REPORT

Task: To code a web scraper which scrapes data from Naukri, monster, indeed, time jobs, shine and LinkedIn.

Libraries used: BeautifulSoup4, Selenium, pandas.

Operating system worked on: Windows

Columns:

1. Recruiter name
2. Recruter tel
3. Recuiter mail id
4. Company website
5. Job location
6. Company name
7. Skill set required
8. Description url
9. Salary offered
10. Experience required
11. Qualification required

Description:

1. Selenium in this task was used to automate the process.
2. BeautifulSoup was used to pull the required data from the html of the required webpage
3. The parser used along with BeautifulSoup is ‘html.parser’.
4. All the scripts are in the folder WebScraper and it is declared as a package.
5. Script main.py is run to initiate the web scraping process.
6. Further in main.py, naukri, monster, indeed, time_jobs, shine and linkedIn gets imported and main method is called which further calls

Procedure:

1. At 1st, URL of the required page is observed, how the URL is changing with the actions of the users like adding filters, clicking on next page.
2. Except for shine and linkedin, a generic URL can be used to add filters and navigate through pages, and for shine and linkedin, selenium was used to automate things.
3. And in the script, here we are using chrome driver, webdriver present in Selenium is used to access the browser.

4. Now, we find all the frames which have our required data using selenium driver and later using for loop, each frames innerHTML is extracted and this innerHTML is given to BeautifulSoup as parameter to get our parsed HTML.
5. In browser, we identify the required data by inspecting the element and find it in the parsed HTML by using find methods of BeautifulSoup.
6. And those extracted data is appended to pandas DataFrame to their respective column names and this repeats for all the Job frames in that webpage and later move to next page till the last page.
7. Now the DataFrame is converted into CSV file.
8. So, now when we run main.py, it takes skill, location and experience input from the user and then changes the URL in each file accordingly and then performs task.