ba-task1

April 13, 2023

1 Task 1

1.1 Web scraping and analysis

This Jupyter notebook includes some code to get you started with web scraping. We will use a package called `BeautifulSoup` to collect the data from the web. Once you've collected your data and saved it into a local `.csv` file you should start with your analysis.

1.1.1 Scraping data from Skytrax

If you visit [https://www.airlinequality.com] you can see that there is a lot of data there. For this task, we are only interested in reviews related to British Airways and the Airline itself.

If you navigate to this link: [https://www.airlinequality.com/airline-reviews/british-airways] you will see this data. Now, we can use `Python` and `BeautifulSoup` to collect all the links to the reviews and then to collect the text data on each of the individual review links.

[2]:

```python
import requests
from bs4 import BeautifulSoup
import pandas as pd
```

[3]:

```python
base_url = "https://www.airlinequality.com/airline-reviews/british-airways"
pages = 10
page_size = 100

reviews = []

# for i in range(1, pages + 1):
for i in range(1, pages + 1):

    print(f"Scraping page {i}")

    # Create URL to collect links from paginated data
    url = f"{base_url}/page/{i}/?sortby=post_date%3ADesc&pagesize={page_size}"

    # Collect HTML data from this page
    response = requests.get(url)
    # Parse content
    content = response.content
    parsed_content = BeautifulSoup(content, 'html.parser')
    for para in parsed_content.find_all("div", {"class": "text_content"}):
        reviews.append(para.get_text())

    print(f"   ---> {len(reviews)} total reviews")
```

```
Scraping page 1
   ---> 100 total reviews
Scraping page 2
   ---> 200 total reviews
Scraping page 3
   ---> 300 total reviews
Scraping page 4
   ---> 400 total reviews
Scraping page 5
   ---> 500 total reviews
Scraping page 6
   ---> 600 total reviews
Scraping page 7
   ---> 700 total reviews
Scraping page 8
   ---> 800 total reviews
Scraping page 9
   ---> 900 total reviews
Scraping page 10
   ---> 1000 total reviews
```

[4]:
```
df = pd.DataFrame()
df["reviews"] = reviews
df.head()
```

[4]:
```
                                             reviews
0   Trip Verified | BA 242 on the 6/2/23. Boardi…
1   Trip Verified |  Not only my first flight in…
2   Trip Verified |  My husband and myself were …
3   Trip Verified | Organised boarding process. …
4   Trip Verified |  Outward journey BA245 Londo…
```

[5]:
```
df.to_csv("BA_reviews.csv")
```

Congratulations! Now you have your dataset for this task! The loops above collected 1000 reviews by iterating through the paginated pages on the website. However, if you want to collect more data, try increasing the number of pages!

The next thing that you should do is clean this data to remove any unnecessary text from each of the rows. For example, “ Trip Verified” can be removed from each row if it exists, as it’s not relevant to what we want to investigate.

[6]:

```
df.reviews= df.reviews.str.split('|',expand=True)[1]
```

[7]:

```
df
```

[7]:

```
                                                 reviews
0     BA 242 on the 6/2/23. Boarding was delayed du…
1      Not only my first flight in 17 years, but al…
2      My husband and myself were flying to Madrid …
3     Organised boarding process. Really friendly c…
4      Outward journey BA245 London to Buenos Aires…
..                                                 …
995   Madrid to London. Credit where it's due. Flew…
996   Venice to Gatwick. I use Snokart luggage whic…
997   First 3 legs were trouble free. Lounges were …
998   Flew London to Budapest with British Airways…
999    Paid for a Vueling Airlines flight from Flor…

[1000 rows x 1 columns]
```

[8]:

```
import re

# Define a function to clean the text
def clean(text):
# Removes all special characters and numericals leaving the alphabets
    text = re.sub('[^A-Za-z]+', ' ', str(text))
    return text

# Cleaning the text in the review column
df['Cleaned Reviews'] = df['reviews'].apply(clean)
df.head()
```

