Web Scraping Project Report

Introduction

Web scraping is the process of extracting data from websites using software tools. It is widely used for collecting publicly available information for research, business analytics, and data-driven decision-making. In this project, we focused on extracting college-related information from indcareer.com The goal was to compile a structured dataset containing key details of colleges from different states and territories in India and store them in an Excel sheet for further analysis.

Project Objectives

- The primary objectives of this project were:
- Extracting college details such as college name, city, phone number, website, and affiliation from Indeed.in.
- Standardizing and organizing the extracted data into an Excel sheet.
- Identifying and overcoming challenges related to web scraping.
- Improving efficiency and accuracy in data extraction.

Challenges Faced and Solutions Implemented

This project was our first hands-on experience with web scraping. While we had theoretical knowledge, practical implementation brought several challenges. Below are the key challenges and the steps taken to overcome them:

1. Extracting Email Addresses

- Challenge: Emails were not directly available due to authentication restrictions.
- Attempts: We used different scraping techniques, including Selenium, but were unable to bypass these restrictions.
- Solution: We explored alternative approaches such as API integration and manual validation but found that many entries had placeholders like "Email Not Available."
- Next Steps: Investigate other APIs that might provide email data or explore official college websites for email retrieval.

2. Missing Affiliation Details

- Challenge: Some colleges lacked affiliation details, making the dataset incomplete.
- Solution: We considered using secondary sources to validate the missing affiliation details manually.
- Next Steps: Enhancing scraping logic or integrating data from additional educational directories.

3. City Field Formatting Issues

- Challenge: The "City" field included extra details such as "Tamil Nadu, India," which affected data clarity.
- Solution: We modified the data processing script to split the field into three separate columns: City, State, and Country.

4. Phone Number Standardization

- Challenge: Extracted phone numbers varied in format, making them difficult to use uniformly.
- Solution: We implemented a script to reformat phone numbers into a standardized international format.

5. Slow Execution and Performance Issues

- Challenge: Running the script took hours, leading to significant delays.
- Solution:
  - Optimized the script by reducing redundant calls.
  - Used asynchronous scraping techniques for improved speed.
  - Handled network interruptions with retry mechanisms to avoid data loss.

6. System and Connectivity Issues

- Challenge: Long execution times led to system crashes and Wi-Fi failures.
- Solution: We optimized the code to reduce execution time and ran the script in smaller batches.

Key Outcomes and Achievements

After several iterations and improvements, we successfully extracted the following details for 25 colleges in Tamil Nadu:

- College Name
- City, State, Country (Formatted properly)
- Phone Number (Standardized format)
- Affiliation Details

- Website Links
  However, due to website restrictions, email extraction remained a challenge, and further investigation is needed to resolve this issue.

In last what we have achieved:

- Expanded dataset coverage beyond Tamil Nadu to include other Indian states as well as union territories to get the information of all the collages in India
- Improve scraping efficiency by using different sets of library and code
- Used machine learning techniques to clean and validate extracted data for better accuracy.
- We have compiled all the excel files in one in which we have all the information about the colleges which are on the indcareer.com website including all states and territories.

Acknowledgment

This project was successfully completed under the guidance of Dr.Thejus Kartha, whose support was invaluable in resolving technical challenges and improving our approach. His mentorship helped us refine our scraping techniques and optimize our data extraction process.

Conclusion

Despite various challenges, we successfully extracted and organized valuable college information. The project provided us with hands-on experience in web scraping, data processing, and problem-solving. While some hurdles, like email extraction, remain unresolved, our learnings from this project will help us build more efficient and scalable scraping solutions in the future.