+ Code + Text

```
import pandas as pd
import matplotlib.pyplot as plt
```

Verified or not a verified trip?

```
data = pd.read_csv("/content/BA_reviews.csv")
data_copy = data.copy()
```

```
data_copy.rename(columns={"Unnamed: 0": "ID"}, inplace = True)
```

```
data_copy.set_index("ID", inplace = True)
```

```
data_copy
```

ID	reviews
0	✅ Trip Verified | I virtually gave up on Brit...
1	✅ Trip Verified | I was pleasantly surprised ...
2	✅ Trip Verified | British Airways is late, th...
3	✅ Trip Verified | Flew from Amman to London on...
4	✅ Trip Verified | This is the worst experience...
...	...
995	Two regular an uneventful flights. Curiously e...
996	✅ Trip Verified | London to Belfast. Another r...
997	✅ Trip Verified | Very full flight on G-BNLP/B...
998	✅ Trip Verified | Warsaw to London. WAW is not...
999	✅ Trip Verified | I booked my flight with Cat...

1000 rows × 1 columns

```
data_copy.isnull().sum()
```

```
reviews    0
dtype: int64
```

```
data_copy.iloc[0]
```

```
reviews    ✅ Trip Verified |  I virtually gave up on Brit...
Name: 0, dtype: object
```

```
data_copy.iloc[995]
```

```
reviews    Two regular an uneventful flights. Curiously e...
Name: 995, dtype: object
```

```
#reviews with validate trips (from and to UK)
data_copy[data_copy['reviews'].str.contains('Trip Verified')]
```

ID	reviews
0	✅ Trip Verified | I virtually gave up on Brit...
1	✅ Trip Verified | I was pleasantly surprised ...
2	✅ Trip Verified | British Airways is late, th...
3	✅ Trip Verified | Flew from Amman to London on...
4	✅ Trip Verified | This is the worst experience...
...	...
994	✅ Trip Verified | Flew London Heathrow to Delh...
996	✅ Trip Verified | London to Belfast. Another r...
997	✅ Trip Verified | Very full flight on G-BNLP/B...
998	✅ Trip Verified | Warsaw to London. WAW is not...
999	✅ Trip Verified | I booked my flight with Cat...

838 rows × 1 columns

```
#reviews with non validate flights(from and to UK)
data_copy[~data_copy['reviews'].str.contains('Trip Verified')]
```

ID	reviews
6	Not Verified | Worst experience ever. Outbound...
12	Not Verified | I've generally been a loyal Go...
21	Not Verified | LHR-LAX. I prefer the Boeing 7...
24	Not Verified | London to Cairo. First, on this...
27	Not Verified | DFW-LHR. Had an easy transfer ...
...	...
830	Not Verified | London Heathrow to Düsseldorf....
898	Not Verified | Los Angeles to London. I booke...
899	Not Verified | The overall flight wasn't too ...
903	Not Verified | First time flying with British...
995	Two regular an uneventful flights. Curiously e...

162 rows × 1 columns

```
data_copy.iloc[995]
```

```
reviews    Two regular an uneventful flights. Curiously e...
Name: 995, dtype: object
```

```
data_copy['reviews_new'] = data_copy['reviews'].str.split('|').str[0]
```

```
data_copy
```

	reviews	reviews_new
ID		
0	✅ Trip Verified | I virtually gave up on Brit...	✅ Trip Verified
1	✅ Trip Verified | I was pleasantly surprised ...	✅ Trip Verified
2	✅ Trip Verified | British Airways is late, th...	✅ Trip Verified
3	✅ Trip Verified | Flew from Amman to London on...	✅ Trip Verified
4	✅ Trip Verified | This is the worst experience...	✅ Trip Verified
...	...	...
995	Two regular an uneventful flights. Curiously e...	Two regular an uneventful flights. Curiously e...
996	✅ Trip Verified | London to Belfast. Another r...	✅ Trip Verified
997	✅ Trip Verified | Very full flight on G-BNLP/B...	✅ Trip Verified
998	✅ Trip Verified | Warsaw to London. WAW is not...	✅ Trip Verified
999	✅ Trip Verified | I booked my flight with Cat...	✅ Trip Verified