[8]:

```
                                             reviews  \
0   BA 242 on the 6/2/23. Boarding was delayed du…
1    Not only my first flight in 17 years, but al…
2    My husband and myself were flying to Madrid …
3   Organised boarding process. Really friendly c…
4    Outward journey BA245 London to Buenos Aires…

                                     Cleaned Reviews
0   BA on the Boarding was delayed due to late ar…
1   Not only my first flight in years but also my…
2   My husband and myself were flying to Madrid o…
3   Organised boarding process Really friendly cr…
4   Outward journey BA London to Buenos Aires Clu…
```

[10]:

```
import nltk

"""This punkt tokenizer divides a text into a list of sentences by using an␣
 ↪unsupervised algorithm to build a model for abbreviation words,
collocations, and words that start sentences. """

nltk.download('punkt')
from nltk.tokenize import word_tokenize
from nltk import pos_tag
nltk.download('stopwords')
from nltk.corpus import stopwords
nltk.download('wordnet')
from nltk.corpus import wordnet
```

```
[nltk_data] Error loading punkt: <urlopen error [WinError 10060] A
[nltk_data]     connection attempt failed because the connected party
[nltk_data]     did not properly respond after a period of time, or
[nltk_data]     established connection failed because connected host
[nltk_data]     has failed to respond>
[nltk_data] Error loading stopwords: <urlopen error [WinError 10060] A
[nltk_data]     connection attempt failed because the connected party
[nltk_data]     did not properly respond after a period of time, or
[nltk_data]     established connection failed because connected host
[nltk_data]     has failed to respond>
[nltk_data] Error loading wordnet: <urlopen error [WinError 10060] A
[nltk_data]     connection attempt failed because the connected party
[nltk_data]     did not properly respond after a period of time, or
[nltk_data]     established connection failed because connected host
[nltk_data]     has failed to respond>
```

[11]:

```
nltk.download('omw-1.4')
nltk.download('averaged_perceptron_tagger')

# POS tagger dictionary
pos_dict = {'J':wordnet.ADJ, 'V':wordnet.VERB, 'N':wordnet.NOUN, 'R':wordnet.
 ↪ADV}
def token_stop_pos(text):
    tags = pos_tag(word_tokenize(text))
    #print(tags)
    newlist = []
    for word, tag in tags:
        if word.lower() not in set(stopwords.words('english')):
          newlist.append(tuple([word, pos_dict.get(tag[0])]))
          #print(tag[0])
          #print(pos_dict.get(tag[0]))
    return newlist
```

```
df['POS tagged'] = df['Cleaned Reviews'].apply(token_stop_pos)
df.head(50)
```

```
[nltk_data] Error loading omw-1.4: <urlopen error [WinError 10060] A
[nltk_data]     connection attempt failed because the connected party
[nltk_data]     did not properly respond after a period of time, or
[nltk_data]     established connection failed because connected host
[nltk_data]     has failed to respond>
[nltk_data] Error loading averaged_perceptron_tagger: <urlopen error
[nltk_data]     [WinError 10060] A connection attempt failed because
[nltk_data]     the connected party did not properly respond after a
[nltk_data]     period of time, or established connection failed
[nltk_data]     because connected host has failed to respond>
```

