TASK REPORT

ON

WEB SCRAPING AND DATASET CREATION

Intern Name : Ganesh Wagh

Internship Domain : Data Analysis / Web Scraping

Organization : Codealpha

Task Name : Task 1 – Web Scraping

Tool Used: Octoparse

Duration : 10 January 2026 – 24 February 2026

Abstract

This report presents the implementation of a web scraping task carried out during the internship. The objective of the task was to extract publicly available data from a website and convert it into a structured dataset suitable for analysis. Automated web scraping was performed using a no-code tool, and the collected data was exported in CSV format. The task demonstrates practical understanding of automated data collection, pagination handling, and dataset creation.

Introduction

In the digital era, large volumes of valuable data are available on websites in unstructured formats. To make this data usable for analysis, it must be extracted and organized systematically. Web scraping is a technique used to automatically collect such data from web pages.
This task focuses on using automated web scraping tools to gather publicly available information from a website and convert it into a structured dataset. Web scraping plays an important role in data analytics, market research, and business intelligence.

Objectives

- The objectives of this task are:
- To understand the concept of web scraping
- To extract data from a publicly accessible website
- To learn automated data collection using no-code tools
- To create a clean and structured dataset
- To document the complete task in a professional manner

Scope of Work

The scope of this task includes:

- Selection of a public demo website
- Extraction of predefined data fields
- Handling pagination for multi-page data
- Exporting data into CSV format

The task does not include database storage, data analysis, or visualization.

Tools and Technologies Used

Tools	Purpose
Octoparse	Automated web scraping
Google Chrome	Website access and inspection
Microsoft Excel	Dataset verification

Website Selection

The following public website was selected for the task:
Website: https://books.toscrape.com

This website is designed for practicing web scraping and provides structured book-related information such as title, price, and availability without requiring login credentials.

Methodology

The task was completed using the following systematic approach:

Step 1: Task Creation
A new custom task was created in Octoparse by entering the website URL
.
Step 2: Automatic Data Detection
Octoparse automatically detected repeating data elements such as book titles, prices, and stock status.

Step 3: Field Selection
The following fields were selected:
Book Title
Price
Stock Availability

Step 4: Pagination Handling
Pagination was enabled to extract data from all pages.

Step 5: Data Extraction
Local extraction mode was used. The process completed successfully without duplicate records.

Step 6: Dataset Export
Extracted data was exported in CSV format.

