Data Scraping of New Geographical Locations, Industries and Job Roles

For capturing data from many websites, web scraping is required. In web scaping process, code is written to send a request to the server that's hosting the web page. Usually, the code downloads that page's source code, just as a browser would but instead of displaying the page visually, we will filter through the page looking for specific HTML elements and extracting whatever content is required.

The following are used to accomplish this:

- **Beautiful Soup library-**

  BeautifulSoup is a python library which will be used to parse the HTML document, and extract the text from the relevant HTML tags. We will import this library, and create an instance of the BeautifulSoup class to parse the page content.

  For example, if we wanted to get all of the titles inside H2 tags from a website, we could write some code to do that. Our code would request the site's content from its server and download it. Then it would go through the page's HTML looking for the H2 tags. Whenever it found an H2 tag, it would copy whatever text is inside the tag, and output it in whatever format we specified.

- **requests library-**

  The first thing we'll need to do to scrape a web page is to download the page for which the Python requests library is used. It makes a GET request to a web server, which will download the HTML contents of a given web page. After running this request, we get a Response object. This object has a status_code property, which if 200 then that indicates the page was downloaded successfully.

- **Pandas library-**

  It is a Python library used for data analysis. Here, we will use the DataFrame class to pass scraped items lists. They are passed as a part of a dictionary where key becomes the column of the DataFrame and each list becomes the column values.

➢ Locations required:

UK

➢ Industries required:

Pharmaceuticals, Software services, FMCG, petrochemicals, Automobiles, constructions, financial services, food and beverages, textiles, Real estate

➢ Steps to follow:

1. Firstly, analyse the URL pattern of the website whose data you are planning to scrape. Like in this case, URL to be scraped is having three variable parameters- job title, location and number of jobs being listed on a single page (incremental window of count 10).
2. Data is then requested from this dynamic URL using requests library.
3. Three lists are prepared- job titles, company, summary.
4. By right clicking on the page then clicking "Inspect" we can find out the names and classes required for extraction, like for list 1 having job titles, firstly name- h2 and class-title is extracted. Then, name- a and class- jobtitle is used to loop the job titles listed.
5. Similarly, for list 2 having company data, name-span and class-company is used from parsed response object.
6. Similarly, for list 3 having summary data, name-div and class-summary is used from parsed response object.
7. These lists are converted to dictionary having three key-value pairs which is then loaded to Pandas DataFrame.
8. This DataFrame will then be stored in CSV format and this scraped data is then used further for scraping new data.