[11]:

```
                                              reviews  \
0    BA 242 on the 6/2/23. Boarding was delayed du…
1     Not only my first flight in 17 years, but al…
2     My husband and myself were flying to Madrid …
3    Organised boarding process. Really friendly c…
4     Outward journey BA245 London to Buenos Aires…
5    Check in agent at LHR was very helpful and fr…
6     Very disappointing. I book BA so I can fly d…
7    Excellent service both on the ground and on b…
8     Good lounge at Cape Town. On time departure…
9     A really excellent journey. Lounge not overc…
10    This flight was one of the worst I have ever…
11   It seems that there is a race to the bottom a…
12    As a Spanish born individual living in Engla…
13    A rather empty and quiet flight to Tel Aviv,…
14    Easy check in and staff member was polite an…
15    Being a silver flyer and booking a flight th…
16    I find BA incredibly tacky and constantly lo…
17    Flew ATL to LHR 8th Jan 2023. Was unlucky en…
18    Great thing about British Airways A380 is th…
19   The staff are friendly. The plane was cold, w…
20   Probably the worst business class experience …
21   Definitely not recommended, especially for bu…
22    BA shuttle service across the UK is still su…
23   I must admit like many others I tend to avoid…
24    When will BA update their Business class cab…
25    Paid £200 day before flight for an upgrade f…
26    BA website did not work (weirdly deleted my …
27    Absolutely terrible experience with British …
28    Vancouver to Delhi via London. We were booke…
29    Old A320 with narrow pitch. Flight perfectly…
30    Another BA Shambles. Started off well with e…
31    BA cancelled my flight home to Heathrow on D…
```

```
32    BA cancelled my flight home, the last flight…
33   Turned up 3.5 hours in advance, Terminal 5 at…
34    Boarding - at gate at LGW they called Group …
35    Missing baggage customer service was the wor…
36    British Airways are not the flag carrier the…
37    Stupidly tried BA again after a five year ga…
38    Seat horribly narrow; 3-4-3 on a 777. Thankf…
39    Glasgow to London delayed by 1 hour. My wife…
40    When I tried to check in online, I was offer…
41    I flew from Prague to LHR. Excellent service…
42    Disappointing again especially on business. …
43    During both the outbound and return flights …
44    I was flying to Warsaw for one day of meetin…
45    Booked a BA holiday to Marrakech, after post…
46   Extremely sub-par service. Highlights: No onl…
47    I virtually gave up on British Airways about…
48    I was pleasantly surprised that the airline …
49    British Airways is late, their website is at…

                                      Cleaned Reviews  \
0    BA on the Boarding was delayed due to late ar…
1    Not only my first flight in years but also my…
2    My husband and myself were flying to Madrid o…
3    Organised boarding process Really friendly cr…
4    Outward journey BA London to Buenos Aires Clu…
5    Check in agent at LHR was very helpful and fr…
6    Very disappointing I book BA so I can fly dur…
7    Excellent service both on the ground and on b…
8    Good lounge at Cape Town On time departure Dr…
9    A really excellent journey Lounge not overcro…
10   This flight was one of the worst I have ever …
11   It seems that there is a race to the bottom a…
12   As a Spanish born individual living in Englan…
13   A rather empty and quiet flight to Tel Aviv v…
14   Easy check in and staff member was polite and…
15   Being a silver flyer and booking a flight thr…
16   I find BA incredibly tacky and constantly loo…
17   Flew ATL to LHR th Jan Was unlucky enough to …
18   Great thing about British Airways A is the ec…
19   The staff are friendly The plane was cold we …
20   Probably the worst business class experience …
21   Definitely not recommended especially for bus…
22   BA shuttle service across the UK is still sur…
23   I must admit like many others I tend to avoid…
24   When will BA update their Business class cabi…
