**By**

**Md Atikujjaman (ATI21508718)**

Yamin Hossan

Kamrul Islam

Khusal khan

Submitted to

**The University of Roehampton**

**Data Engineering**

**CMP020X304S**

Abstract

This brief study describes a web scraping technique used to obtain data about data engineering books from Amazon.co.uk. This course covers the techniques and moral aspects associated with scraping, data processing, and storage.

Declaration

we hereby certify that this report constitutes our work, that where the language of others is used, quotation marks so indicate, and that appropriate credit is given where we have used the language, ideas, expressions, or writings of others.

we declare that this report describes the original work that has not been previously presented for the award of any other degree of any other institution.

Md Atikujjaman , Yamin Hossain, Kamrul Islam, Khusal khan

**Date:** 14/03/2024

**Table of Contents**

**Task: 1**

* 1. **. Introduction:** ..................................................................................Page5

**1.2Methodology** ......................................................................... Page 5- 6

**1**.**3. Discussion and Conclusion:** ..................................................................Page 6

**Task:2**

**2.1. Objective**..............................................................................................

**2.2Methodology** ......................................................................................vii

**3**.**3. Rationale:………………..** ......................................................................vii

**Task:3**

**2.1. Introduction:** ..............................................................................................

**2.2Methodology** ......................................................................................vii

**3**.**3. Discussion and Conclusion:** ......................................................................vii

|  |
| --- |
| **TASK 1** |

**1.1: Introduction:** Web data extraction is essential to many fields, such as market research and data science. For academic purposes, this project demonstrates the extraction of structured data from an unstructured online environment using web scraping to gather data on data engineering books from Amazon.co.uk.

**1.2:Methodology:**

**collecting the webpage & Webpage HTML Parsing:**   
  
**Tools Used:**  
**Requests Library:** For submitting GET requests via HTTP to URLs.   
**BeautifulSoup (bs4 Library):** For content analysis and parsing in HTML.  
**Process:**   
**HTTP GET Request:** To get around limitations on scraping, this request is sent with a User-Agent header to mimic a browser.   
**HTML Parsing:** Utilised BeautifulSoup to get information from certain HTML elements by filtering them depending on their characteristics.

**Screenshots:**

A screenshot of a computer

Description automatically generated

**Data Extraction:**   
  
**Process:** The BeautifulSoup object was used to get data like title, author, date, price, and star rating. Missing data was carefully handled to make sure the data was correct and complete.

**Screenshot:**A screenshot of a computer

Description automatically generated

**Creating a Data Frame and Storing Data:**   
  
**Tool Used:** pandas’ library.

**Process:** After the data was retrieved, it was sorted into a list of tuples, converted into a Data Framework for more efficient data handling, and saved as a CSV file.

**Screenshot:**

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

**1.3: Discussion and Conclusion:** This technique emphasizes the significance of ethical online scraping processes, which include following terms of service and considering server load. By using a methodical method for extracting and storing data, it is possible to get important datasets from internet sources, which may enhance academic study and analysis.

|  |
| --- |
| **TASK 2** |

**2.1: Objective:**

To convert extracted e-commerce data into a PostgreSQL database, allowing for complex data manipulation.

**2.2: Methodology:**

**PostgreSQL Installation:** The initial step involves installing PostgreSQL to provide an efficient basis to handle relational data.

**Screenshot: A black screen with white text

Description automatically generated**

**A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated**

**Data Transfer:**

The Python script uses the psycopg2 package to automate the transfer of scraped e-commerce data into a PostgreSQL database. After establishing a connection to the server, it processes the CSV file, inserts each book entry, and configures the database and table structure. This technique ensures that raw data will be efficiently transformed into a structured database format, allowing for in-depth querying and further analysis. Data integrity and the reliability of the transfer process are ensured by the script's error handling.

**Screenshot;**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A white screen with black text

Description automatically generated**

A screenshot of a computer

Description automatically generated

**Database design Design:** A clean and well-organized data structure was ensured by carefully outlining the database design using draw.io.

**Screenshot:A diagram of a data flow

Description automatically generated**

**SQL Query Execution:**   
A SQL query in PostgreSQL fetched titles with top ratings from the books table, sorted by date, to showcase efficient data sorting and retrieval capabilities.

**Screenshot:**

**A computer screen shot of a black screen

Description automatically generatedA screenshot of a computer

Description automatically generated**

**2.3: Rationale:**

**Python Integration:** The need for programmable database operations, which expedited the data import process, led to the use of Python in conjunction with the psycopg2 package.

**Schema Visualisation:** Verifying data integrity and elucidating the links among data in the database required schema visualization.

**SQL Efficiency:** Sorting data with a SQL query highlighted PostgreSQL's robust sorting and retrieval capabilities, essential for data analysis and insight gathering.