Web Scraping Project :

INDEED.COM DATA JOBS SCRAPING

Odomero Omokahfe • 437967
Karen Gurupira •446087

Introduction

As a recent graduate it is difficult to find an entry level job. The aim of our project was to scrape data from Indeed.com useful to recent graduates in the data science and analysis field. In the world that is widely adapting remote working as the new norm, our project focused on remote jobs only which would allow eligible candidates to apply despite of their residential location.

Data scrapped also includes salaries offered by employers. This would be helpful to a graduate as they can use it as a guide when putting forward their remuneration requests.

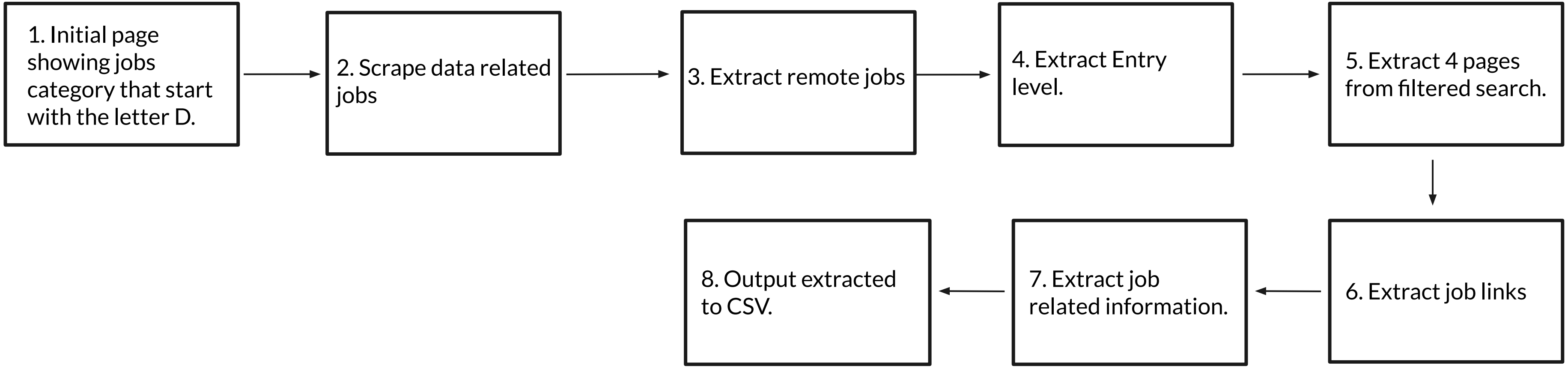
The information scrapped was targeted for students who;

- Hold a Master's degree in Data Science or any related field.
- Looking for entry level jobs.
- In search for Remote work.

Indeed is an American worldwide employment website for job listings launched in November 2004. The website provides a wide search engine for different types of jobs from all over the world.

Scraping Process.

- Web scraping extracts underlying HTML code and data from a web page. In this project, three scraper tools were used in the process of collecting structured web data in an automated fashion, thus computing the same output. The following crawlers were used; BeautifulSoup, Scrapy and Selenium.
- Each scraper followed the steps detailed below.



1. Initial page showing jobs category that start with the letter D.

2. Scrape data related jobs

3. Extract remote jobs

4. Extract Entry level.

5. Extract 4 pages from filtered search.

6. Extract job links

7. Extract job related information.

8. Output extracted to CSV.

Scraper Mechanics and Performance

Beautiful Soup

Beautifulsoup is a parsing library which summons content from URL allowing us to parse certain parts of the page with no hassle.
Unlike the Selenium crawler, BeautifulSoup does not operate automatically unless an infinite loop with a certain criteria is put.

Scrapy

Scrappy is a complete framework scraper which operate automatically when constraints are specified. Because scapy is a complete framework, it includes certain tools that make web-scraping easier. These tools include:

- Feed exports which allows for scraped data to be stored in formats like CSV,JSON and XML.
- Selectors that enable scrapy to be more efficient than BeautifulSoup.

We observed that Scrapy seem to run faster than BeautifulSoup and Selenium for data extraction.

Selenium

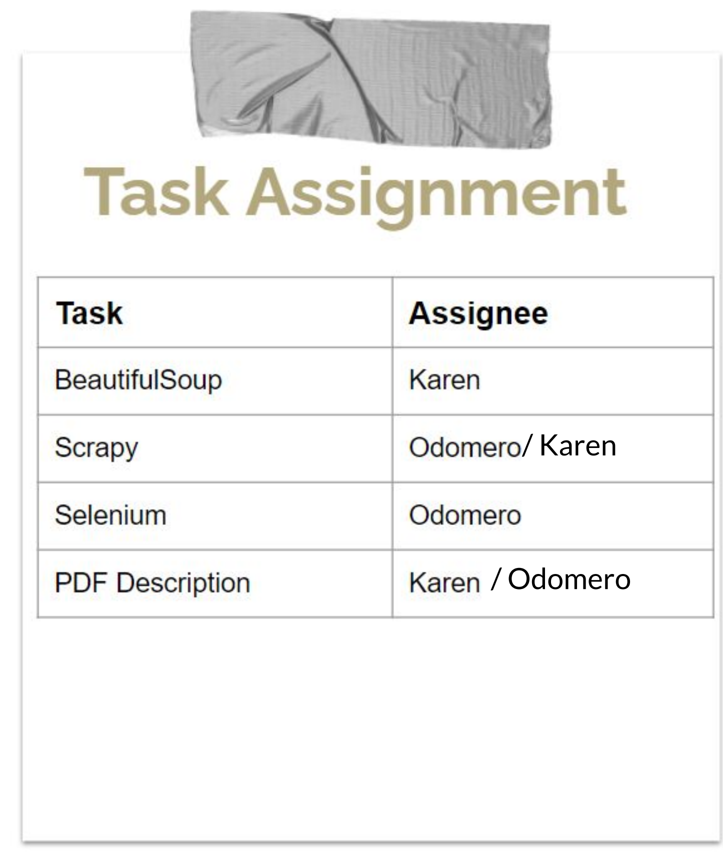
Selenium is a package that is used to automate web browser interaction. We used Selenium geckodriver(Firefox browser driver) to interact with Indeed.com to retrieve all data remote entry level jobs.
As interesting as it was automating the data extraction process with Selenium, it had a longer run time compared to the other two Scrapers.

Data output and analysis

The data output exhibits a total of about 200 remote entry level data related jobs as at the time of preparing this report. Further analysis of the data output reveals that some jobs do not have any salary or job-type information. Despite this, a useful and efficient analysis of the data output can be performed and job seekers can utilise the information to directly access job application portals to make submit applications.

The data output from the scrapers shows the company name, salary/salary range for the vacancies, the job title, the job type, and the hyperlink to the job listings. The data generated can be used for, and not limited to;

- Analysing how salaries differ from company to company, within the Data job industry
- Analyzing how salaries differ for full time jobs and contract jobs.
- Analyzing which data jobs category has more job vacancies between Data Analyst, Data Entry Clerk, Data Entry operator and Database Administrator.



Task Assignment

Task	Assignee
BeautifulSoup	Karen
Scrapy	Odomero/ Karen
Selenium	Odomero
PDF Description	Karen / Odomero