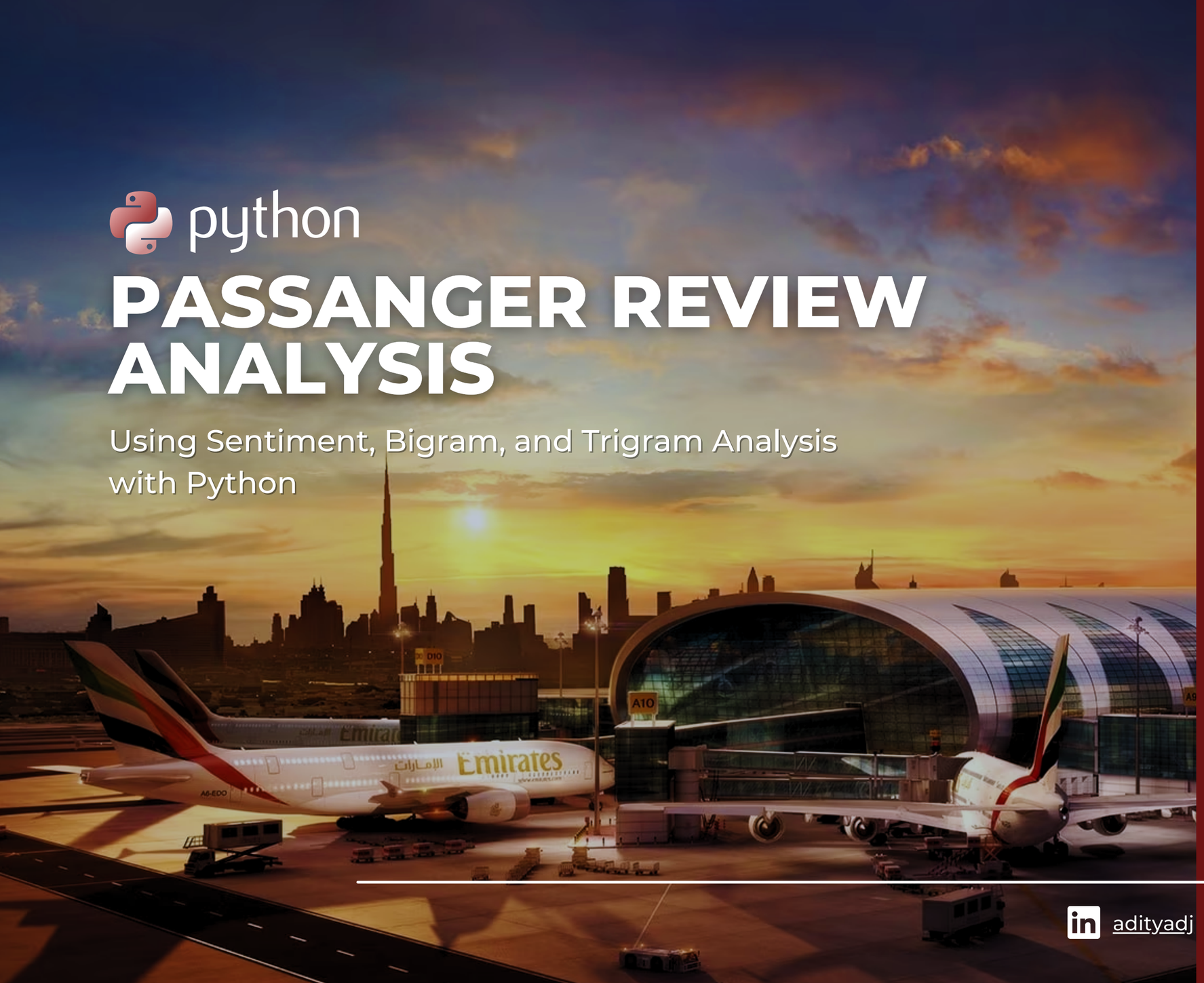


Aditya D. Jati

adityadamarjati.adj@gmail.com

A data enthusiast who certified as *Data Analyst Associate*. Registered as Bachelor of Engineering from Institut Teknologi Sepuluh Nopember. Familiar with *Python*, *SQL*, *Tableau* and *SAP Fundamentals* for Business Analyst.

I created this with the aim of providing learning for myself and anyone who is interested in. So, happy learning!

