Documentation

Jagadeep Boora

1.

Explanation of the Scraping Approach:

1. Libraries Used: BeautifulSoup (bs4) is used by the code to parse HTML text and extract pertinent information. queries are used to make HTTP queries, and pandas are used to manipulate data. Furthermore, CSV files and regular expressions are handled by re-libraries and CSV, respectively.
2. User-Agent and Headers: To prevent the website from blocking the scraping process, the code configures the proper User-Agent and Accept-Language headers to resemble a genuine browser request.
3. Data Retrieval Function: A product's URL is used by the return_data() function to retrieve product information. The product name, picture link, cost, rating, description, review title, and review content are all retrieved. The review paragraph's "READ MORE" wording has also been cleaned up.
4. Product Link Fetching: The supplied URLs are retrieved using the get_data() function. After that, iterating through each link, it gets the HTML content and uses the return_data() function to extract the product data.
5. Managing Pagination: To manage pagination, the code recognizes and navigates across several product listing pages. Every page has a product link, which it gathers and uses to extract data.

   Note: The code written has tuned in such way that we can extract more than 40 products (this is the number of products in a single page). The links of the pages are being stored for later usage. We can specify the exact number of products required which makes the code reusable for multiple methods or links.

6. Writing Data to CSV: Finally, the extracted product data is written to CSV files with the specified file names.

The organized nature of CSV data, which offers clearly defined fields like Product Name, Price, and Rating and makes processing and analysis simple, makes them ideal for fine-tuning models. Model training pipelines are made simpler by CSV's tabular representation, which allows for seamless integration into a variety of machine learning frameworks. Moreover, the data preparation process may be streamlined by loading CSV data into pandas DataFrames with ease for preprocessing operations like feature engineering and data cleaning. Furthermore, because CSV is widely used as a format for exchanging data, it is accessible from a variety of tools and programming languages, which encourages experimentation and cooperation in the creation and assessment of models.

2. Challenges faced: I encountered a number of difficulties when trying to integrate the data cleaning processes into the code blocks in charge of data gathering. It was challenging to preserve code readability and clarity while ensuring a smooth integration of the cleaning process with data retrieval. Another set of difficulties was structuring the code structure to maximize maintainability and efficiency because the way code segments are arranged has a big impact on overall performance and ease of maintenance.

3. Tools and Libraries used:

1. BeautifulSoup (**bs4**): Used for parsing HTML content and extracting data from web pages.
2. requests: Used for making HTTP requests to fetch web pages.
3. pandas: Used for data manipulation and analysis. (removed later)
4. CSV: Used for reading from and writing to CSV files.
5. re: Used for working with regular expressions.

Ethical Considerations:

In order to adhere to the terms of service of the website and take legal considerations into account while scraping, I made sure that I comprehended the instructions included in the robots.txt file. The aim of the scraper was to make it transparent to the website's servers by configuring it with a valid User-Agent string in HTTP headers. In order to prevent server overload, rate-limiting techniques regulated the number of requests and gave priority to handling HTTP status codes, particularly 503 faults. The scraper adhered to ethical data usage guidelines and respected pagination structures. Frequent evaluations of the scraping procedure ensured continued adherence to the terms of service on the website and regulatory obligations, promoting openness and responsibility throughout the scraping endeavours.