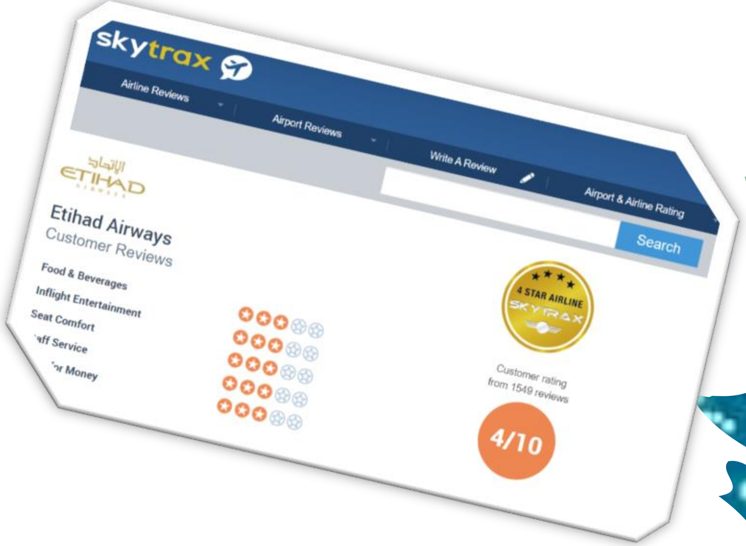
Social Data Mining Techniques

Airlines Analysis

Introduction

Data Source : Skytrax (Airlinequality)

4000+ Reviews



skytrax

Airline Reviews | Airport Reviews | Write A Review | Airport & Airline Rating

Search

الإتحاد ETIHAD

Etihad Airways
Customer Reviews

Food & Beverages
Inflight Entertainment
Seat Comfort
aff Service
or Money

4 STAR AIRLINE SKYTRAX

Customer rating from 1549 reviews

4/10

https://www.airlinequality.com/airline-reviews

Objective

- Web scrape the data
- Text Analysis
- Store data in the remote database
- Visualize in Power BI