CLICK TO FIND ME :

adityadj | @adityadj | @adityadj98 | Quora

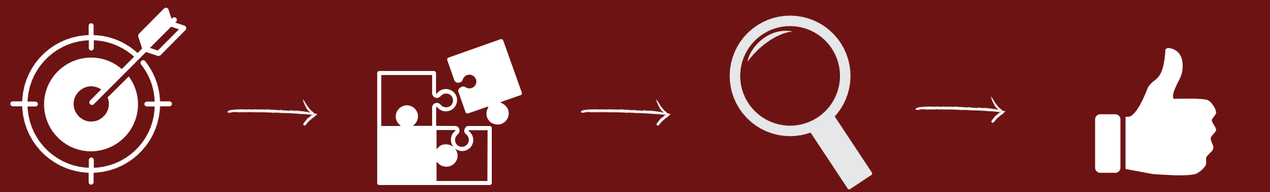
python 

Task

Your manager is concerned about the service provided by **Emirates Airlines**. Therefore, He want to know the opinions of thousands of passengers regarding the service they have received. They want to know which aspects of the service are of concern to the customers.

With the data available so far, which is the reviews from Skytrax, is it possible to find out which aspects of the service are of concern to the customers?

Outline



Airlines :

Emirates

Data sources :

SKYTRAX

www.airlinequality.com

adityadj | @adityadj | @adityadj98 | Quora

python 

Objective

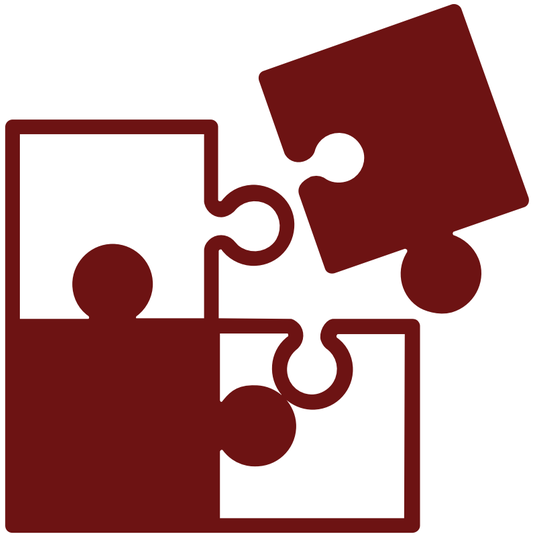
adityadj | @adityadj | @adityadj98 | Quora

OBJECTIVE

Scraping Data From Skytrax

Get insight about which services are concerned by the passenger

python 



Preparation

adityadj | @adityadj | @adityadj98 | Quora

PREPARATION

First, I try to determine which data will be taken for analysis.

Beautifulsoup

Then, I try to scrap the data using beautifulsoup python-package.

pandas

Then stored all of them as dataframe using pandas

8/10

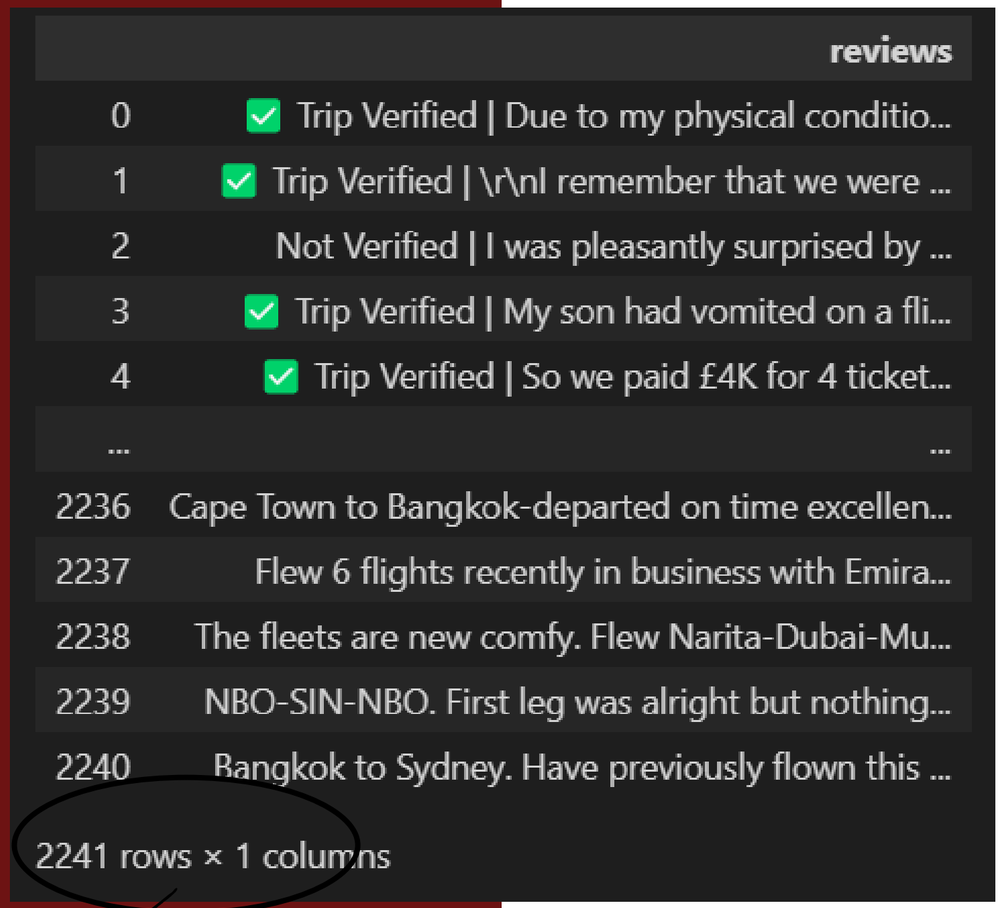
"an exceptional flight"