Screenshots

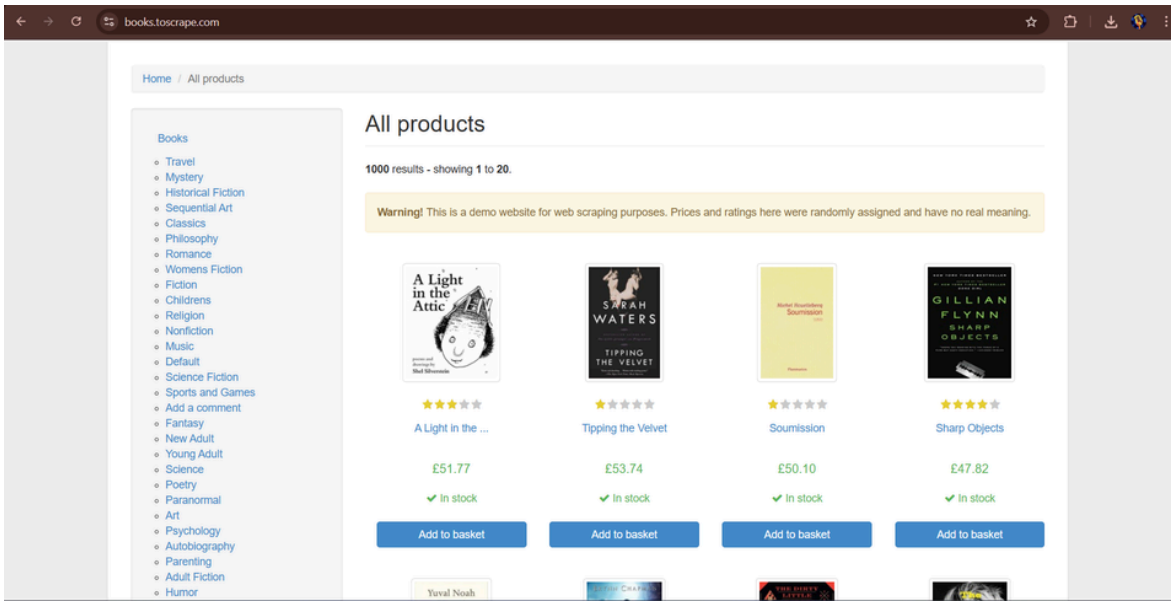

Figure 1: Selected target website for web scraping (books.toscrape.com

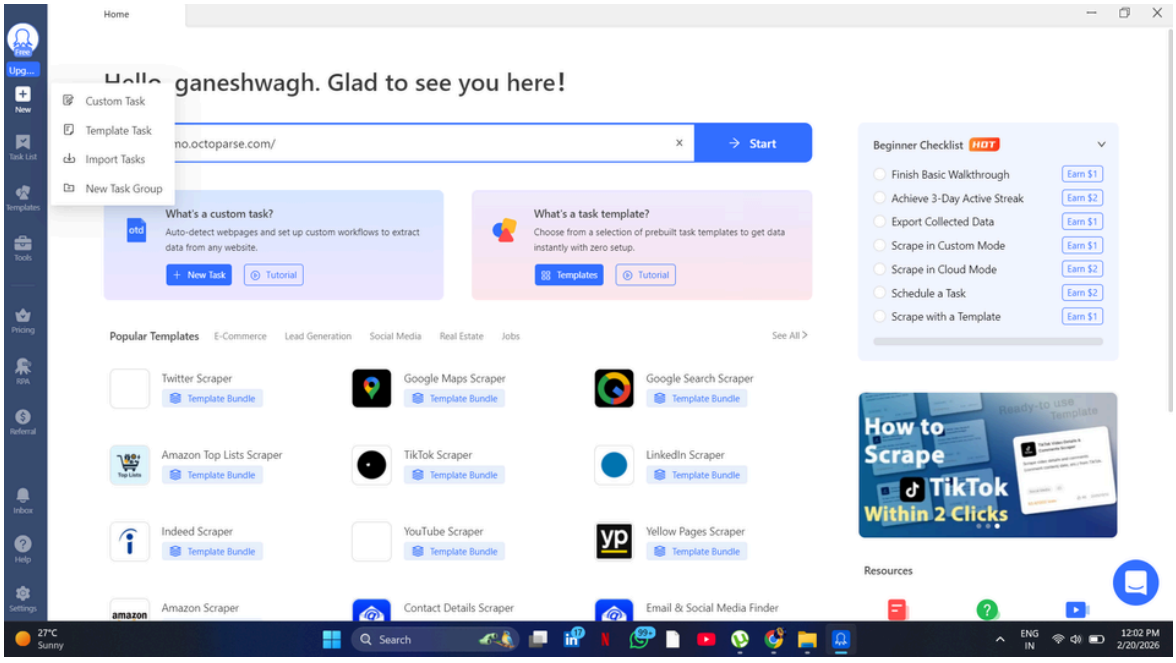

Figure 2: Creation of new web scraping task in Octoparse

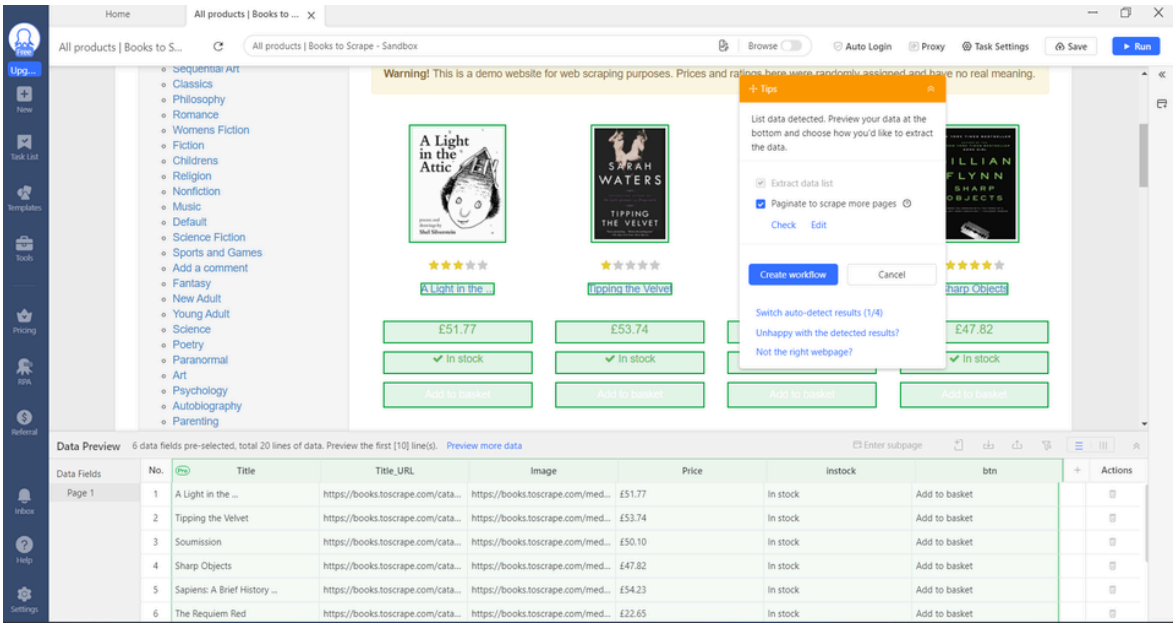