Identified Airlines

★★★
3 STAR AIRLINE
SKYTRAX

QATAR
AIRWAYS القطرية

الاتحاد
ETIHAD

American Airlines

TURKISH AIRLINES

Austrian

flydubai

Lufthansa

UNITED
AIRLINES

Tools Used

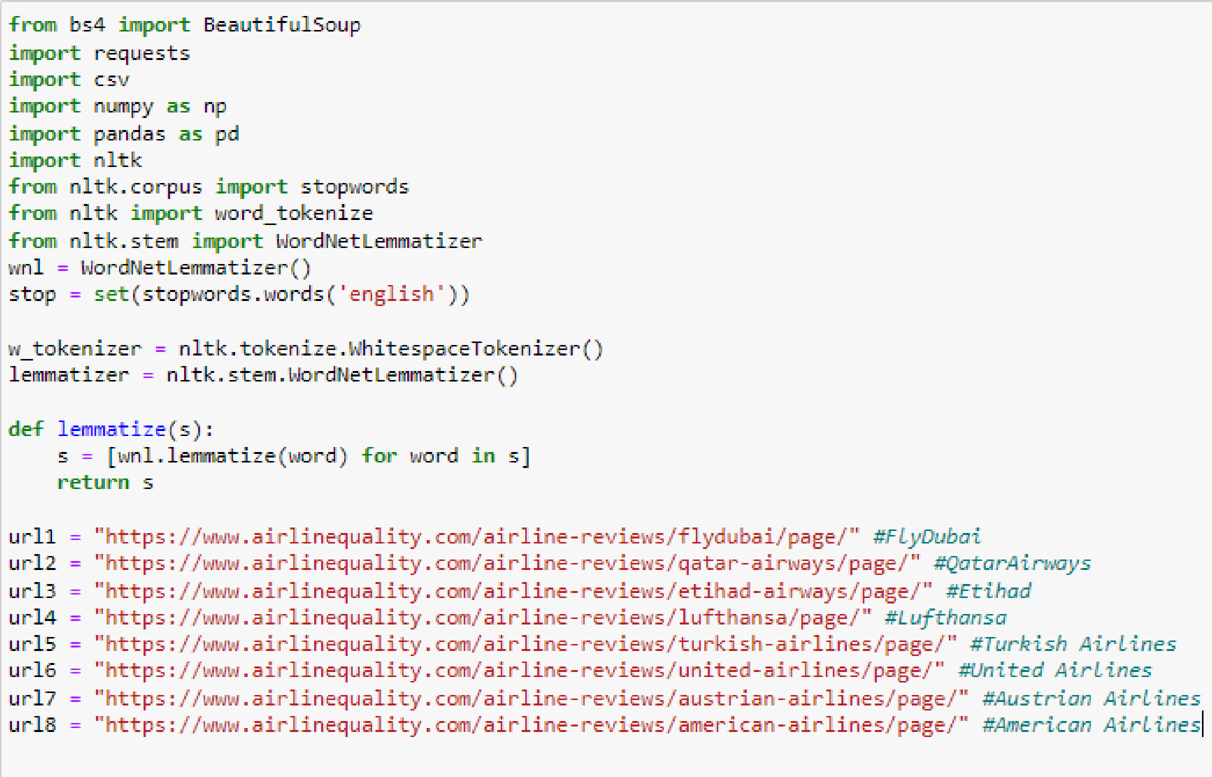
jupyter

Beautiful Soup

mongoDB

TextBlob

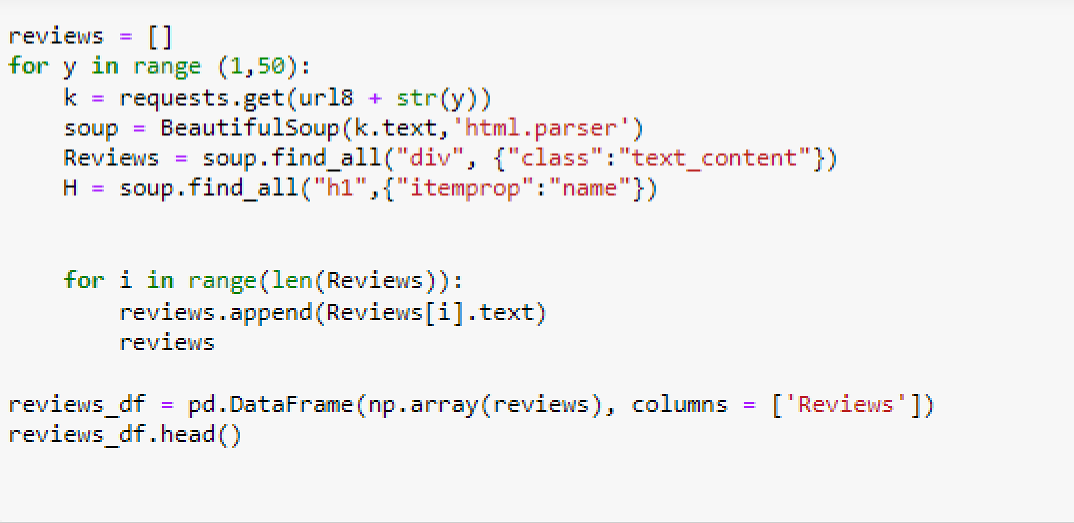
Coding Snippets



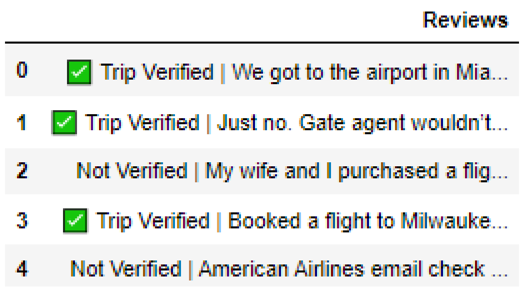
```python
from bs4 import BeautifulSoup
import requests
import csv
import numpy as np
import pandas as pd
import nltk
from nltk.corpus import stopwords
from nltk import word_tokenize
from nltk.stem import WordNetLemmatizer
wnl = WordNetLemmatizer()
stop = set(stopwords.words('english'))

w_tokenizer = nltk.tokenize.WhitespaceTokenizer()
lemmatizer = nltk.stem.WordNetLemmatizer()

def lemmatize(s):
    s = [wnl.lemmatize(word) for word in s]
    return s

url1 = "https://www.airlinequality.com/airline-reviews/flydubai/page/" #FlyDubai
url2 = "https://www.airlinequality.com/airline-reviews/qatar-airways/page/" #QatarAirways
url3 = "https://www.airlinequality.com/airline-reviews/etihad-airways/page/" #Etihad
url4 = "https://www.airlinequality.com/airline-reviews/lufthansa/page/" #Lufthansa
url5 = "https://www.airlinequality.com/airline-reviews/turkish-airlines/page/" #Turkish Airlines
url6 = "https://www.airlinequality.com/airline-reviews/united-airlines/page/" #United Airlines
url7 = "https://www.airlinequality.com/airline-reviews/austrian-airlines/page/" #Austrian Airlines
url8 = "https://www.airlinequality.com/airline-reviews/american-airlines/page/" #American Airlines
```

Importing all the required libraries &
defining each airline review URL

Scrape the reviews from the already
defined links

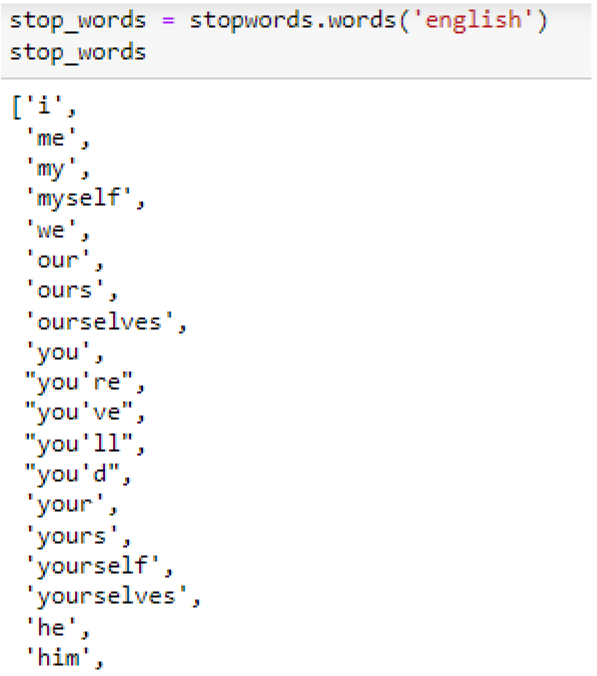


```python
reviews = []
for y in range (1,50):
    k = requests.get(url8 + str(y))
    soup = BeautifulSoup(k.text,'html.parser')
    Reviews = soup.find_all("div", {"class":"text_content"})
    H = soup.find_all("h1",{"itemprop":"name"})


    for i in range(len(Reviews)):
        reviews.append(Reviews[i].text)
        reviews

reviews_df = pd.DataFrame(np.array(reviews), columns = ['Reviews'])
reviews_df.head()
```



	Reviews
0	✅ Trip Verified | We got to the airport in Mia...
1	✅ Trip Verified | Just no. Gate agent wouldn't...
2	Not Verified | My wife and I purchased a flig...
3	✅ Trip Verified | Booked a flight to Milwauke...
4	Not Verified | American Airlines email check ...

Coding Snippets

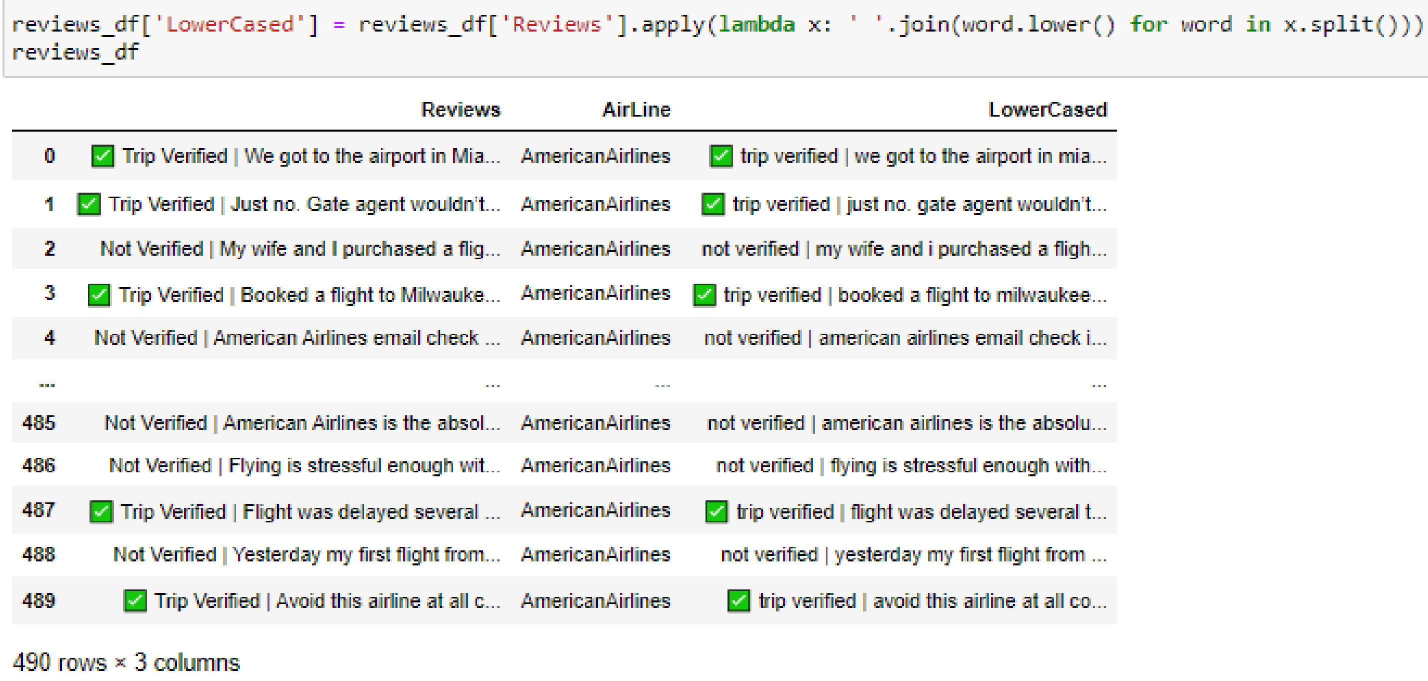


```
stop_words = stopwords.words('english')
stop_words
```

```
['i',
 'me',
 'my',
 'myself',
 'we',
 'our',
 'ours',
 'ourselves',
 'you',
 "you're",
 "you've",
 "you'll",
 "you'd",
 'your',
 'yours',
 'yourself',
 'yourselves',
 'he',
 'him',
```