Zeenith Moola (South Africa) 25th March 2023

✅ *Trip Verified* | I had an exceptional flight, the staff were considerate and polite. Cabin staff Aaliyah was kind and generous and showed thoughtfulness. I will definitely continue to travel with Emirates again

Type Of Traveller	Solo Leisure
Seat Type	Economy Class
Route	Dubai to Durban
Date Flown	March 2023
Seat Comfort	★★★☆☆
Cabin Staff Service	★★★★★
Food & Beverages	★★★★☆
Inflight Entertainment	★★★☆☆
Ground Service	★★★★☆
Wifi & Connectivity	★★★☆☆
Value For Money	★★★★☆
Recommended	✔

Output :



	reviews
0	✅ Trip Verified | Due to my physical conditio...
1	✅ Trip Verified | \r\nI remember that we were ...
2	Not Verified | I was pleasantly surprised by ...
3	✅ Trip Verified | My son had vomited on a fli...
4	✅ Trip Verified | So we paid £4K for 4 ticket...
...	...
2236	Cape Town to Bangkok-departed on time excellen...
2237	Flew 6 flights recently in business with Emira...
2238	The fleets are new comfy. Flew Narita-Dubai-Mu...
2239	NBO-SIN-NBO. First leg was alright but nothing...
2240	Bangkok to Sydney. Have previously flown this ...

2241 rows × 1 columns

It mean we just get 2241 data review from scraping process before

Preparation

python

Finally we can get the whole data in pandas dataframe. With this form we can process data more easily both cleaning data and visualizing it because it is more flexible.

But we got another situation that must be concerned

There are 2 major text patterns that will interfere with the analysis. This can be said to be annoying because it can lead to ambiguity when doing word counting. The disturbing pattern is as follows :

✅ Trip Verified | Not Verified |

We can removed that unimportant part by creating a function.

Creating a function

```python
# So I want to removed the unimportant part using this code
def process_reviews(reviews):
    if reviews.startswith('✅ Trip Verified'):
        return reviews.split('|', 1)[1]
    elif reviews.startswith('Not Verified'):
        return reviews.split('|', 1)[1]
    else:
        return reviews
```

Creating a function to perform data cleaning will make it easier for us if we want to do the same thing for the next process without the need to rewrite the code.

Result :

	reviews
0	Due to my physical condition, I fly in Busin...
1	\r\nI remember that we were very satisfied wi...
2	I was pleasantly surprised by level of servi...
3	My son had vomited on a flight where my part...
4	So we paid £4K for 4 tickets to be upgraded ...
...	...
2236	Cape Town to Bangkok-departed on time excellen...
2237	Flew 6 flights recently in business with Emira...
2238	The fleets are new comfy. Flew Narita-Dubai-Mu...
2239	NBO-SIN-NBO. First leg was alright but nothing...
2240	Bangkok to Sydney. Have previously flown this ...

2241 rows × 1 columns

	reviews	Cleaned Reviews	POS tagged	Lemma	Sentiment	Analysis
0	Due to my physical condition, I fly in Busin...	Due to my physical condition I fly in Busines...	[(Due, a), (physical, a), (condition, n), (fly...	Due physical condition fly Business always r...	0.1531	Neutral
1	\r\nI remember that we were very satisfied wi...	I remember that we were very satisfied with t...	[(remember, v), (satisfied, a), (airline, n), ...	remember satisfied airline year ago time dis...	0.1779	Neutral
2	I was pleasantly surprised by level of servi...	I was pleasantly surprised by level of servic...	[(pleasantly, r), (surprised, v), (level, n), ...	pleasantly surprise level service experience...	0.9022	Positive
3	My son had vomited on a flight where my part...	My son had vomited on a flight where my partn...	[(son, n), (vomited, v), (flight, n), (partner...	son vomit flight partner leave seat request ...	0.5106	Positive
4	So we paid £4K for 4 tickets to be upgraded ...	So we paid K for tickets to be upgraded to bu...	[(paid, v), (K, n), (tickets, n), (upgraded, v...	pay K ticket upgrade business class family w...	0.7579	Positive

Regex

First I try to clean the data from any special character for example : "" !@#$%^&*()

I try to do this process by using regex package from python.

