**Assignment:**

1. Given a CSV of g2crowdurls (a URL shortening service), write a python

module that uses playwright-browser to scrape the contents of a company’s

details from G2Crowd. The goal is to scrape the company page and return

the output in a structured format. Output can be a JSON or a CSV file with

fields and values.

2. Please ensure the following points as well:

1. All API calls should be asynchronous

2. Exception handling needs to be present wherever applicable.

**Answer**

As per your assignment tried my best for the result. Giving all code details and all steps which are required to run the code.

**Code details:**

1. The code imports necessary modules: **csv** for working with CSV files, **json** for working with JSON data, **asyncio** for asynchronous programming, and **async\_playwright** from Playwright library for browser automation.
2. The **scrape\_company\_details** function is defined. It takes a URL as a parameter and is responsible for scraping company details from the webpage corresponding to that URL. It uses Playwright to launch a Chromium browser, navigate to the URL, wait for a specific element with the class name 'company-details' to load, and then extract the company name, description, and rating from the webpage. It returns a dictionary with the scraped data.
3. The **main** function is defined. It serves as the entry point of the program. It specifies the input CSV file (**g2crowd\_urls.csv**) and the output JSON file (**scraped\_data.json**).
4. An empty list called **scraped\_data** is created to store the scraped data.
5. The input CSV file is opened using a context manager. The contents of the file are read using the **csv.DictReader** class, and the URLs from the 'URL' column are extracted into a list called **urls**.
6. The code then iterates over each URL in the **urls** list. For each URL, it attempts to scrape the company details by calling the **scrape\_company\_details** function and passing the URL as an argument. The scraped data is appended to the **scraped\_data** list.
7. If an error occurs while scraping data from a URL, an exception is caught and an error message is printed.
8. Once all the URLs have been processed, the code attempts to save the scraped data to a JSON file using the **json.dump** function. The data is written to the **scraped\_data.json** file with an indentation of 4 spaces.
9. The **main** function is run using **asyncio.run()** to execute the asynchronous code.

**Step for running code**

Folder contain 4 files.

1. g2crowd\_urls.csv
2. code details
3. output.json
4. scrape\_g2crowd.py
5. requirements.txt

1. g2crowd\_urls.csv file contain urls of g2 crowd websites.

2. code details file have all information about code.

3. output.json is for storing the scraped data from web in json format.

4. Main code present in scrape\_g2crowd.py file. For running this file open in vs code. Before the running this file install all requirements which are contain in requirements.txt file.

**Result**

I apologies, I tried my best for result. I cannot get where I’m wrong.

Code is giving following errors.

An error occurred while scraping company details: Protocol error (Page.navigate): Cannot navigate to invalid URL

=========================== logs ===========================

navigating to "URL", waiting until "load"

============================================================

An error occurred while scraping company details: Timeout 10000ms exceeded.

=========================== logs ===========================

waiting for locator("xpath=//div[contains(@class, \"css-u003cdiv\")]") to be visible

============================================================

An error occurred while scraping company details: Timeout 10000ms exceeded.

=========================== logs ===========================

waiting for locator("xpath=//div[contains(@class, \"css-u003cdiv\")]") to be visible

============================================================

An error occurred while scraping company details: Timeout 10000ms exceeded.

=========================== logs ===========================

waiting for locator("xpath=//div[contains(@class, \"css-u003cdiv\")]") to be visible

============================================================

An error occurred while scraping company details: Timeout 10000ms exceeded.

=========================== logs ===========================

waiting for locator("xpath=//div[contains(@class, \"css-u003cdiv\")]") to be visible

============================================================

An error occurred while scraping company details: Timeout 10000ms exceeded.

=========================== logs ===========================

waiting for locator("xpath=//div[contains(@class, \"css-u003cdiv\")]") to be visible

============================================================

An error occurred while scraping company details: Timeout 10000ms exceeded.

=========================== logs ===========================

waiting for locator("xpath=//div[contains(@class, \"css-u003cdiv\")]") to be visible

============================================================

Scraping completed and data saved to output.json

Top of Form