Checking on stop words within the scraped data

Switching the scraped data to lower cases

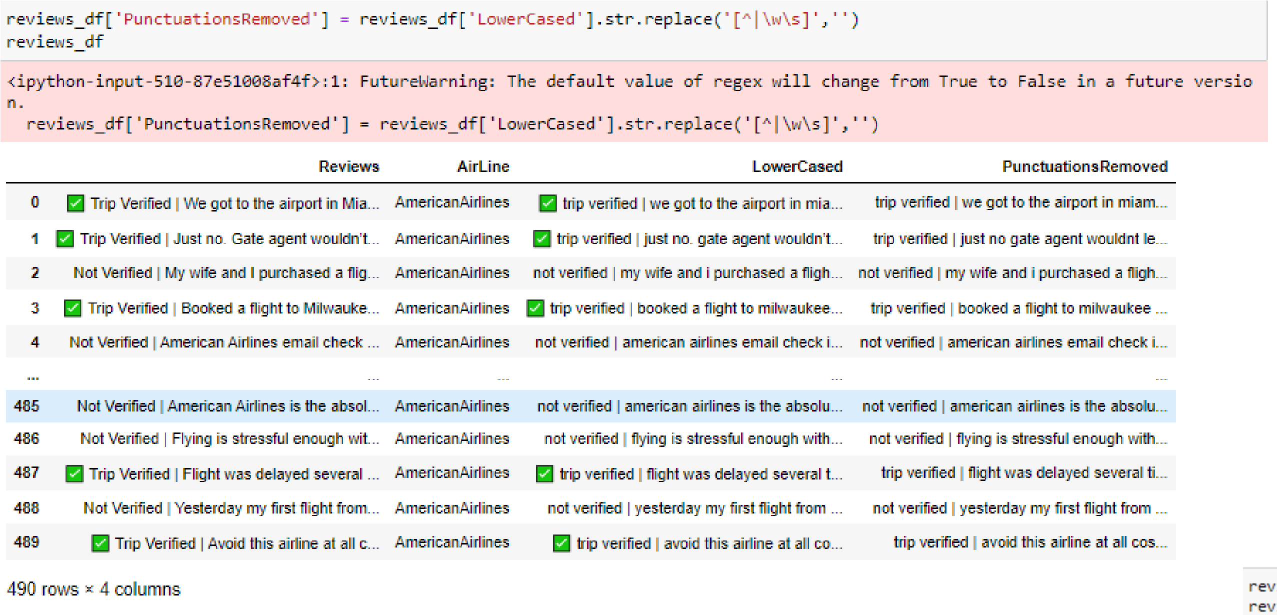


```
reviews_df['LowerCased'] = reviews_df['Reviews'].apply(lambda x: ' '.join(word.lower() for word in x.split()))
reviews_df
```

	Reviews	AirLine	LowerCased
0	✅ Trip Verified | We got to the airport in Mia...	AmericanAirlines	✅ trip verified | we got to the airport in mia...
1	✅ Trip Verified | Just no. Gate agent wouldn't...	AmericanAirlines	✅ trip verified | just no. gate agent wouldn't...
2	Not Verified | My wife and I purchased a flig...	AmericanAirlines	not verified | my wife and i purchased a fligh...
3	✅ Trip Verified | Booked a flight to Milwauke...	AmericanAirlines	✅ trip verified | booked a flight to milwaukee...
4	Not Verified | American Airlines email check ...	AmericanAirlines	not verified | american airlines email check i...
...	...	...	...
485	Not Verified | American Airlines is the absol...	AmericanAirlines	not verified | american airlines is the absolu...
486	Not Verified | Flying is stressful enough wit...	AmericanAirlines	not verified | flying is stressful enough with...
487	✅ Trip Verified | Flight was delayed several ...	AmericanAirlines	✅ trip verified | flight was delayed several t...
488	Not Verified | Yesterday my first flight from...	AmericanAirlines	not verified | yesterday my first flight from ...
489	✅ Trip Verified | Avoid this airline at all c...	AmericanAirlines	✅ trip verified | avoid this airline at all co...