NLTK

The nltk.corpus package defines a collection of corpus reader classes, which can be used to access the contents of a diverse set of corpora.

Using Natural Language Toolkit (NLTK), It will be removed unimportant word like "was", "were", "I", "you" etc.

Lemmatization

Lemmatization is the process of transforming a word to its base form (lemmatization) by removing word inflections such as suffixes and prefixes, resulting in a base word called a lemma.

The purpose of lemmatization is to obtain a base word that has the same meaning as other forms of the word.

Vander Sentiment

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a rule-based sentiment analysis tool that is specifically designed for analyzing the sentiment of text

I used this package to make function to defined a word to a score "is it negative or positive review?"

output (pandas data frame):

```
Positive    1407
Negative     620
Neutral      214
Name: Analysis, dtype: int64
```

Preparation

python

adityadj | @adityadj | @adityadj98 | Quora

By using data from the "Lemma" column from the previous data frame, I tried to create another 3 data frames to get clues in the form of insights from data reviews written by passangers.

"CLICK FOR ALL CODE"

You can click those if you need to view my code

Word Counting

	Word	Frequency
0	flight	4686
1	emirates	3326
2	dubai	2909
3	seat	2221
4	service	2137
...	...	...
8217	zh	1
8218	zimbabwe	1
8219	zonal	1
8220	zoo	1
8221	zuerst	1

8222 rows × 2 columns

This data frame is prepared for counting a word. Just word by word.

Bigram Counting

	Bigram	Frequency
0	via dubai	704
1	business class	674
2	cabin crew	551
3	fly emirates	409
4	verified review	379
...	...	...
104884	nice hand	1
104885	hand out	1
104886	out airline	1
104887	however think	1
104888	seating unfriendly	1

104889 rows × 2 columns

Bigram is an n-gram in natural language processing and text analysis that consists of two words that appear consecutively in a text. In a bigram, the text is divided into a sequence of consecutive word pairs without overlapping.

Trigram Counting

	Trigram	Frequency
0	via Dubai Emirates	104
1	time fly Emirates	49
2	Bangkok via Dubai	46
3	London via Dubai	41
4	business class seat	39
...	...	...
163231	Bangkok reassure would	1
163232	reassure would make	1
163233	make connection would	1
163234	connection would ground	1
163235	seating unfriendly service	1

163236 rows × 2 columns

Trigram is an n-gram in natural language processing and text analysis that consists of three words that appear consecutively in a text. In a trigram, the text is divided into a sequence of consecutive three-word units without overlapping.

python 

Exploration

adityadj | @adityadj | @adityadj98 | Quora

EXPLOREATION

Previously we have created 4 data frames that are ready to be visualized. From data visualization, we can explore what one data means and what it means when connecting data.

Package for visualization

matplotlib

Matplotlib is a Python package for data visualization that provides a wide range of tools to analyze and present data, including various types of charts such as scatter plots, line plots, and histograms. It is highly flexible and customizable to meet specific user needs and can be integrated with other Python libraries like NumPy and Pandas.

