Web scraping to gain company insights

**Here is the background information on your task**

British Airways (BA) is the flag carrier airline of the United Kingdom (UK). Every day, thousands of BA flights arrive to and depart from the UK, carrying customers across the world. Whether it’s for holidays, work or any other reason, the end-to-end process of scheduling, planning, boarding, fuelling, transporting, landing, and continuously running flights on time, efficiently and with top-class customer service is a huge task with many highly important responsibilities.

As a data scientist at BA, it will be your job to apply your analytical skills to influence real life multi-million-pound decisions from day one, making a tangible impact on the business as your recommendations, tools and models drive key business decisions, reduce costs and increase revenue.

Customers who book a flight with BA will experience many interaction points with the BA brand. Understanding a customer's feelings, needs, and feedback is crucial for any business, including BA.

This first task is focused on scraping and collecting customer feedback and reviewing data from a third-party source and analysing this data to present any insights you may uncover.

**Here is your task**

**Scrape data from the web**  
The first thing to do will be to scrape review data from the web. For this, you should use a website called [Skytrax](https://www.airlinequality.com/" \t "_blank).

The team leader wants you to focus on reviews specifically about the airline itself. You should collect as much data as you can in order to improve the output of your analysis. To get started with the data collection, you can use the **“Jupyter Notebook”** in the Resources section below to run some Python code that will help to collect some data.

**Analyse data**  
Once you have your dataset, you need to prepare it. The data will be very messy and contain purely text. You will need to perform data cleaning in order to prepare the data for analysis. When the data is clean, you should perform your own analysis to uncover some insights. As a starting point, you could look at ~~topic modelling,~~ sentiment analysis or wordclouds to provide some insight into the content of the reviews. It is recommended to complete this task using Python, however, you can use any tool that you wish. You can use some of the documentation websites provided in the Resources section below to analyse the data.

Please ensure that you have created a folder called "data" and mapped your file path.

**Present insights**  
Your manager would like you to summarise your findings within a single PowerPoint slide, so that they can present the results at the next board meeting. You should create visualisations and metrics to include within this slide, as well as clear and concise explanations in order to quickly provide the key points from your analysis. Use the **“PowerPoint Template”** provided to complete the slide.

Once you’ve completed your PowerPoint, please submit your document below.

NLP is used in text processing mainly and there are many kinds of tasks that are made easier using NLP. Eg: In chatbots, Autocorrection, Speech Recognition, Language translator, Social media monitoring, Hiring and recruitment, Email filtering, etc.

**Estimated time for task completion: 1.5 hours, depending on your learning style.**

1. Web scraping of reviews

2. Topic Modelling:

Some of the important points or topics which makes text processing easier in NLP:

Removing stopwords and punctuation marks

Stemming

Lemmatization

Encoding them to ML language using Countvectorizer or Tfidf vectorizer

Topic modeling techniques

The techniques listed below are some of the most common and popular techniques that are used to perform topic modeling in NLP:

Latent Semantic Analysis (LSA)

Probabilistic Latent Semantic Analysis (pLSA)

Latent Dirichlet Allocation (LDA)

Topic modelling is done using LDA(Latent Dirichlet Allocation). Topic modelling refers to the task of identifying topics that best describes a set of documents. These topics will only emerge during the topic modelling process (therefore called latent). And one popular topic modelling technique is known as Latent Dirichlet Allocation (LDA).

Topic modelling is an unsupervised approach of recognizing or extracting the topics by detecting the patterns like clustering algorithms which divides the data into different parts. The same happens in Topic modelling in which we get to know the different topics in the document. This is done by extracting the patterns of word clusters and frequencies of words in the document.

So based on this it divides the document into different topics. As this doesn’t have any outputs through which it can do this task hence it is an unsupervised learning method. This type of modelling is very much useful when there are many documents present and when we want to get to know what type of information is present in it. This takes a lot of time when done manually and this can be done easily in very little time using Topic modelling.

Latent Dirichlet Allocation:

In LDA, latent indicates the hidden topics present in the data then Dirichlet is a form of distribution. Dirichlet distribution is different from the normal distribution. When ML algorithms are to be applied the data has to be normally distributed or follows Gaussian distribution. The normal distribution represents the data in real numbers format whereas Dirichlet distribution represents the data such that the plotted data sums up to 1. It can also be said as Dirichlet distribution is a probability distribution that is sampling over a probability simplex instead of sampling from the space of real numbers as in Normal distribution.

Now when topic modelling is to get the different topics present in the document. LDA comes to as a savior for doing this task easily instead of performing many things to achieve it. As LDA brings the words in the topics with their distribution using Dirichlet distribution. Hence the name Latent Dirichlet Allocation. The words assigned(or allocated) to the topic with their distribution using Dirichlet distribution.