INFO-SCRAPER

PROJECT SYNOPSIS

OF MINOR PROJECT

BACHELOR OF TECHNOLOGY

INFORMATION & TECHNOLOGY

SUBMITTED BY:

ISHA ARORA (2004930)

JAIYATI NANDA (2004932)

HARMAN PREET (2004921)

GURMANNAT WALIA (2004912)

GURU NANAK DEV ENGINEERING COLLEGE,

LUDHIANA

Contents

1. Introduction

The process of extracting data from websites is termed as web Scraping which includes the use of programming languages and libraries such as BeautifulSoup to get and analyse HTML and some formats of web page and retrieve appropriate content from it. Use of web scraping tools automate the extraction of data and then that data is stored in a structured format such as a spreadsheet or database. This helps in quick searching and filtering of the information.

This project focuses on getting the useful information from a website skillout which is a job search platform that allows users to find internship opportunities across a variety of industries and locations. However, manually collecting data from Skillout can be time-taking and ineffective, especially while looking for internships in multiple locations or industries. This is where web scraping comes in.

Overall, web scraping Skillout internship website can be a powerful tool for college students and professionals, letting them to gather internship rolls quickly and efficiently from one of the leading job search platforms. This can help students find internships that are related to their interests and career goals, and give them a competitive advantage in the job market.

2. Rationale

Colleges and universities can benefit from scraping the Skillout internship website for several reasons, including:

- Finding relevant internships: The skillout website offers a huge database of internships in a variety of fields and locations. Colleges can quickly locate relevant internships for their students based on their interests, abilities, career objectives by scraping this page. By automating the search and application processes can save time and resources.
- Staying up-to-date with internship trends: Colleges may keep up with the most recent internship trends, such as which industries are hiring the most interns or what kinds of talents are in demand, by routinely scraping the Skillout website and colleges can gain a competitive advantage by being the first to identify new internship opportunities and trends. This can help them attract more students and enhance their reputation as a leading provider of internship services.

3 Objectives

Depending on the requirements of the college or organisation, several goals may be set for scraping the Skillout internship website. Common goals include the following:

- To provide a wider pool of internships.
- To automate the internship search process.

4. Feasibility Study

A feasibility study of web scraping involves analysing whether it is technically possible, and legally permissible to scrape data from a website.

- Technical feasibility: The technical feasibility of web scraping a website requires knowledge of a website structure and format or type of data to be scrapped. It would also involve the type of tools and technologies which are best suited for a certain type of websites or data extraction task.
- Legal feasibility: Online scraping may violate a website's terms of service or, in some circumstances, be unlawful. It's crucial to follow the terms of service of the website and refrain from using the data that has been scraped for any immoral or unlawful activities.
- Economic feasibility: Online scraping can take a lot of time, especially if while scraping a big website or regularly gathering data. It is crucial to have the required time, people, and computing resources to finish the project.

5. Methodology Used

The methodology for scraping involves the following steps:

- Identifying the target website and specifying the pages from which data is to be extracted and then choosing a scraping tool including BeautifulSoup and many more.
- Writing a piece of code to extract the necessary information from the webpage is the next step. This may involve identifying the HTML elements that contain the data that is to be extracted such as job titles, company names and job description.
- Once the data has been extracted, store it in a structured format, like a spreadsheet or database. Since new internship opportunities are regularly updated, it is vital to update the data frequently.
- After that, the data obtained can be used to provide students a competitive edge in the job market by assisting them in locating internships that are tailored to their interests and career aspirations.

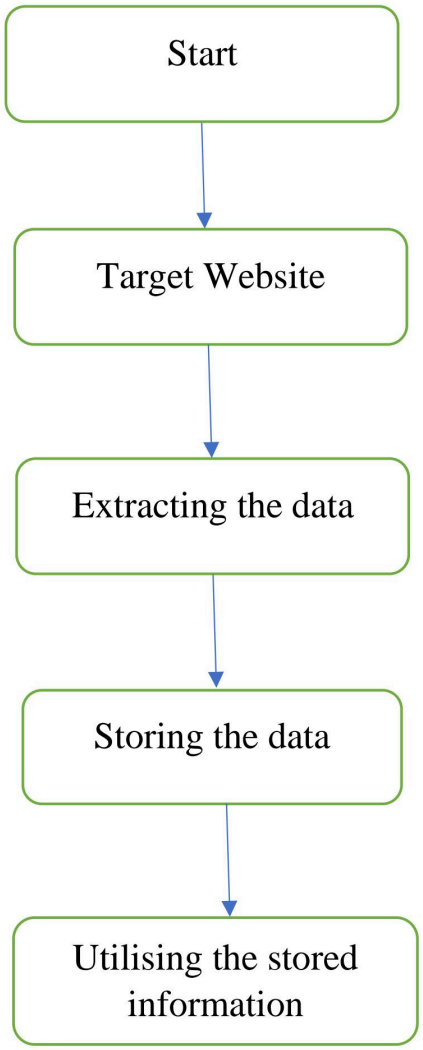

FIG.1 Flowchart

6. Facilities required

Hardware Requirements

- CPU: 2x 32-bit 2.8 GHz 8.00 GT/s CPUs or above
- RAM: 2GB or above
- A reliable internet connection

Software Requirements

- Browser: Google Chrome
- SQL database
- PyCharm 2.7or above

7. Expected Outcomes

- A list of all the available internships.
- Emphasising the abilities and experiences that match the requirements of the position.
- Analysis of trends in the kinds of internships that are available.

8. References

[1] V. Singrodia, A. Mitra and S. Paul, "A Review on Web Scrapping and its Applications," 2019 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2019, pp. 1-6. Retrieved from *https://www.tandfonline.com/doi/full/10.1080/10691898.2020.1787116.*

[2]Inoue, Megumi & Li, Meng-Hao & Hashemi, Mahdi & Yu, Yang & Jonnalagadda, Jahnavi & Kulkarni, Rajendra & Kestenbaum, Matthew & Mohess, Denise & Koizumi, Naoru. (2023). Opinion and Sentiment Analysis of Palliative Care in the Era of COVID-19. Healthcare. 11. 855. 10.3390/healthcare11060855. Available at *https://researchdata.wisc.edu/news/an-introduction-to-web-scraping-for...*

[3] D. PRATIBA, A. M.S., A. DUA, G. K. SHANBHAG, N. BHANDARI and U. SINGH, "Web Scraping And Data Acquisition Using Google Scholar," 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), Bengaluru, India, 2018, pp. 277-281, doi: 10.1109/CSITSS.2018.8768777 Retrieved *from Web Scraping And Data Acquisition Using Google Scholar | IEEE Conference Publication | IEEE Xplore*.

[4] Sharma, Rizul. (2020). Web Data Scraping. 10.13140/RG.2.2.15125.76009. Retrieved from *https://www.researchgate.net/publication/342011184*