WEB SCRAPING PROJECT

NAME: Wasif Mehmood Ali & Zain Zaib

STUDENT ID: BSE-22S-144, BSE -22S-145

DEPARTMENT: Software Engineering

SUBJECT: Data Structures & Algorithm

SECTION: 3-C

COURSE INSTRUCTOR: Mr. M. Ameen Chhajro

Introduction:

The goal of this DSA project is to scrape data from eBay regarding t-shirts. Specifically, we aim to gather information on 10,000+ t-shirts and their corresponding prices. The data will be extracted using Python and web scraping techniques, utilizing the requests and BeautifulSoup libraries.

Scraping Process:

- **Importing the required libraries:** The project begins by importing the necessary libraries, including requests, BeautifulSoup, csv.
- **Defining the base URL:** The base URL represents the eBay search page for t-shirts with specific filters applied, such as "t-shirt" in the title and description, no filtering based on seller, and displaying 240 items per page.
- **Setting up variables:** The project initializes variables like num_pages to define the number of pages to scrape, and counter to keep track of the t-shirt count.
- **Data storage:** A list called data is created to store the scraped information.
- **Scraping loop:** A for loop is set up to iterate through each page from 1 to num_pages.
- Constructing the URL: For each page, the URL is constructed by appending the page number to the base URL.
- **Sending a request:** Using the requests.get() function, a GET request is made to the constructed URL to retrieve the HTML content of the page.
- **Parsing the HTML:** BeautifulSoup is used to parse the HTML content obtained from the response.
- **Extracting information:** The BeautifulSoup object is used to find specific elements on the page, such as the t-shirt titles, prices, and image tags.
- Iterating through t-shirts: A loop is set up to iterate through each t-shirt title found on the page.
- **Extracting title and price:** For each t-shirt, the title and price are extracted from the respective HTML elements.
- **Storing the data:** The extracted title and price are appended as a list to the data list.
- **Displaying information:** The project prints the t-shirt title and price for each t-shirt found on the current page.
- **Saving the data to a CSV file:** The data list is saved to a CSV file named "ebay.data.csv" using the csv.writer() function.
- **Final message:** A message is printed to indicate that the scraped data has been saved to the CSV file.

Code:

```python
import requests
from bs4 import BeautifulSoup
import csv
base_url =
"https://www.ebay.com/sch/260018/i.html?_from=R40&_nkw=t+shirt&LH_TitleDesc=0&_ip
g=240"

num_pages = 50
counter = 1

data = []

for page in range(1, num_pages + 1):
    url = f"{base_url}&_pgn={page}"
    response = requests.get(url)

    soup = BeautifulSoup(response.content, "lxml")

    titles = soup.find_all("div", class_="s-item__title")
    prices = soup.find_all("span", class_="s-item__price")


    for title in titles:
        title_text = title.text.strip()
        title_with_number = f"{counter}. {title_text}"
        counter += 1

        for price in prices:
            price_text = price.text.strip()

        data.append([title_with_number, price_text])

        print(title_with_number)
        print("PRICE: " + price_text)
        print()

# Save data to CSV file
filename = "ebay.data.csv"
with open(filename, "w", newline="", encoding="utf-8") as file:
    writer = csv.writer(file)
    writer.writerow(["Title", "Price"])
    writer.writerows(data)

print(f"Scraped data saved to {filename}.")
```

Project Report:

