A

SYNOPSIS

of

MINOR PROJECT

on

Web Scraping

Submitted by

Rajveer Singh Dhabhai (21EGICS090)

Project Guide
Ms. Charu Kavadia

Head of Department
Dr. Mayank Patel

Geetanjali Institute of Technical Studies, Dabok , Udaipur (Raj.)
Department of Computer Science and Engineering
October,2023

Contents

Problem Statement

To develop a Python script for web scraping product data (names and prices) from Amazon and saving it as an Excel file.

Brief Description

The project involves creating a Python script that utilizes BeautifulSoup for parsing Amazon's HTML content, extracting product names and prices, organizing them into a Pandas DataFrame, and finally exporting the data to an Excel file for easy analysis and comparison.

Objective and Scope

Objective:

- Automate the extraction of product information from Amazon.
- Store the data in a structured format for further analysis or integration into other applications.

Scope:

- fetch product names and prices from a specified Amazon page.
- Develop a Python script capable of handling typical HTML structures of Amazon product pages.
- Handle basic error scenarios (e.g., connection issues, missing elements).
- Save data locally as an Excel file for user convenience.

Methodology

- **Scraping**: Utilize BeautifulSoup library to navigate and extract data from Amazon's HTML documents efficiently.

- **Data Handling**: Employ pandas DataFrame to structure and manipulate the extracted data, ensuring clarity and ease of use.

- **Error Handling**: Implement mechanisms to detect and manage errors, ensuring script reliability and resilience.

- **Export**: Utilize pandas' built-in functionality to export the structured data into Excel format, maintaining data integrity and usability.

Hardware and Software Requirements

- **Hardware Requirements:**
  - Standard computer system with adequate processing power and storage.
  - RAM: 512 MB (**RECOMMENDED 2GB**)
  - HARD DRIVE: MINIMUM (2 GB FREE SPACE)
  - PROCESSOR: ANY WORKING WITH WINDOWS 8 AND ABOVE (**Intel-i3/Ryzen 5300U +)**
  - Reliable internet connection for accessing and scraping Amazon's website.

- **Software Requirements:**
  - Python (version 3.6 or higher) **installed with necessary libraries: BeautifulSoup, pandas.**
  - Web browser for testing and validation purposes.
  - **Excel** software for viewing and analysing the exported data.

Technologies

- **Python**: A versatile programming language chosen for its robust libraries and ease of scripting.

- **BeautifulSoup**: A Python library for parsing HTML and XML documents, essential for extracting data from complex web pages like Amazon.

- **pandas**: A powerful data analysis and manipulation library for Python, ideal for organizing scraped data into structured formats.

- **Excel**: Widely used spreadsheet software for viewing, analysing, and manipulating structured data exported from Python scripts.

Testing Techniques

- **Unit Testing**: Validate individual functions and components of the script to ensure they perform as expected.

- **Integration Testing**: Verify the script's functionality as a whole, simulating different scenarios and edge cases to detect and resolve potential issues.

- **User Acceptance Testing (UAT)**: Involve end-users or stakeholders to validate the script's usability and effectiveness in meeting project requirements.

Project Contribution

- **Personal Contribution**: Designed and implemented the Python script for web scraping, data extraction, and Excel export functionality.

- **Impact**: Facilitated easy access to structured Amazon product data for potential use in price comparison, market analysis, or other applications.