Project assessment proposal: Flight Scrapping

Goal

There are webpages such as Skyscanner that scraps different flight sites in order to tell you the cheaper flights between two cities. Our goal is not to replicate it, but to improve a feature that the web doesn't have.

The thing is that with the web you can only look for the flights in one date and one location at the time. So if, for instance, you have flexibility of dates to make your trip (imagine you can depart in a range of 3 days and come back in another range of 3 days) and you do not have any preferences between going to either some different airports inside the country you are visiting (imagine there are 5 airports which you are fine to go because you are doing a trip around the country), you have to search date by date and airport by airport in order to see all possibilities (for our example would be 45 different searches).

Our scope is that, given the range of days you are able to depart, the range of the days you are able to come back, the city of departure and the country you are visiting, our algorithm ranks you the flights given two criteria: the price and the total amount of hours of trip.

The data source and web technology it is based upon

The data would be taken from https://www.skyscanner.com mainly. Its webpage is based on html technology.

The approach/technology that you will use for downloading

When a flight search is done, the page makes a query such as:

https://www.skyscanner.es/transporte/vuelos/bcn/lgw/170812/170902/.html#results

As we can see, it is looking flights from Barcelona – El Prat to London – Gatwick, departing the 12th of August and returning the 2nd of September. We could, with the input data of our algorithm, change the query of the webpage with all the possibilities and make web scrapping from its content each time.

How you plan to clean the data (scraping)

First, we would look at the html code in order to get familiar with the structure of the page. After that, we could use *regex* in order to get the parts we are interested to know about (prices and time of flights).

How you would analyze the extracted information

So, once we have extracted the information from the webs of all possible combinations of dates and airports we input, we would create a table with variables such as:

- From_City: City of departure (Optional)
- To_Country: Country of destination (Optional)
- Airport: Airport of destination
- Depart: Date of departure
- Return: Date of return
- Price: Cost of the ticket (€)
- Time_Depart: Time of the first flight
- Time_Return: Time of the second flight
- Time: Sum of the time of both flights (Optional)
- Score: 0-10 grade based on a combination of cost and time (Optional)

And we would sort it depending of the cost, the time, or a combination of both, so that the user quickly knows which are the best options.