Data Scrapping from HolidayIQ using Scrapy

Scrapy is a Python framework for large scale web scraping. It gives you all the tools you need to efficiently extract data from websites, process them as required and store them in the desired structure and format. Scrapy uses XPATH as the query language to select nodes from an XML or HTML document.

In the Visual Studio Code editor, there is a ‘spiders’ folder. Create a new spider(filename.py). At first, we must Import scrapy. As spiders are classes, we must define a class and inherit from the ‘scrapy.Spider’ class. Each spider is divided into 3 parts. The 1st part is spider name/identity, the 2nd part is responsible for sending URL(s) and the 3rd part to receive the GET response.

In the request part we enter the url(s) to be scraped.

The response part is responsible to catch the response returned by each request sent. We need to specify the xpath container in the response selector. We also specify the XPATHS for the various individual items that we need to extract.

Static Pagination

To automate static pagination, get the XPATH for the ‘NEXT” link which is an href attribute in static paging. We then need to join the href with the url of the current page and initiate a scrapy request.

Dynamic Pagination

To automate dynamic pagination, we need to transform the input URLs using the request URL we get from the XHR. ­­We also need to key in the Authorization header for the request to work outside the browser.