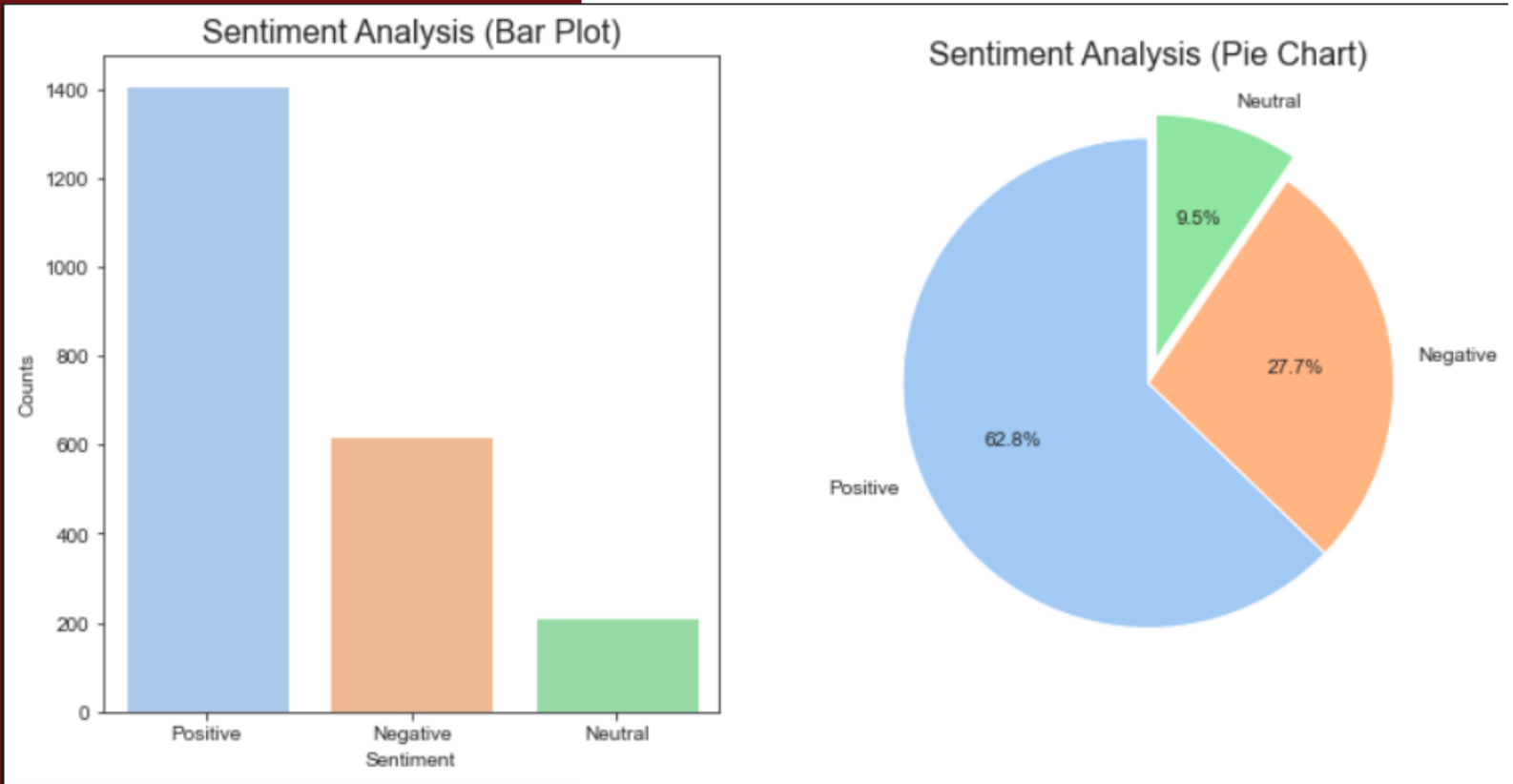
seaborn

Seaborn is a data visualization library based on Matplotlib that provides a higher-level interface to create more attractive and informative statistical graphics. It simplifies the process of creating complex visualizations, has automatic color palettes, and is designed to work well with Pandas dataframes. Seaborn is a valuable tool for data scientists and researchers who want to create informative and visually appealing visualizations with minimal coding effort.

