Madi Reddy Alekya

11610976
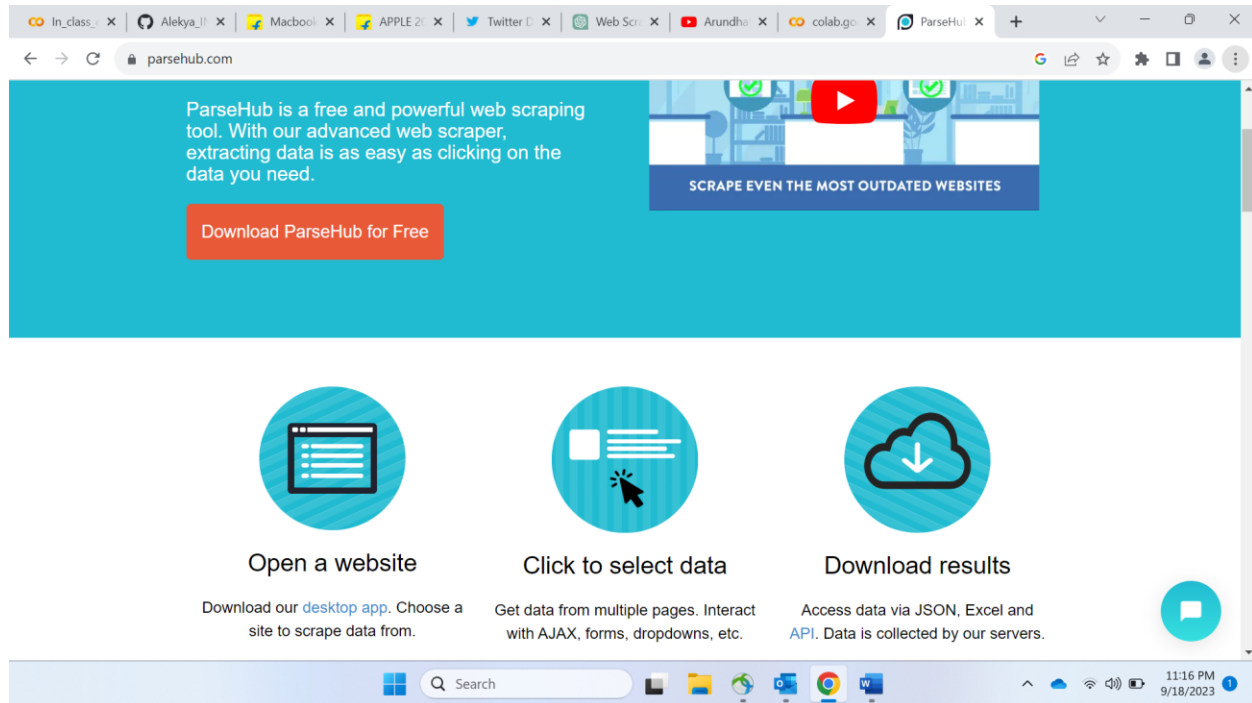
INFO 5731 – Computational Methods

Exercise 02

Introduction to ParseHub:

ParseHub is a powerful and user-friendly web scraping tool that allows you to extract data from websites without the need for coding. It offers a point-and-click interface, making it accessible to users with little or no programming experience. ParseHub can scrape data from various websites, including e-commerce sites, news websites, and more. You can then export the scraped data in formats like CSV, Excel, JSON, or Google Sheets.

Steps to Follow for Web Scraping Using ParseHub:

Install ParseHub:

- Visit the ParseHub website (https://www.parsehub.com/) and sign up for an account.
- Download and install the ParseHub desktop app, which is available for both Windows and macOS.

Create a New Project:

Launch ParseHub and click on "New Project."

Enter the URL of the website you want to scrape.

Select Data to Scrape:

Use the ParseHub interface to select the elements on the webpage you want to scrape. You can click on the data you want to extract, and ParseHub will automatically detect similar items.

Train the Scraper:

Define the data you want to extract by creating "selectors" for each piece of information you need. For example, if you're scraping a product page, you can create selectors for the product name, price, and description.

Pagination (Optional):