1000 rows × 2 columns

```
data_copy.reviews_new.unique()
```

```
array(['✅ Trip Verified ', 'Not Verified ', '❎ Not Verified ',
       'Two regular an uneventful flights. Curiously enough, though, with the exact same crew! The crew were
very nice and the service is very attentive and polite, but I just cannot take it that British Airways has
chosen to provide a service just like low cost carriers, where everything is charged for, apart from luggage.
On the second leg of the trip, the aircraft felt extremely warm and for some odd reason, row 6 where I was
sitting did not have AC.'],
      dtype=object)
```

```
data_copy['reviews_new'] = data_copy['reviews_new'].str.replace('\W', ' ', regex=True)
```

```
<>:1: DeprecationWarning: invalid escape sequence \W
<>:1: DeprecationWarning: invalid escape sequence \W
<>:1: DeprecationWarning: invalid escape sequence \W
<ipython-input-355-6586f97376b2>:1: DeprecationWarning: invalid escape sequence \W
  data_copy['reviews_new'] = data_copy['reviews_new'].str.replace('\W', ' ', regex=True)
```

```
data_copy
```

	reviews	reviews_new
ID		

```
data_copy.reviews_new.unique()
```

```
array(['  Trip Verified ', 'Not Verified ', '  Not Verified ',
       'Two regular an uneventful flights  Curiously enough  though  with the exact same crew  The crew were very nice and the service is very attentive and polite  but I just cannot take it that British Airways has chosen to provide a service just like low cost carriers  where everything is charged for  apart from luggage On the second leg of the trip  the aircraft felt extremely warm and for some odd reason  row 6 where I was sitting did not have AC '],
      dtype=object)
```

```
data_copy['reviews_new']=data_copy['reviews_new'].str.strip()
```

996	✅ Trip Verified | London to Belfast. Another f...	Trip Verified

```
data_copy['reviews_new']=data_copy['reviews_new'].replace('Two regular an uneventful flights  Curiously enough  though  with the exact
```

998	✅ Trip Verified | Warsaw to London. WAW is not ...	Trip Verified

```
data_copy.reviews_new.unique()
```

```
array(['Trip Verified', 'Not Verified'], dtype=object)
```

```
del data_copy['reviews']
counts = data_copy.apply(pd.Series.value_counts, axis=1)[['Trip Verified', 'Not Verified']].fillna(0)
data_copy=pd.concat((data_copy, counts.astype(int)), axis=1)

#data.join(counts)
data_copy
```

	reviews_new	Trip Verified	Not Verified
ID			
0	Trip Verified	1	0
1	Trip Verified	1	0
2	Trip Verified	1	0
3	Trip Verified	1	0
4	Trip Verified	1	0
...	...	...	...
995	Trip Verified	1	0
996	Trip Verified	1	0
997	Trip Verified	1	0
998	Trip Verified	1	0
999	Trip Verified	1	0

1000 rows × 3 columns

```
data_copy= data_copy.groupby(['reviews_new']).sum()
```

```
data_copy
```

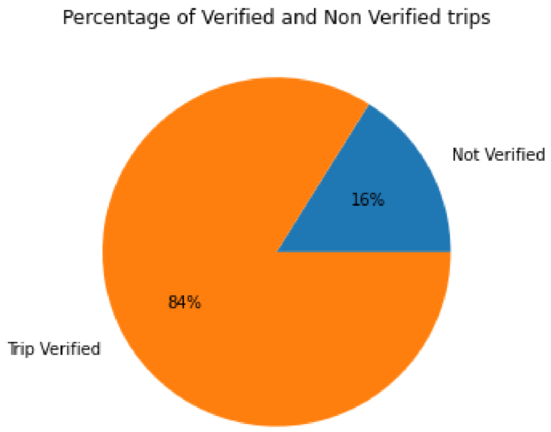
	Trip Verified	Not Verified
reviews_new		

```
data_copy.rename(columns={"Trip Verified": "SUM"}, inplace = True)
del data_copy['Not Verified']
data_copy.iloc[0]= 161
```

```
data_copy
```

	SUM
reviews_new	
Not Verified	161
Trip Verified	839

```
plt.rcParams["figure.figsize"] = (20,5)
plt.pie(data_copy['SUM'], labels= ['Not Verified',"Trip Verified"], labeldistance=1.15, autopct='%1.0f%%')
plt.title('Percentage of Verified and Non Verified trips')
plt.show()
```