Exploration

python

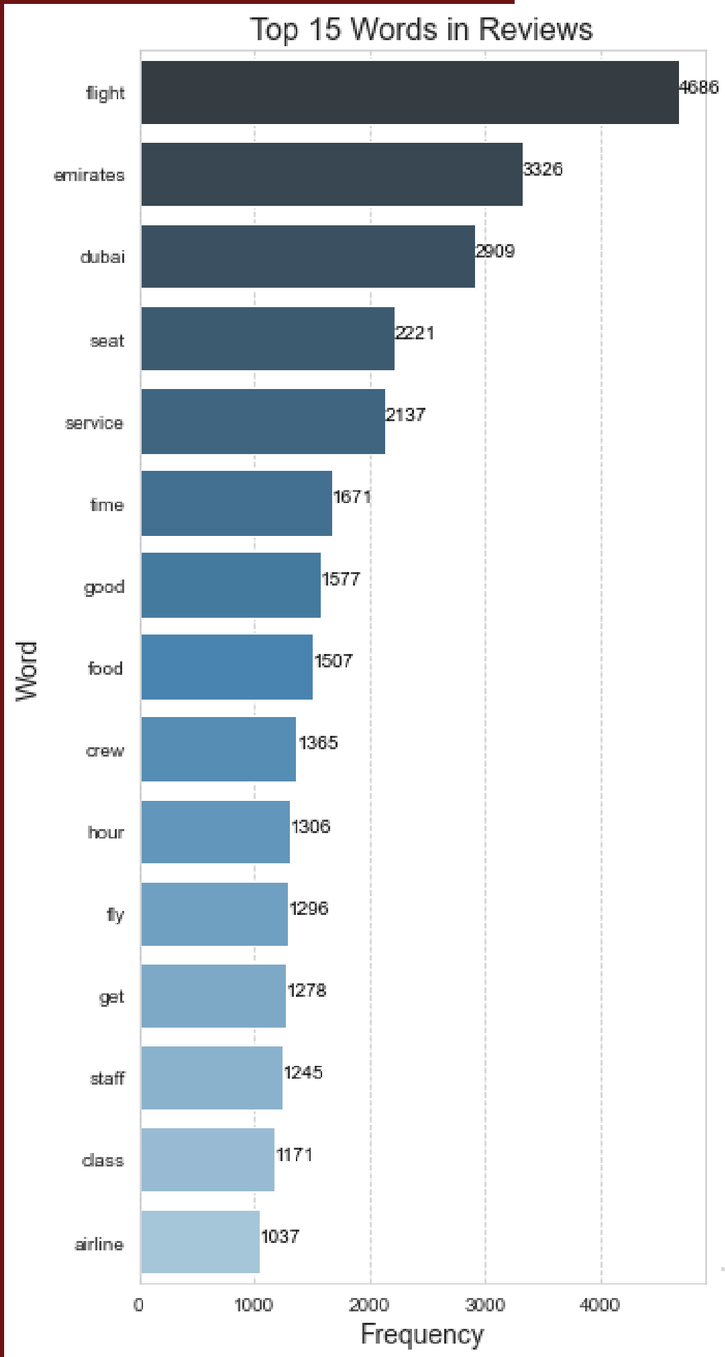
Sentiment Analysis



The given data represents the results of sentiment analysis of text reviews on Emirates Airlines' services. In this analysis, the text reviews were classified into three sentiment categories, namely positive, negative, and neutral.

The data shows that out of all the reviews, 1407 were categorized as positive, 620 were categorized as negative, and 214 were categorized as neutral. This means that the majority of reviews about Emirates Airlines' services are positive.

This can be seen as an indication that most users of Emirates Airlines are satisfied with the services provided, including flight services and other services available at Emirates Airlines. However, although most of the reviews are positive, there is still a small percentage of reviews that are categorized as negative, which can be a consideration for Emirates Airlines to continue improving and enhancing the quality of the services provided.