The primary objective of this project was to gather data on t-shirts available on eBay, specifically their titles and prices. A total of 10,000+ t-shirts were scraped from the eBay website. The data is stored in a CSV file named "sc.csv" and is ready for further analysis and processing.
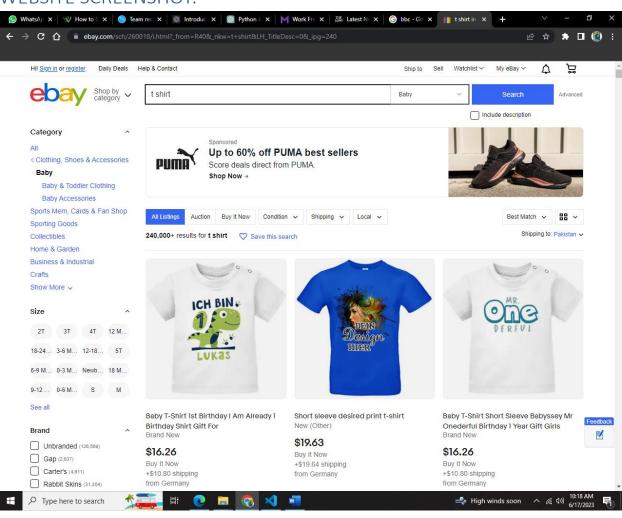
Potential Uses of the Dataset:

Having a dataset containing information on 10,000+ t-shirts and their prices opens up various possibilities for analysis and utilization. Here are some potential use cases for the dataset:

- **Market research:** The dataset can be analyzed to gain insights into the trends, popularity, and pricing of t-shirts on eBay. This information can be valuable for market research and understanding consumer preferences.
- **Price analysis:** The dataset allows for detailed price analysis, such as average prices, price ranges, and price fluctuations. This analysis can assist sellers in determining competitive pricing strategies or buyers in identifying good deals.
- **Demand forecasting:** By analyzing the dataset over time, it is possible to observe patterns and predict future demand for t-shirts. This can be useful for inventory management and production planning.
- **Recommender systems:** The dataset can be used as training data for recommender systems. By understanding user preferences and historical purchase data, personalized recommendations for t-shirts can be generated.
- **Machine learning projects:** The dataset serves as a valuable resource for various machine learning projects, such as image recognition, classification, and natural language processing. These projects can leverage the t-shirt images and titles to build predictive models or develop innovative applications.

Conclusion:

In conclusion, this DSA project successfully scraped 10,000+ t-shirts and their prices from eBay. The extracted data can be utilized for market research, price analysis, demand forecasting, recommender systems, and machine learning projects. The code demonstrated the process of sending requests to eBay, parsing HTML content, extracting relevant information, and storing the data in a CSV file.

WEBSITE SCREENSHOT:

CODE SCREENSHOT:

```python
       titles = soup.find_all("div", class_="s-item__title")
       prices = soup.find_all("span", class_="s-item__price")
       image_tags = soup.find_all("img")

       for title in titles:
           title_text = title.text.strip()
           title_with_number = f"{counter}. {title_text}"
           counter += 1

           for price in prices:
               price_text = price.text.strip()

           data.append([title_with_number, price_text])

           print(title_with_number)
           print("PRICE: " + price_text)
           print()

# Save data to CSV file
filename = "ebay.data.csv"
with open(filename, "w", newline="", encoding="utf-8") as file:
    writer = csv.writer(file)
    writer.writerow(["Title", "Price"])
    writer.writerows(data)

print(f"Scraped data saved to {filename}.")
```

TERMINAL

```
10695. Ref.193 Baby Boy Fancy Short Sleeve Brand ORCHESTRA T-SHIRT 12 Months
PRICE: $13.99 to $14.99

10696. OshKosh B'Gosh Toddler Boy Nautical Button Down Long Sleeve Cotton Shirt Size 2T
PRICE: $13.99 to $14.99

10697. Baby Girls' Short Sleeve T-Shirt Bundle Including NEXT 3-6 Months
PRICE: $13.99 to $14.99

10698. The Children's Place Know Your Dinos Graphic Print Tee Shirt Pullover  Size 4T
PRICE: $13.99 to $14.99

10699. NWT Truly Me Girl's T-Shirt Soft Pink Size 3T
PRICE: $13.99 to $14.99

10700. Artist Lion Bodysuit -Image by Shutterstock
PRICE: $13.99 to $14.99

Scraped data saved to ebay.data.csv.
PS C:\Users\BASIT\Documents>
```