25   Paid day before flight for an upgrade from ec…
26   BA website did not work weirdly deleted my fl…
```

```
27   Absolutely terrible experience with British A…
28   Vancouver to Delhi via London We were booked …
29   Old A with narrow pitch Flight perfectly on t…
30   Another BA Shambles Started off well with exc…
31   BA cancelled my flight home to Heathrow on De…
32   BA cancelled my flight home the last flight o…
33   Turned up hours in advance Terminal at London…
34   Boarding at gate at LGW they called Group to …
35   Missing baggage customer service was the wors…
36   British Airways are not the flag carrier they…
37   Stupidly tried BA again after a five year gap…
38   Seat horribly narrow on a Thankfully the flig…
39   Glasgow to London delayed by hour My wife and…
40   When I tried to check in online I was offered…
41   I flew from Prague to LHR Excellent service a…
42   Disappointing again especially on business Th…
43   During both the outbound and return flights w…
44   I was flying to Warsaw for one day of meeting…
45   Booked a BA holiday to Marrakech after postin…
46   Extremely sub par service Highlights No onlin…
47   I virtually gave up on British Airways about …
48   I was pleasantly surprised that the airline c…
49   British Airways is late their website is atro…

                                             POS tagged
0   [(BA, n), (Boarding, n), (delayed, v), (due, a…
1   [(first, a), (flight, n), (years, n), (also, r…
2   [(husband, n), (flying, v), (Madrid, n), (rd, …
3   [(Organised, v), (boarding, v), (process, n), …
4   [(Outward, n), (journey, n), (BA, n), (London,…
5   [(Check, n), (agent, n), (LHR, n), (helpful, a…
6   [(disappointing, a), (book, n), (BA, n), (fly,…
7   [(Excellent, a), (service, n), (ground, n), (b…
8   [(Good, a), (lounge, n), (Cape, n), (Town, n),…
9   [(really, r), (excellent, a), (journey, n), (L…
10  [(flight, n), (one, None), (worst, a), (ever, …
11  [(seems, v), (race, n), (bottom, a), (amongst,…
12  [(Spanish, a), (born, v), (individual, a), (li…
13  [(rather, r), (empty, a), (quiet, a), (flight,…
14  [(Easy, a), (check, n), (staff, n), (member, n…
15  [(silver, n), (flyer, n), (booking, v), (fligh…
16  [(find, v), (BA, a), (incredibly, r), (tacky, …
17  [(Flew, n), (ATL, n), (LHR, n), (th, n), (Jan,…
18  [(Great, a), (thing, n), (British, n), (Airway…
19  [(staff, n), (friendly, r), (plane, n), (cold,…
20  [(Probably, r), (worst, a), (business, n), (cl…
21  [(Definitely, r), (recommended, v), (especiall…
22    [(BA, n), (shuttle, a), (service, n), (across,…
23    [(must, None), (admit, v), (like, None), (many…
24    [(BA, v), (update, v), (Business, n), (class, …
25    [(Paid, n), (day, n), (flight, n), (upgrade, n…
26    [(BA, n), (website, n), (work, v), (weirdly, r…
27    [(Absolutely, r), (terrible, a), (experience, …
28    [(Vancouver, n), (Delhi, n), (via, None), (Lon…
29    [(Old, n), (narrow, a), (pitch, n), (Flight, n…
30    [(Another, None), (BA, n), (Shambles, n), (Sta…
31    [(BA, n), (cancelled, v), (flight, n), (home, …
32    [(BA, n), (cancelled, v), (flight, n), (home, …
33    [(Turned, v), (hours, n), (advance, a), (Termi…
34    [(Boarding, v), (gate, n), (LGW, n), (called, …
35    [(Missing, v), (baggage, n), (customer, n), (s…
36    [(British, a), (Airways, n), (flag, n), (carri…
37    [(Stupidly, r), (tried, v), (BA, n), (five, No…
38    [(Seat, n), (horribly, r), (narrow, a), (Thank…
39    [(Glasgow, n), (London, n), (delayed, v), (hou…
40    [(tried, v), (check, v), (online, n), (offered…
41    [(flew, v), (Prague, n), (LHR, n), (Excellent,…
42    [(Disappointing, v), (especially, r), (busines…
43    [(outbound, n), (return, v), (flights, n), (of…
44    [(flying, v), (Warsaw, n), (one, None), (day, …
45    [(Booked, v), (BA, n), (holiday, n), (Marrakec…
46    [(Extremely, r), (sub, a), (par, n), (service,…
47    [(virtually, r), (gave, v), (British, n), (Air…
48    [(pleasantly, r), (surprised, v), (airline, n)…
49    [(British, a), (Airways, n), (late, a), (websi…
```