490 rows × 3 columns

Coding Snippets



```
reviews_df['PunctuationsRemoved'] = reviews_df['LowerCased'].str.replace('[^|\w\s]','')
reviews_df
```

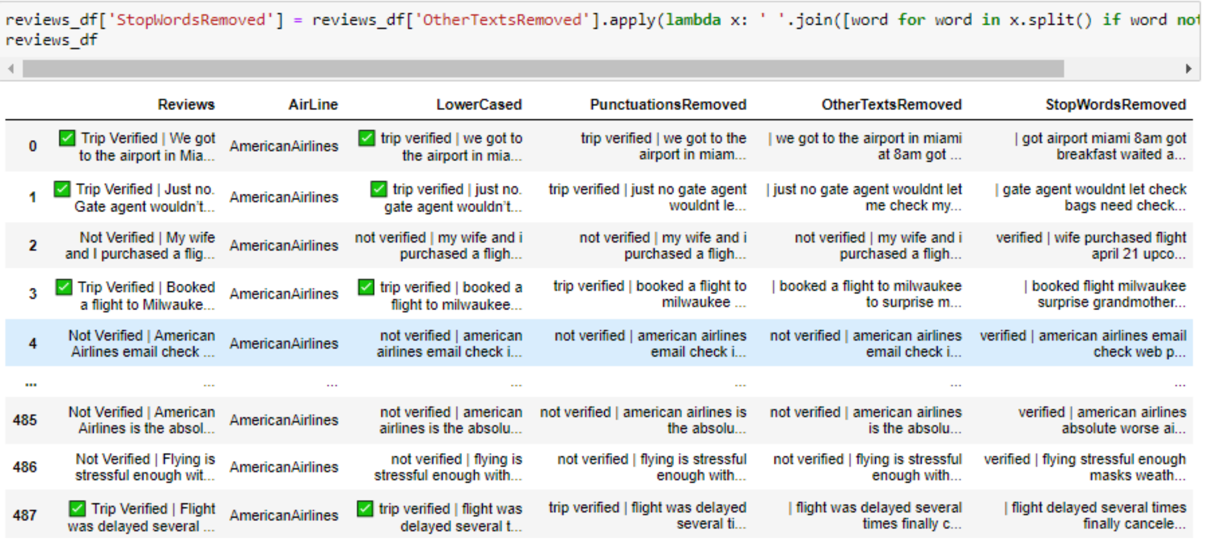
```
<ipython-input-510-87e51008af4f>:1: FutureWarning: The default value of regex will change from True to False in a future versio
n.
  reviews_df['PunctuationsRemoved'] = reviews_df['LowerCased'].str.replace('[^|\w\s]','')
```

	Reviews	AirLine	LowerCased	PunctuationsRemoved
0	✅ Trip Verified | We got to the airport in Mia...	AmericanAirlines	✅ trip verified | we got to the airport in mia...	trip verified | we got to the airport in miam...
1	✅ Trip Verified | Just no. Gate agent wouldn't...	AmericanAirlines	✅ trip verified | just no. gate agent wouldn't...	trip verified | just no gate agent wouldnt le...
2	Not Verified | My wife and I purchased a flig...	AmericanAirlines	not verified | my wife and i purchased a fligh...	not verified | my wife and i purchased a fligh...
3	✅ Trip Verified | Booked a flight to Milwauke...	AmericanAirlines	✅ trip verified | booked a flight to milwaukee...	trip verified | booked a flight to milwaukee ...
4	Not Verified | American Airlines email check ...	AmericanAirlines	not verified | american airlines email check i...	not verified | american airlines email check i...
...	...	...	...	...
485	Not Verified | American Airlines is the absol...	AmericanAirlines	not verified | american airlines is the absolu...	not verified | american airlines is the absolu...
486	Not Verified | Flying is stressful enough wit...	AmericanAirlines	not verified | flying is stressful enough with...	not verified | flying is stressful enough with...
487	✅ Trip Verified | Flight was delayed several ...	AmericanAirlines	✅ trip verified | flight was delayed several t...	trip verified | flight was delayed several ti...
488	Not Verified | Yesterday my first flight from...	AmericanAirlines	not verified | yesterday my first flight from ...	not verified | yesterday my first flight from ...
489	✅ Trip Verified | Avoid this airline at all c...	AmericanAirlines	✅ trip verified | avoid this airline at all co...	trip verified | avoid this airline at all cos...

490 rows × 4 columns

Removing all the ☑ and punctuation marks from the Lowercased data

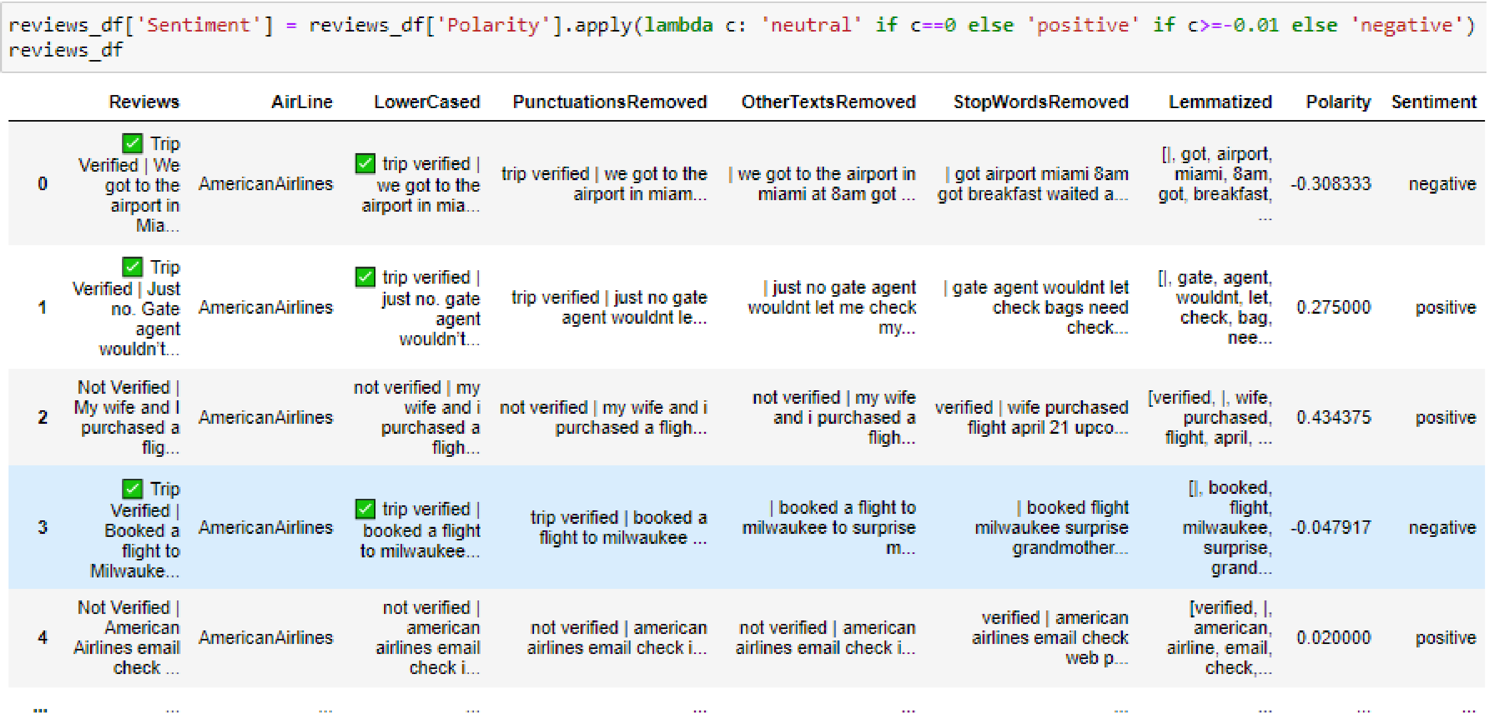
Removing all the stop words



```
reviews_df['StopWordsRemoved'] = reviews_df['OtherTextsRemoved'].apply(lambda x: ' '.join([word for word in x.split() if word not
reviews_df
```