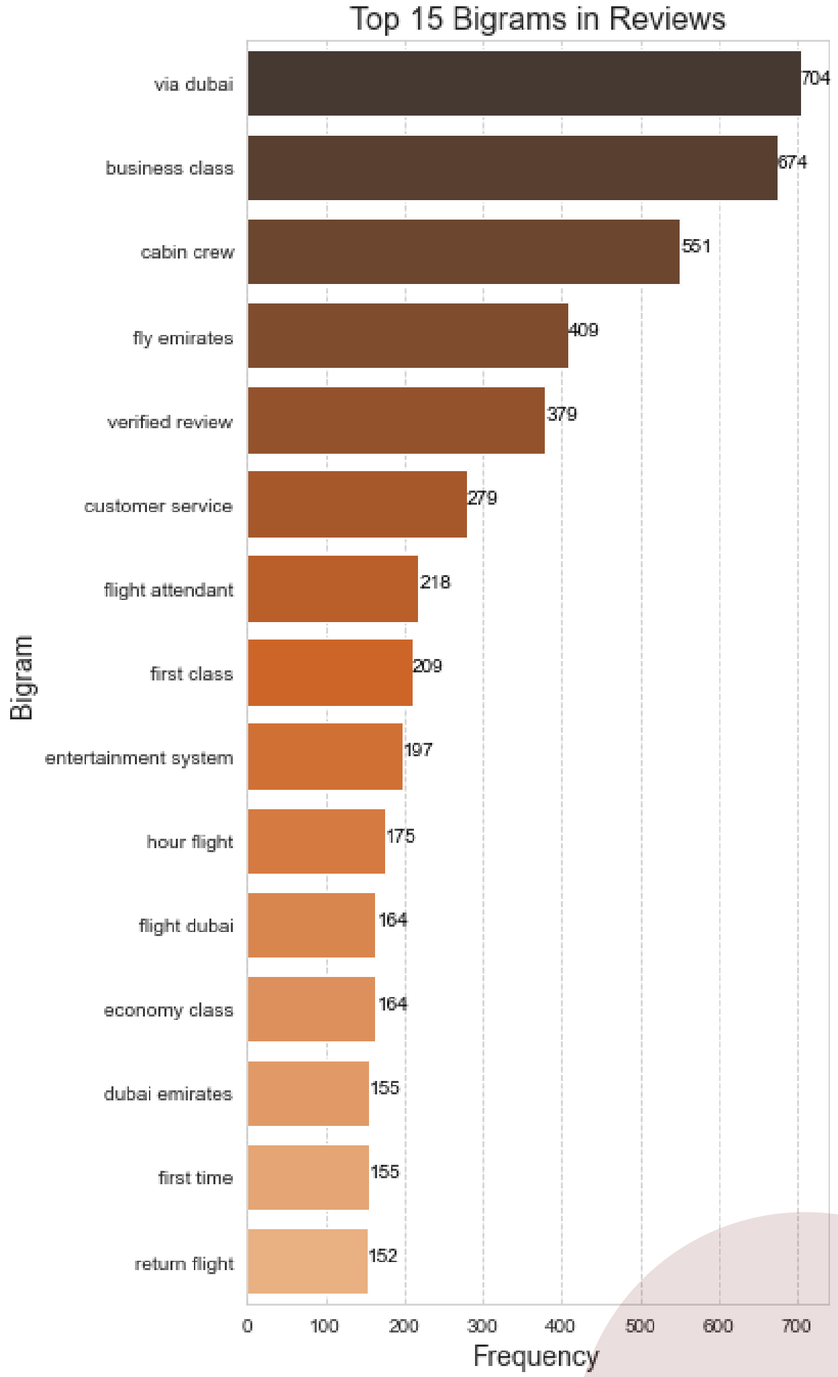
Exploration

python

Word Counting Analysis

This data frame shows the most frequently occurring words in the analyzed text. From this data, we can see that the word "flight" appears the most, with a count of 4686, followed by "emirates" with a count of 3326, and "dubai" with a count of 2909. Other common words include "seat" (2221) and "service" (2137).

Based on this analysis, we can conclude that the text primarily discusses experiences related to flying with Emirates and visiting Dubai.



