**Information Retrieval Project**

Amazon Scrapper

BY

|  |  |
| --- | --- |
| **Name** | **ID** |
| Mohab Ashraf | 210439 |
| Ahmed Hossam | 213581 |
| Yousef Mohamed | 211777 |

Faculty of Computer Science

Group B

Dr. Moataz Samy

May 15, 2023

* **Introduction**

In this paper we document in detail how our scraper functions with code snippets. Our the scraper was built for the Amazon website, using Selenium 4.0, an automation tool, and Custom Tkinter as for an interactive GUI, the main focus of this project was to learn the intricate methods of collecting available online data from B2C online ecommerce websites. The data covers products based on a user keyword through the interface.

* **Brief Summary**

Web scraping is the process of extracting data from websites using a bot or web crawler. It is a form of data collection in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later analysis.

* **Explanation**

Firstly, we choose a specific Python library called Selenium. This library package consists of a toolset that allows you to automate web browser interaction. A CSV was chosen to import the collected data into.

Starting up, building the GUI interface was simple, as CTK provides u developer friendly toolkit.

A screen shot of a computer

Description automatically generated with medium confidence

A picture containing text, screenshot, font

Description automatically generated

The “scrape\_button” calls the “Scrape\_Amazon” function, which in turn, returns a list of data.

A screen shot of a computer code

Description automatically generated with low confidence

The sub-tool used in parsing, is the common “WebDriver” toolkit through a “Chrome” browser. Which is best known for its well maneuverability and evasion of captchas, adding a time.sleep() of 1 second and using direct “Headless” mode, which creates a Chrome WebDriver instance to help with going undetected and parse faster instead of rendering the browser GUI each time, even though the products’ page was conveniently static HTML.

We accessed the products page directly by embedding a string variable “keyword” into the domain link directly. Using the “.get()” function, we receive the user input string directly.

Then, Products elements are located using CSS Selector and by Iterating over each product element to extract relevant information such as product page URL, product name, company name, price, rating, and product image, after gathering the relevant information, the information is extracted and stored in a List, this List is called “Products”.

A screen shot of a computer program

Description automatically generated with low confidence

Any Exceptions or Failed Scrapes during the scraping process, will be counted by a Counter “ct” and displayed along with the scraped products, and the WebDriver Instance.

A screen shot of a computer program

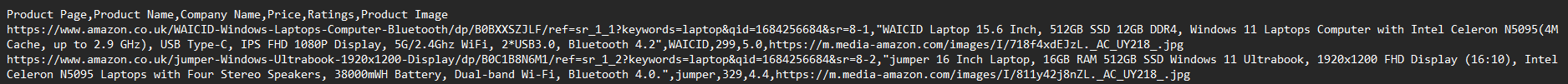
Description automatically generated with low confidence

Finally, a CSV File is creating afterwards to be used to store the relevant information retrieved by the Amazon Scraper through the “Scrape Amazon” function.

A picture containing text, screenshot, font

Description automatically generated

CSV file Structure:



* **Conclusion**

To conclude, the realm of technology will always advance as information becomes the key to discovering how we interact. Providing various needs to access information faster and generate a summary is in much need to the industry, we’ve learned a lot through this project, alongside the ethical means we have to follow.