	Reviews	AirLine	LowerCased	PunctuationsRemoved	OtherTextsRemoved	StopWordsRemoved
0	✅ Trip Verified | We got to the airport in Mia...	AmericanAirlines	✅ trip verified | we got to the airport in mia...	trip verified | we got to the airport in miam...	| we got to the airport in miami at 8am got ...	| got airport miami 8am got breakfast waited a...
1	✅ Trip Verified | Just no. Gate agent wouldn't...	AmericanAirlines	✅ trip verified | just no. gate agent wouldn't...	trip verified | just no gate agent wouldnt le...	| just no gate agent wouldnt let me check my...	| gate agent wouldnt let check bags need check...
2	Not Verified | My wife and I purchased a flig...	AmericanAirlines	not verified | my wife and i purchased a fligh...	not verified | my wife and i purchased a fligh...	not verified | my wife and i purchased a fligh...	verified | wife purchased flight april 21 upco...
3	✅ Trip Verified | Booked a flight to Milwauke...	AmericanAirlines	✅ trip verified | booked a flight to milwaukee...	trip verified | booked a flight to milwaukee ...	| booked a flight to milwaukee to surprise m...	| booked flight milwaukee surprise grandmother...
4	Not Verified | American Airlines email check ...	AmericanAirlines	not verified | american airlines email check i...	not verified | american airlines email check i...	not verified | american airlines email check i...	verified | american airlines email check web p...
...	...	...	...	...	...	...
485	Not Verified | American Airlines is the absol...	AmericanAirlines	not verified | american airlines is the absolu...	not verified | american airlines is the absolu...	not verified | american airlines is the absolu...	verified | american airlines absolute worse ai...
486	Not Verified | Flying is stressful enough wit...	AmericanAirlines	not verified | flying is stressful enough with...	not verified | flying is stressful enough with...	not verified | flying is stressful enough with...	verified | flying stressful enough masks weath...
487	✅ Trip Verified | Flight was delayed several ...	AmericanAirlines	✅ trip verified | flight was delayed several t...	trip verified | flight was delayed several ti...	| flight was delayed several times finally c...	| flight delayed several times finally cancele...

Coding Snippets

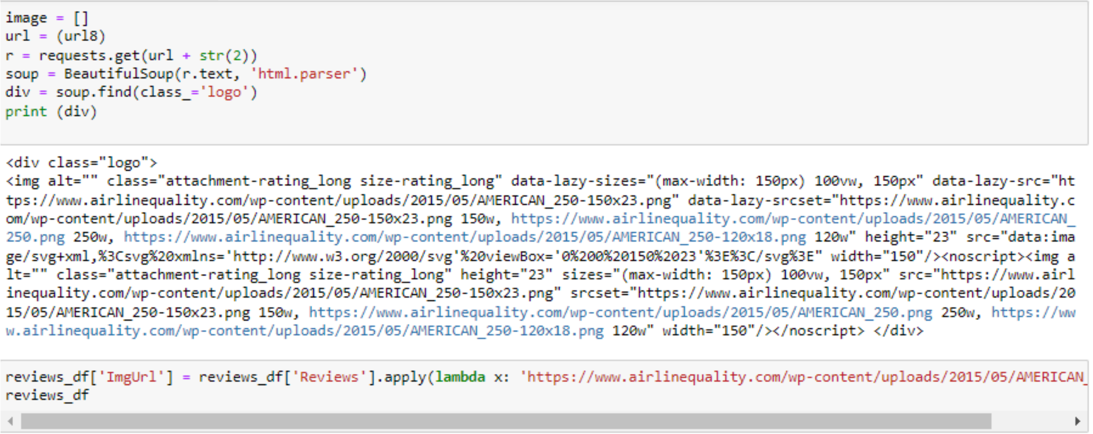


```
reviews_df['Sentiment'] = reviews_df['Polarity'].apply(lambda c: 'neutral' if c==0 else 'positive' if c>=-0.01 else 'negative')
reviews_df
```

	Reviews	AirLine	LowerCased	PunctuationsRemoved	OtherTextsRemoved	StopWordsRemoved	Lemmatized	Polarity	Sentiment
0	✅ Trip Verified | We got to the airport in Mia...	AmericanAirlines	✅ trip verified | we got to the airport in mia...	trip verified | we got to the airport in miam...	| we got to the airport in miami at 8am got ...	| got airport miami 8am got breakfast waited a...	[|, got, airport, miami, 8am, got, breakfast, ...	-0.308333	negative
1	✅ Trip Verified | Just no. Gate agent wouldn't...	AmericanAirlines	✅ trip verified | just no. gate agent wouldn't...	trip verified | just no gate agent wouldnt le...	| just no gate agent wouldnt let me check my...	| gate agent wouldnt let check bags need check...	[|, gate, agent, wouldnt, let, check, bag, nee...	0.275000	positive
2	Not Verified | My wife and I purchased a flig...	AmericanAirlines	not verified | my wife and i purchased a fligh...	not verified | my wife and i purchased a fligh...	not verified | my wife and i purchased a fligh...	verified | wife purchased flight april 21 upco...	[verified, |, wife, purchased, flight, april, ...	0.434375	positive
3	✅ Trip Verified | Booked a flight to Milwauke...	AmericanAirlines	✅ trip verified | booked a flight to milwaukee...	trip verified | booked a flight to milwaukee ...	| booked a flight to milwaukee to surprise m...	| booked flight milwaukee surprise grandmother...	[|, booked, flight, milwaukee, surprise, grand...	-0.047917	negative
4	Not Verified | American Airlines email check ...	AmericanAirlines	not verified | american airlines email check i...	not verified | american airlines email check i...	not verified | american airlines email check i...	verified | american airlines email check web p...	[verified, |, american, airline, email, check,...	0.020000	positive
...	...	...	...	...	...	...	...	...	...

