COMP39/9900 Computer Science/IT Capstone Project
School of Computer Science and Engineering, UNSW
Course Convenor: Dr. Basem Suleiman(b.suleiman@unsw.edu.au)

Project Number: 79

Project Title: Web Platform for Data Scraping Tool

Project Clients: Dr. Baseem Suleiman, Namit Khurana

Project specializations: Software Development; Web Application, Development; Data Crawling; Web Scraping

Number of groups: 2 groups

Background:

With the increasing demand for niche-specific data, businesses and individuals require quick and customizable solutions to gather web data efficiently. Current web scraping solutions often require users to have technical knowledge or use multiple tools to retrieve, process, and organize the data they need. This creates a barrier for non-technical users or businesses lacking development resources. The demand is especially high in domains like finance, medicine, retail, and weather, where real-time and accurate data is crucial for decision-making.

This project aims to solve this gap by developing an all-in-one web scraping platform, allowing users from any industry or niche to request data tailored to their needs, while handling the complexities of web scraping, data cleaning, and formatting.

Project Goals:

The main objective is to create a web platform that provides:

1. **Customizable Web Data Scraping**: Users submit their requirements (industry, specific websites, data format), and the platform scrapes the web to gather the requested information.

2. **User-Friendly Interface**: A simple, intuitive interface that allows users to define their data needs without requiring technical expertise.

3. **Flexible Data Delivery**: The platform should provide data in multiple formats (JSON, CSV, Excel, or API integration), depending on the user's preferences.

4. **Automated Data Processing**: Data scraped from the web will undergo cleaning, structuring, and transformation to ensure it is ready for use.

5. **Compliance and Security**: Ensure that web scraping adheres to legal guidelines, such as respecting robots.txt, and data is handled securely.

Requirements and Scope:

1. **Users**: Any individual or business needing niche-specific data from sectors like finance, medicine, retail, or weather.
2. **Features**:
    - **Custom Query Submission**: Users can define the type of data they need and any specific websites or sources.
    - **Data Formatting**: Offer various formats (JSON, CSV, XML, etc.) or customized formats based on user requests.
    - **Automated Scheduling**: Ability to set up automated scrapes for real-time updates at regular intervals.
    - **User Authentication & Dashboard**: Users can track progress, manage their scraping requests, and access their results from a personalized dashboard.
3. **Performance**:
    - **Scalability**: Handle multiple requests concurrently, from different industries and websites.
    - **Efficiency**: Ensure quick scraping and processing with the ability to scrape data from a wide range of websites, regardless of the complexity of the page structure.

Required Knowledge and skills:

Developing this requires some knowledge in frontend and backend technologies to build a user-friendly and functional interface. Knowledge of web scraping tools such as Scrapy, BeautifulSoup, and Selenium is good for efficient data extraction. Additionally, experience in data cleaning and processing with tools like Pandas is needed to structure data effectively. Database management, API development are also a part of the required skill set

Expected outcomes/deliverables:

The project will deliver a functional web scraping platform with complete source code and user-friendly documentation. It will include guides for both users and administrators, covering everything from data request submission to managing and scaling the platform.

Supervision:

Dr. Baseem Suleiman and Namit Khurana.