Data Cleaning for LDA model

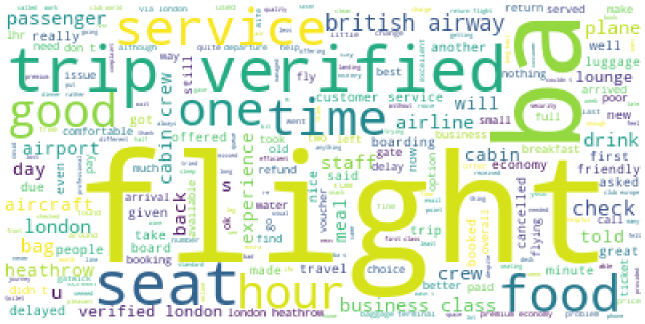
```
# Load the regular expression library
import re
# Remove punctuation

data['reviews'] = data['reviews'].str.replace('\W', ' ', regex=True)
# Convert the titles to lowercase
data['reviews'] = data['reviews'].map(lambda x: x.lower())
# Print out the first rows
data['reviews'].head()
```

```
<>:5: DeprecationWarning: invalid escape sequence \W
<>:5: DeprecationWarning: invalid escape sequence \W
<>:5: DeprecationWarning: invalid escape sequence \W
<ipython-input-367-3127f5cbca4f>:5: DeprecationWarning: invalid escape sequence \W
  data['reviews'] = data['reviews'].str.replace('\W', ' ', regex=True)
0      trip verified    i virtually gave up on brit...
1      trip verified    i was pleasantly surprised ...
2      trip verified    british airways is late  th...
3      trip verified   flew from amman to london on...
```

```
4      trip verified   this is the worst experience...
Name: reviews, dtype: object
```

Remove punctuation/lower casing

```
from wordcloud import WordCloud
long_string = ','.join(list(data['reviews'].values))
wordcloud = WordCloud(background_color="white", max_words=5000, contour_width=3, contour_color='steelblue')
wordcloud.generate(long_string)
wordcloud.to_image()
```



Exploratory Analysis

```
import gensim
from gensim.utils import simple_preprocess
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
stop_words = stopwords.words('english')
stop_words.extend(['from', 'subject', 're', 'edu', 'use'])
def sent_to_words(sentences):
    for sentence in sentences:
        # deacc=True removes punctuations
        yield(gensim.utils.simple_preprocess(str(sentence), deacc=True))
def remove_stopwords(texts):
    return [[word for word in simple_preprocess(str(doc))
             if word not in stop_words] for doc in texts]
data = data.reviews.values.tolist()
data_words = list(sent_to_words(data))
data_words = remove_stopwords(data_words)
print(data_words[:1][0][:30])
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
['trip', 'verified', 'virtually', 'gave', 'british', 'airways', 'three', 'years', 'ago', 'writing', 'avios', '
```

Prepare data for LDA Analysis

```
import gensim.corpora as corpora
id2word = corpora.Dictionary(data_words)
texts = data_words
corpus = [id2word.doc2bow(text) for text in texts]
print(corpus[:1][0][:30])
```

```
[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 2), (7, 1), (8, 1), (9, 1), (10, 1), (11, 1), (12, 1), (1
```