Finding both Polarity & Sentiment for the reviews

Appending image URL of respective Airline into DataFrame

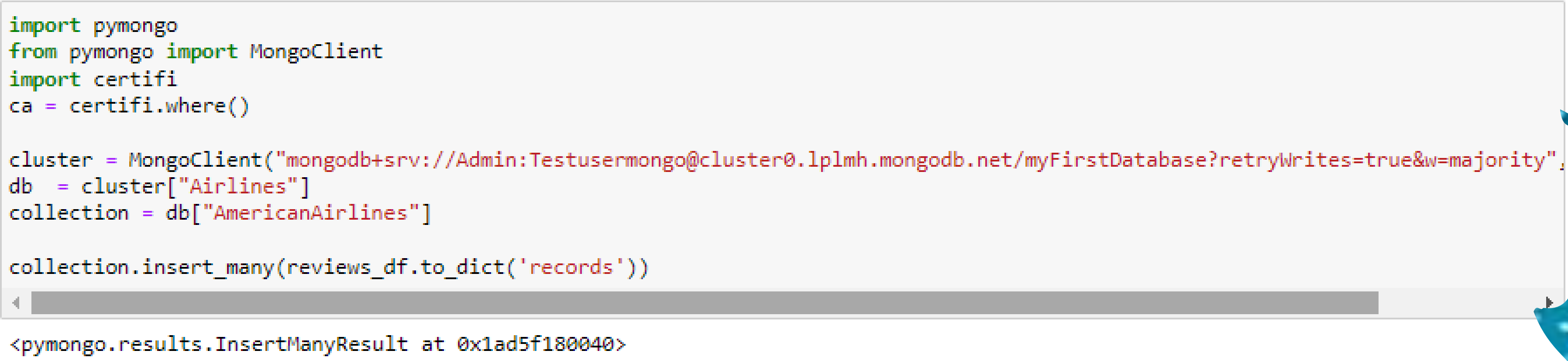


```
image = []
url = (url8)
r = requests.get(url + str(2))
soup = BeautifulSoup(r.text, 'html.parser')
div = soup.find(class_='logo')
print (div)
```

```
<div class="logo">
<img alt="" class="attachment-rating_long size-rating_long" data-lazy-sizes="(max-width: 150px) 100vw, 150px" data-lazy-src="ht
tps://www.airlinequality.com/wp-content/uploads/2015/05/AMERICAN_250-150x23.png" data-lazy-srcset="https://www.airlinequality.c
om/wp-content/uploads/2015/05/AMERICAN_250-150x23.png 150w, https://www.airlinequality.com/wp-content/uploads/2015/05/AMERICAN_
250.png 250w, https://www.airlinequality.com/wp-content/uploads/2015/05/AMERICAN_250-120x18.png 120w" height="23" src="data:ima
ge/svg+xml,%3Csvg%20xmlns='http://www.w3.org/2000/svg'%20viewBox='0%200%20150%2023'%3E%3C/svg%3E" width="150"/><noscript><img a
lt="" class="attachment-rating_long size-rating_long" height="23" sizes="(max-width: 150px) 100vw, 150px" src="https://www.airl
inequality.com/wp-content/uploads/2015/05/AMERICAN_250-150x23.png" srcset="https://www.airlinequality.com/wp-content/uploads/20
15/05/AMERICAN_250-150x23.png 150w, https://www.airlinequality.com/wp-content/uploads/2015/05/AMERICAN_250.png 250w, https://ww
w.airlinequality.com/wp-content/uploads/2015/05/AMERICAN_250-120x18.png 120w" width="150"/></noscript> </div>
```

```
reviews_df['ImgUrl'] = reviews_df['Reviews'].apply(lambda x: 'https://www.airlinequality.com/wp-content/uploads/2015/05/AMERICAN_
reviews_df
```

Coding Snippets

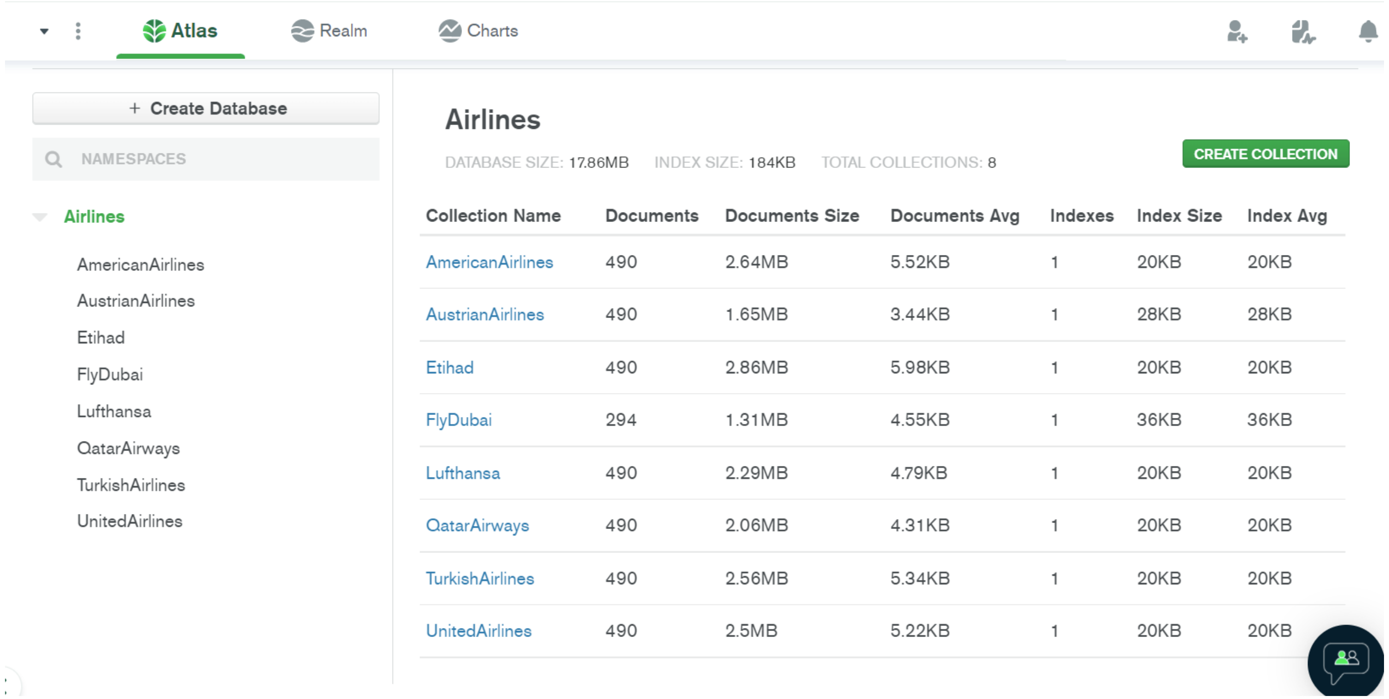


```
import pymongo
from pymongo import MongoClient
import certifi
ca = certifi.where()

cluster = MongoClient("mongodb+srv://Admin:Testusermongo@cluster0.lplmh.mongodb.net/myFirstDatabase?retryWrites=true&w=majority",
db  = cluster["Airlines"]
collection = db["AmericanAirlines"]

collection.insert_many(reviews_df.to_dict('records'))
```

```
<pymongo.results.InsertManyResult at 0x1ad5f180040>
```