[12]:

```python
# Obtaining the stem words – Lemmatization

from nltk.stem import WordNetLemmatizer
wordnet_lemmatizer = WordNetLemmatizer()
def lemmatize(pos_data):
    lemma_rew = " "
    for word, pos in pos_data:
     if not pos:
        lemma = word
        lemma_rew = lemma_rew + " " + lemma
     else:
        lemma = wordnet_lemmatizer.lemmatize(word, pos=pos)
        lemma_rew = lemma_rew + " " + lemma
    return lemma_rew

df['Lemma'] = df['POS tagged'].apply(lemmatize)
df.head()
```

[12]:

```
                                             reviews  \
0  BA 242 on the 6/2/23. Boarding was delayed du…
1   Not only my first flight in 17 years, but al…
2   My husband and myself were flying to Madrid …
3  Organised boarding process. Really friendly c…
4   Outward journey BA245 London to Buenos Aires…

                                     Cleaned Reviews  \
0  BA on the Boarding was delayed due to late ar…
1  Not only my first flight in years but also my…
2  My husband and myself were flying to Madrid o…
3  Organised boarding process Really friendly cr…
4  Outward journey BA London to Buenos Aires Clu…

                                          POS tagged  \
0  [(BA, n), (Boarding, n), (delayed, v), (due, a…
1  [(first, a), (flight, n), (years, n), (also, r…
2  [(husband, n), (flying, v), (Madrid, n), (rd, …
3  [(Organised, v), (boarding, v), (process, n), …
4  [(Outward, n), (journey, n), (BA, n), (London,…

                                               Lemma
0   BA Boarding delay due late arrival incoming …
1   first flight year also first time back Engla…
2   husband fly Madrid rd February Legal matter …
3   Organised board process Really friendly crew…
4   Outward journey BA London Buenos Aires Club …
```

[ ]:

```

```

[13]:

```python
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
analyzer = SentimentIntensityAnalyzer()


# function to calculate vader sentiment
def vadersentimentanalysis(review):
    vs = analyzer.polarity_scores(review)
    return vs['compound']

df['Sentiment'] = df['Lemma'].apply(vadersentimentanalysis)

# function to analyse
def vader_analysis(compound):
    if compound >= 0.5:
        return 'Positive'
    elif compound < 0 :
        return 'Negative'
    else:
        return 'Neutral'
df['Analysis'] = df['Sentiment'].apply(vader_analysis)
df.head()
```

[13]:

```
                                             reviews  \
0  BA 242 on the 6/2/23. Boarding was delayed du…
1   Not only my first flight in 17 years, but al…
2   My husband and myself were flying to Madrid …
3  Organised boarding process. Really friendly c…
4   Outward journey BA245 London to Buenos Aires…

                                     Cleaned Reviews  \
0  BA on the Boarding was delayed due to late ar…
1  Not only my first flight in years but also my…
2  My husband and myself were flying to Madrid o…
3  Organised boarding process Really friendly cr…
4  Outward journey BA London to Buenos Aires Clu…

                                          POS tagged  \
0 [(BA, n), (Boarding, n), (delayed, v), (due, a…
1 [(first, a), (flight, n), (years, n), (also, r…
2 [(husband, n), (flying, v), (Madrid, n), (rd, …
3 [(Organised, v), (boarding, v), (process, n), …
4 [(Outward, n), (journey, n), (BA, n), (London,…

                                               Lemma  Sentiment  Analysis
0   BA Boarding delay due late arrival incoming …     0.9493  Positive
1   first flight year also first time back Engla…     0.9869  Positive
2   husband fly Madrid rd February Legal matter …     0.9799  Positive
3   Organised board process Really friendly crew…     0.9371  Positive
4   Outward journey BA London Buenos Aires Club …    -0.1119  Negative
```

[14]:

```
vader_counts = df['Analysis'].value_counts()
vader_counts
```

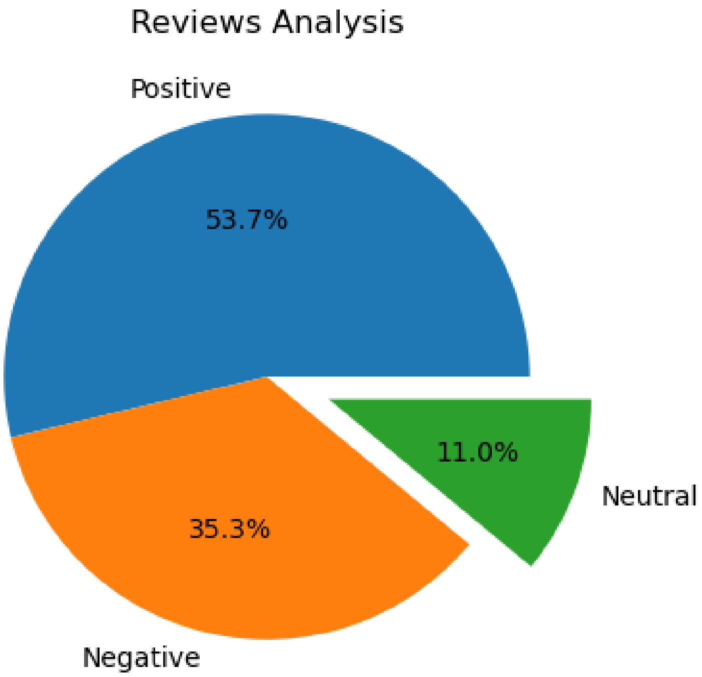
[14]:

```
Positive    537
Negative    353
Neutral     110
Name: Analysis, dtype: int64
```