If the data spans multiple pages (e.g., search results), you can set up pagination by instructing ParseHub to navigate through multiple pages of the website.
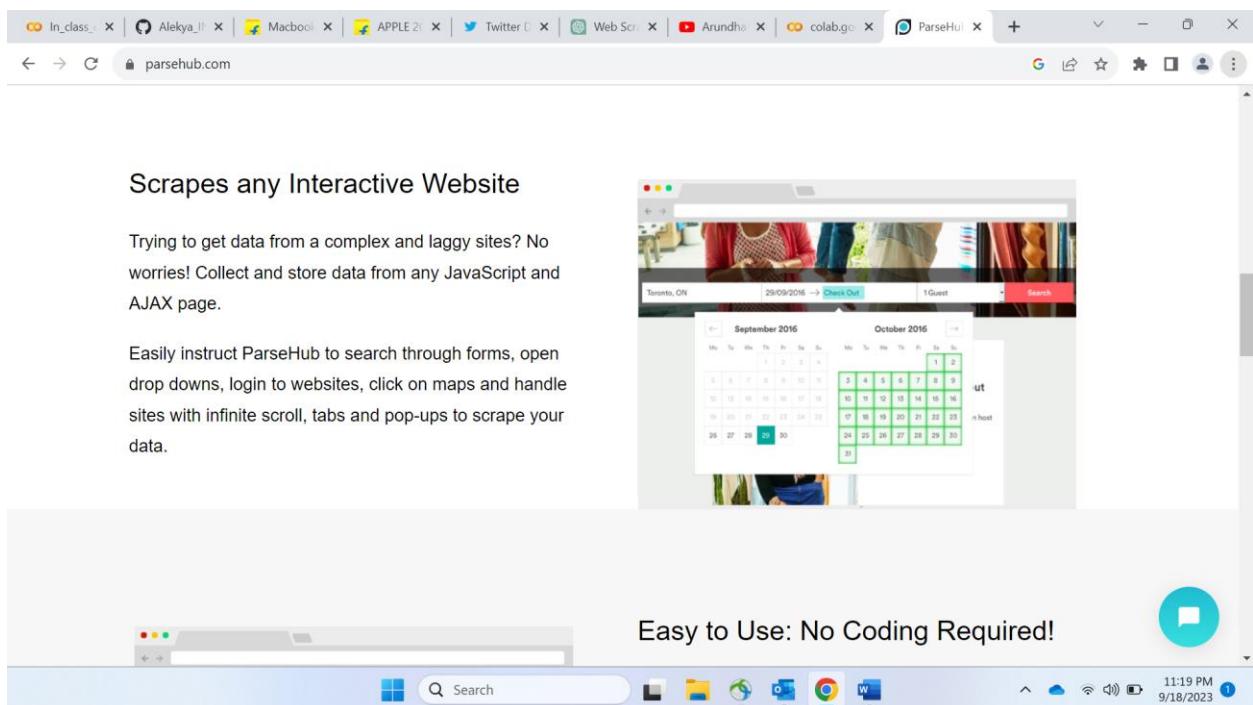
Run the Scraper:

Click on the "Run" button to start the scraping process. ParseHub will visit the website, collect the data as per your instructions, and display it in the app.

Review and Export Data:

Review the scraped data to ensure it's accurate.

Once satisfied, click on "Save & Run" to run the scraper in the cloud (if needed).
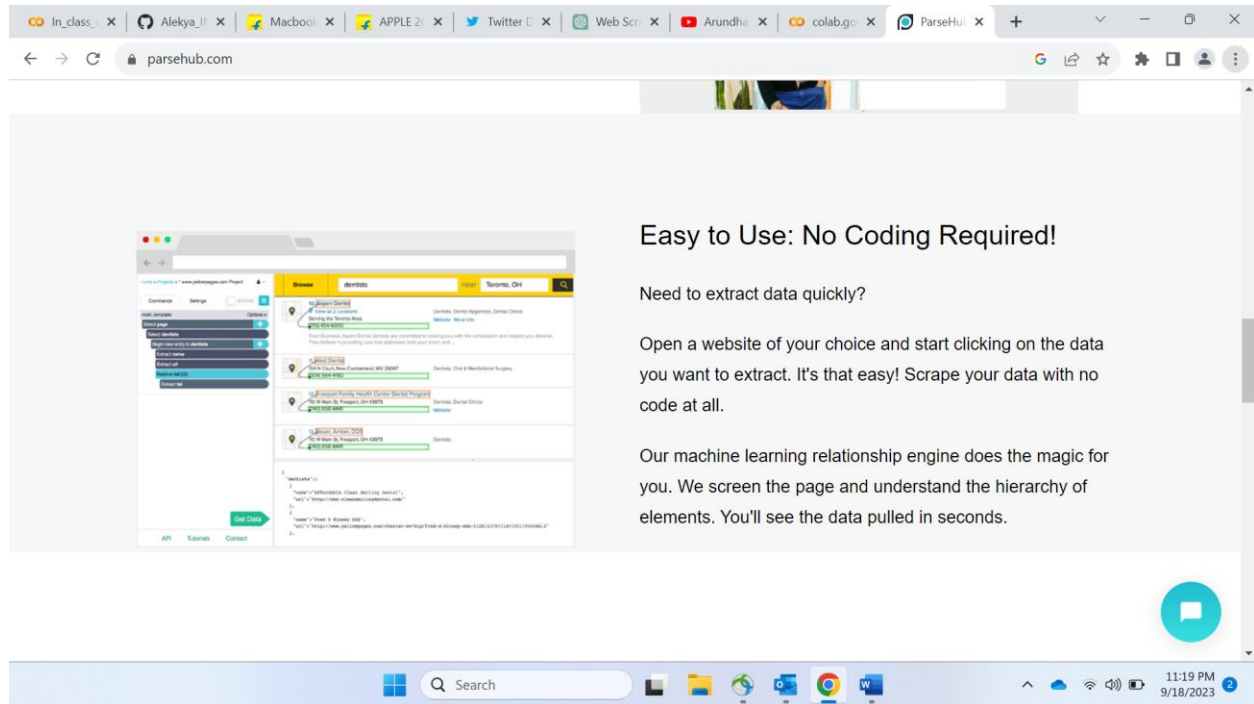
Finally, click on "Download Data" to export the data in your preferred format, such as CSV, Excel, JSON, or Google Sheets.

Final Output of the Data Collected:

- The final output of the data collected using ParseHub can be saved in various formats, including:
- CSV (Comma-Separated Values): This format is commonly used for data interchange and can be easily imported into spreadsheet software like Excel or Google Sheets.

- Excel: Data can be saved in Excel format, which is suitable for further analysis, charting, and reporting.
- JSON (JavaScript Object Notation): JSON format is widely used for structured data storage and exchange in web applications and can be used for more complex data structures.
- Google Sheets: You can export the data directly to Google Sheets, making it accessible and shareable with collaborators.

By following these steps, you can efficiently scrape data from websites using ParseHub and export it in your preferred format for analysis or other purposes.