Exporting DataFrame to MongoDB

Data Storage – Mongo DB



Airlines

- AmericanAirlines
- AustrianAirlines
- Etihad
- FlyDubai
- Lufthansa
- QatarAirways
- TurkishAirlines
- UnitedAirlines

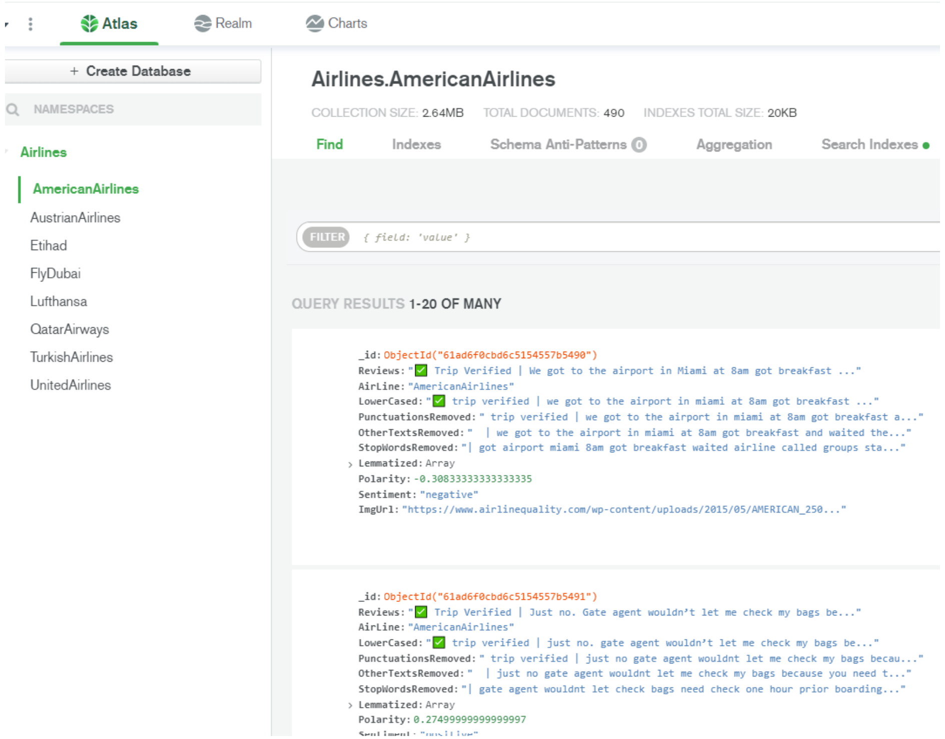
Airlines

DATABASE SIZE: 17.86MB INDEX SIZE: 184KB TOTAL COLLECTIONS: 8

Collection Name	Documents	Documents Size	Documents Avg	Indexes	Index Size	Index Avg
AmericanAirlines	490	2.64MB	5.52KB	1	20KB	20KB
AustrianAirlines	490	1.65MB	3.44KB	1	28KB	28KB
Etihad	490	2.86MB	5.98KB	1	20KB	20KB
FlyDubai	294	1.31MB	4.55KB	1	36KB	36KB
Lufthansa	490	2.29MB	4.79KB	1	20KB	20KB
QatarAirways	490	2.06MB	4.31KB	1	20KB	20KB
TurkishAirlines	490	2.56MB	5.34KB	1	20KB	20KB
UnitedAirlines	490	2.5MB	5.22KB	1	20KB	20KB

Total of 08 Airline's reviews

Preview of the data stored in a collection



Airlines

- AmericanAirlines
- AustrianAirlines
- Etihad
- FlyDubai
- Lufthansa
- QatarAirways
- TurkishAirlines
- UnitedAirlines