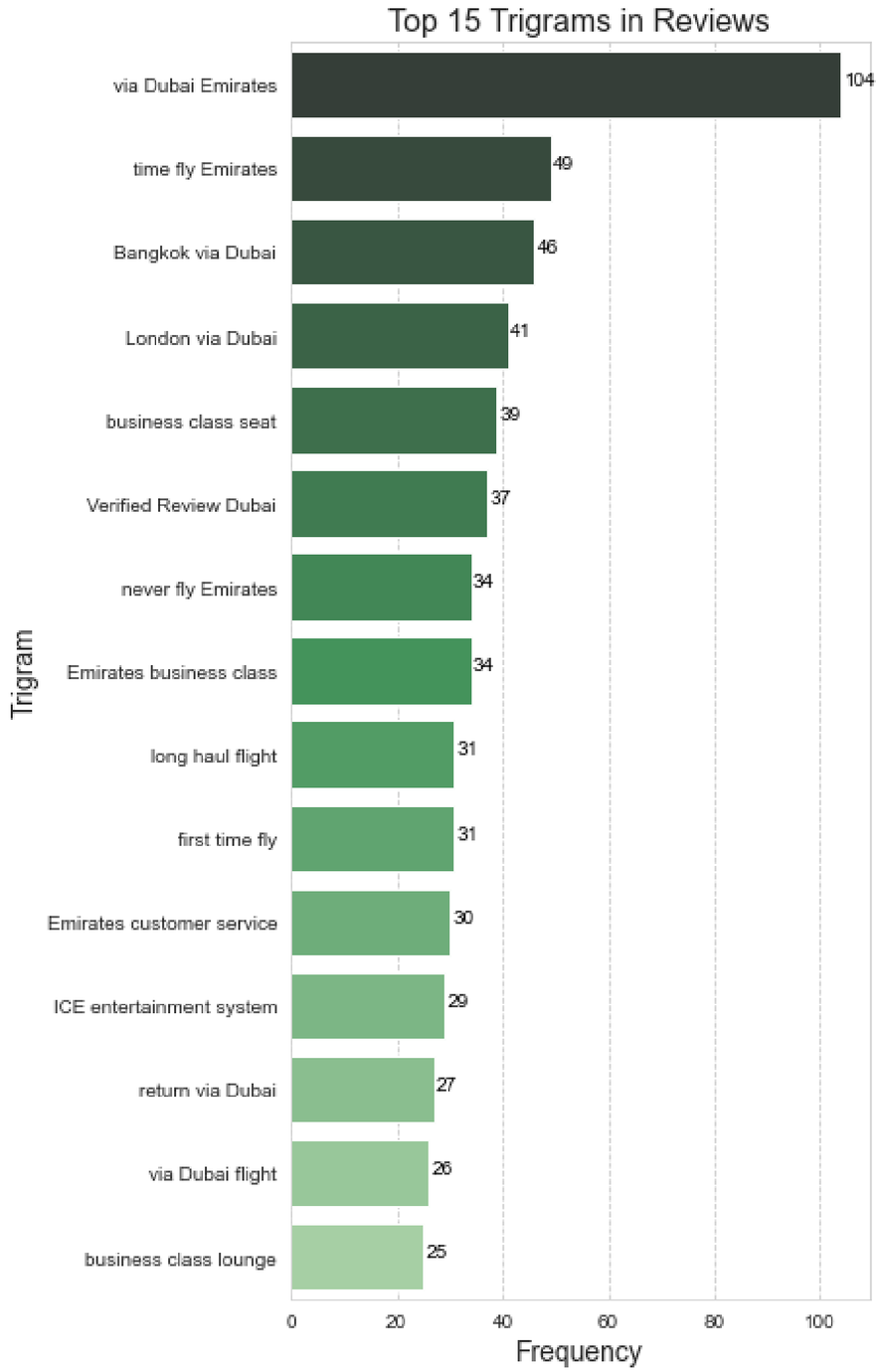
Exploration

python

Bigram Analysis

This data frame shows the most frequently occurring pairs of words in the analyzed text. From this data, we can see that the most common bigram is "via dubai" with a frequency of 704, followed by "business class" with a frequency of 674, and "cabin crew" with a frequency of 551. Other common bigrams include "fly emirates" (409) and "verified review" (379).

Based on this analysis, we can conclude that the text primarily discusses experiences related to flying via Dubai with business class and interactions with the cabin crew.



Exploration

python

Trigram Analysis

This data frame shows the most frequently occurring groups of three words in the analyzed text. From this data, we can see that the most common trigram is "via dubai emirates" with a frequency of 104, followed by "business class seat" with a frequency of 60, and "emirates business class" with a frequency of 56. Other common trigrams include "time fly emirates" (49) and "bangkok via dubai" (46).

Based on this analysis, we can conclude that the text primarily discusses experiences related to flying via Dubai with business class, comfortable seating, and Emirates' services.

python 

Conclusion

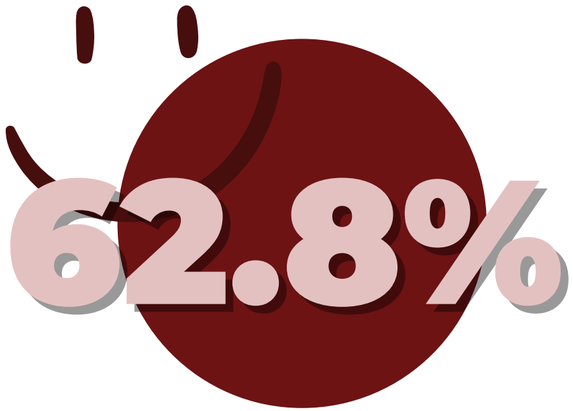
adityadj | @adityadj | @adityadj98 | Quora

Conclusion

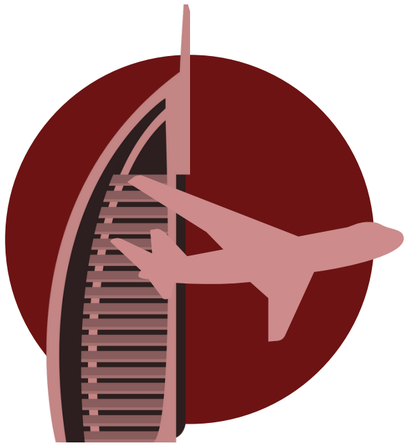
Exploration

python

Scraping data from Syntrax is very possible using python with the beautifulsoup and pandas packages.



62.8% of reviews describe the service as positive for Emirates Airlines. While 27.7% are negative and the others are neutral.



Flights to Dubai with business class are a topic that is always mentioned by reviewers. Then followed by comfortable seats and cabin crew.

I think there is still much to be evaluated from my work. But hopefully it can be useful. thank you!

THANK YOU

adityadj | @adityadj | @adityadj98 | Quora