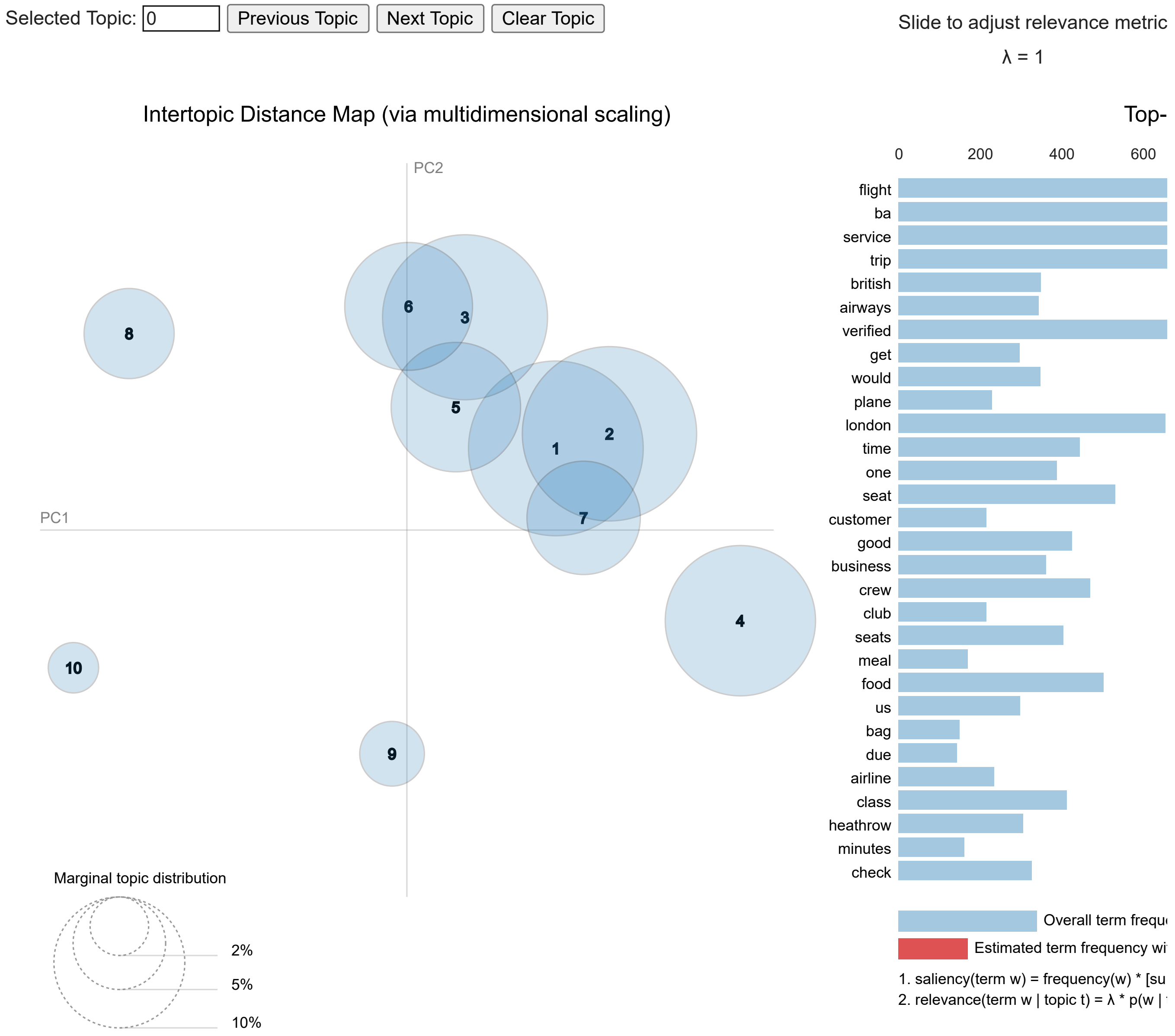
LDA model training

```
from pprint import pprint
num_topics = 10
lda_model = gensim.models.LdaMulticore(corpus=corpus,
                                       id2word=id2word,
                                       num_topics=num_topics)
pprint(lda_model.print_topics())
doc_lda = lda_model[corpus]
```

Analyzing LDA model results

```
import pyLDAvis.gensim_models as gensimvis
import os
import pickle
import pyLDAvis
pyLDAvis.enable_notebook()
LDAvis_data_filepath = os.path.join('reviews_words'+str(num_topics)+'.csv')
if 1 == 1:
    LDAvis_prepared = gensimvis.prepare(lda_model, corpus, id2word)
    with open(LDAvis_data_filepath, 'wb') as f:
        pickle.dump(LDAvis_prepared, f)

with open(LDAvis_data_filepath, 'rb') as f:
    LDAvis_prepared = pickle.load(f)
pyLDAvis.save_html(LDAvis_prepared, 'reviews_words'+ str(num_topics) +'.html')
LDAvis_prepared
```

```
/usr/local/lib/python3.7/dist-packages/pyLDAvis/_prepare.py:247: FutureWarning: In a future version of pandas
  by='saliency', ascending=False).head(R).drop('saliency', 1)
```



1. saliency(term w) = frequency(w) * [su
2. relevance(term w | topic t) = λ * p(w |