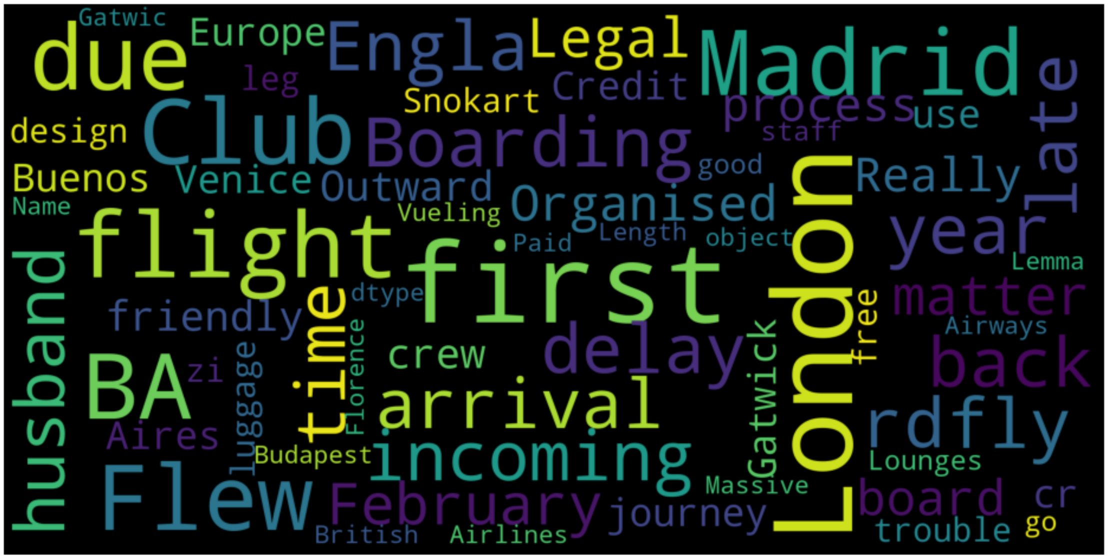
[15]:

```
import matplotlib.pyplot as plt
%matplotlib inline
plt.figure(figsize=(15,7))

plt.subplot(1,3,2)
plt.title("Reviews Analysis")
plt.pie(vader_counts.values, labels = vader_counts.index, explode = (0, 0, 0.
 ↪25), autopct='%1.1f%%', shadow=False)
```

```
[15]: ([<matplotlib.patches.Wedge at 0x178f87e95d0>,
        <matplotlib.patches.Wedge at 0x178f87e9bd0>,
        <matplotlib.patches.Wedge at 0x178f87ea2f0>],
       [Text(-0.12757508092656847, 1.0925770447554624, 'Positive'),
        Text(-0.25006438374722234, -1.071199236361342, 'Negative'),
        Text(1.2701889961293427, -0.45729630887634853, 'Neutral')],
       [Text(-0.06958640777812825, 0.5959511153211613, '53.7%'),
        Text(-0.13639875477121216, -0.584290492560732, '35.3%'),
        Text(0.799748627192549, -0.287927305588812, '11.0%')])
```



```
[22]: from wordcloud import WordCloud, STOPWORDS
      stopwords = set(STOPWORDS)

      def show_wordcloud(data):
          wordcloud = WordCloud(
              background_color='black',
              stopwords=stopwords,
              max_words=200,
              max_font_size=40,
              min_font_size = 1,
              scale=5,
              random_state=5)

          wordcloud=wordcloud.generate(str(data))

          fig = plt.figure(1, figsize=(12, 12))
          plt.axis('off')

          plt.imshow(wordcloud)
          plt.show()

      show_wordcloud(df.Lemma)
```



```
[ ]:
```