Airlines.AmericanAirlines

COLLECTION SIZE: 2.64MB TOTAL DOCUMENTS: 490 INDEXES TOTAL SIZE: 20KB

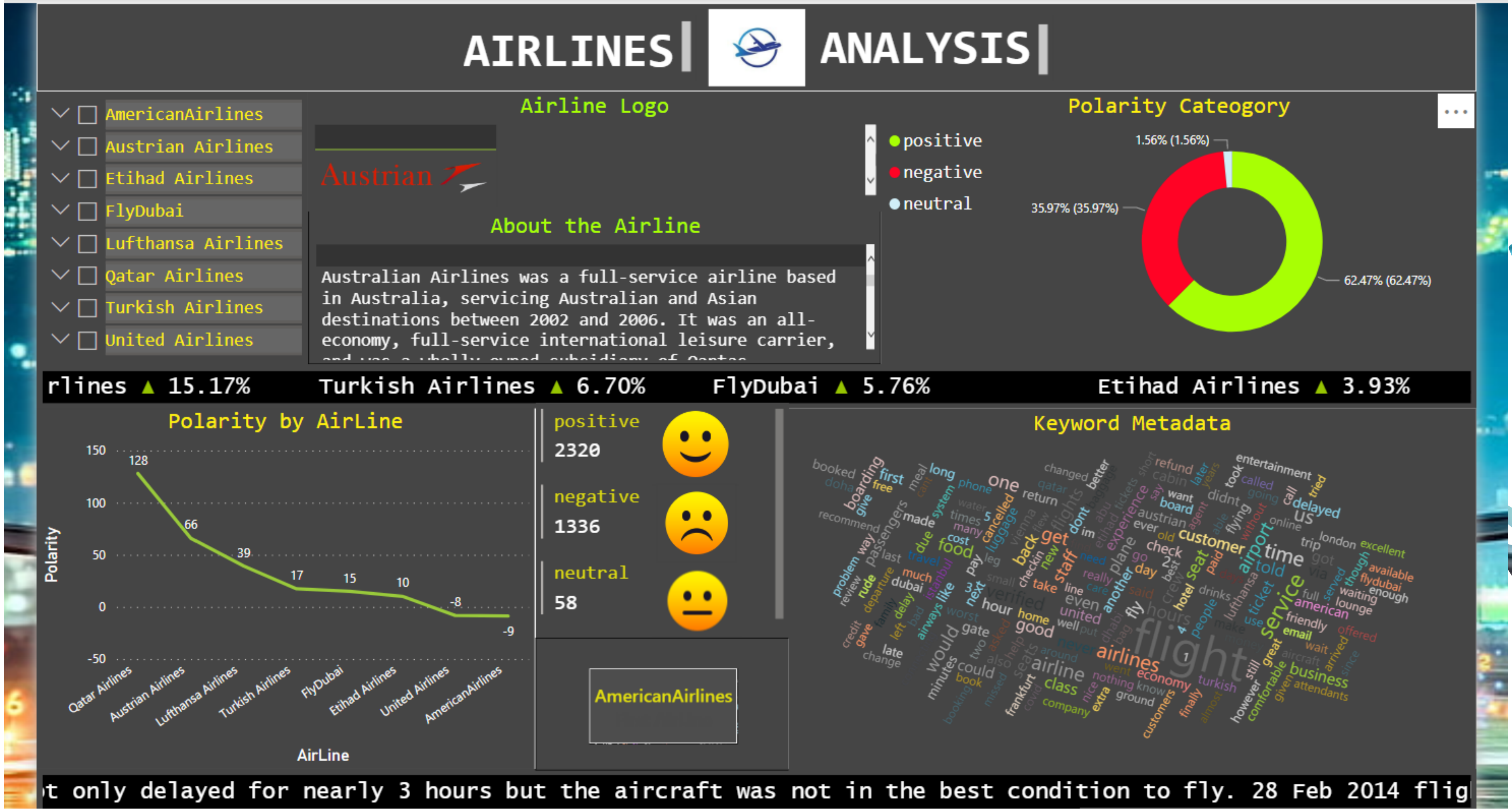
{ field: 'value' }

QUERY RESULTS 1-20 OF MANY

```
_id: ObjectId("61ad6f0cbd6c5154557b5490")
Reviews: "✅ Trip Verified | We got to the airport in Miami at 8am got breakfast ..."
AirLine: "AmericanAirlines"
LowerCased: "✅ trip verified | we got to the airport in miami at 8am got breakfast ..."
PunctuationsRemoved: " trip verified | we got to the airport in miami at 8am got breakfast a..."
OtherTextsRemoved: "  | we got to the airport in miami at 8am got breakfast and waited the..."
StopWordsRemoved: "| got airport miami 8am got breakfast waited airline called groups sta..."
> Lemmatized: Array
Polarity: -0.30833333333333335
Sentiment: "negative"
ImgUrl: "https://www.airlinequality.com/wp-content/uploads/2015/05/AMERICAN_250..."
```

```
_id: ObjectId("61ad6f0cbd6c5154557b5491")
Reviews: "✅ Trip Verified | Just no. Gate agent wouldn’t let me check my bags be..."
AirLine: "AmericanAirlines"
LowerCased: "✅ trip verified | just no. gate agent wouldn’t let me check my bags be..."
PunctuationsRemoved: " trip verified | just no gate agent wouldnt let me check my bags becau..."
OtherTextsRemoved: "  | just no gate agent wouldnt let me check my bags because you need t..."
StopWordsRemoved: "| gate agent wouldnt let check bags need check one hour prior boarding..."
> Lemmatized: Array
Polarity: 0.27499999999999997
Sentiment: "positive"
```

Power BI – Visualization



AIRLINES | ANALYSIS |

- □ AmericanAirlines
- □ Austrian Airlines
- □ Etihad Airlines
- □ FlyDubai
- □ Lufthansa Airlines
- □ Qatar Airlines
- □ Turkish Airlines
- □ United Airlines

Airline Logo

Austrian

About the Airline

Australian Airlines was a full-service airline based in Australia, servicing Australian and Asian destinations between 2002 and 2006. It was an all-economy, full-service international leisure carrier, and was a wholly owned subsidiary of Qantas

Polarity Cateogory

- ● positive
- ● negative
- ● neutral

rlines ▲ 15.17%   Turkish Airlines ▲ 6.70%   FlyDubai ▲ 5.76%   Etihad Airlines ▲ 3.93%

Polarity by AirLine

positive	2320
negative	1336
neutral	58

AmericanAirlines

Keyword Metadata

t only delayed for nearly 3 hours but the aircraft was not in the best condition to fly. 28 Feb 2014 flig

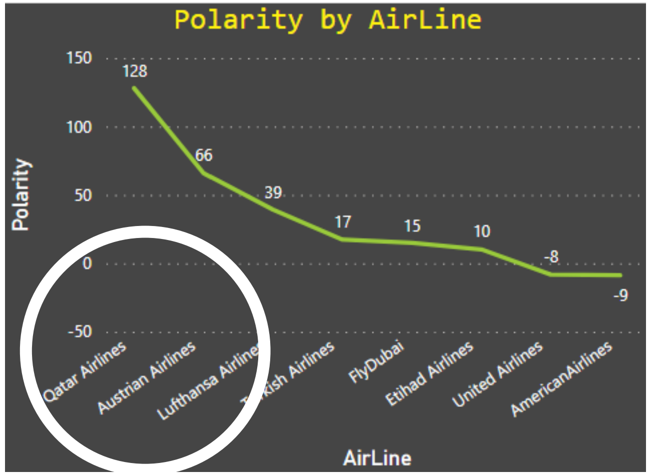
Conclusion

- Qatar Airways has been on the customer’s favorite list out of the ones that we choose to analyse, having showed a 432 positive reviews with just 50 negative reviews.

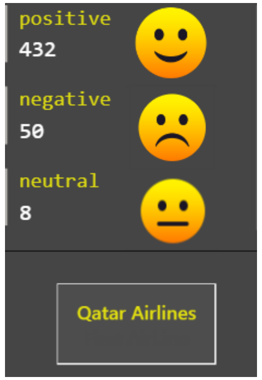
- Opposingly American Airlines, as per our analysis, ranked as the least preferred airline with much negative reviews totalling to 250 negative (51%) reviews from customers. However said airline have also shown a positive polarity count of 230 as well.

- Turkish Airline, Flydubai & Etihad is being on the customers positive list, and are among the similar preference list as it shows a 50-58% of positive sentiment

- Excellent, Business, Service, good and service are some of the key words we have identified from the keyword metadata expressing that it was the customers favourite choice and worst, never, covid are few frequent words identified within reviews from American airline showing that it’s most negative sentiment is best verified.



Top 03 Airlines derived from this analysis



positive
432 🙂

negative
50 ☹

neutral
8 😐

Qatar Airlines

Thank you