Figure 3: Automatic detection of data fields by Octoparse

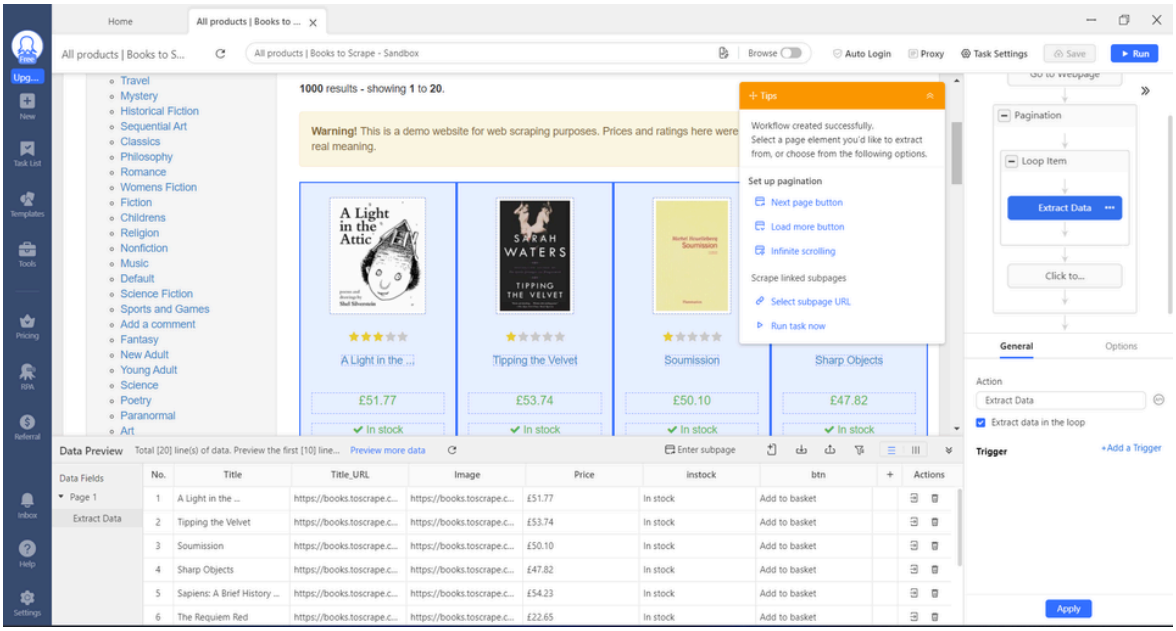

Figure 4: Creation of scraping workflow and pagination configuration

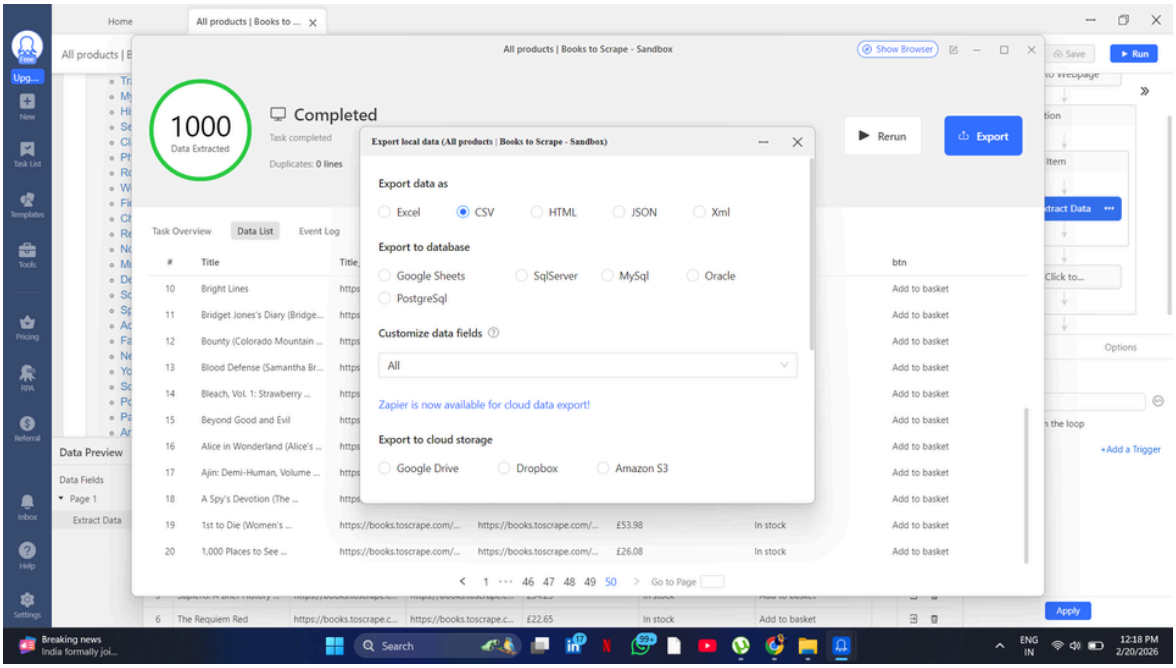

Figure 5: Preview of extracted data fields in Octoparse

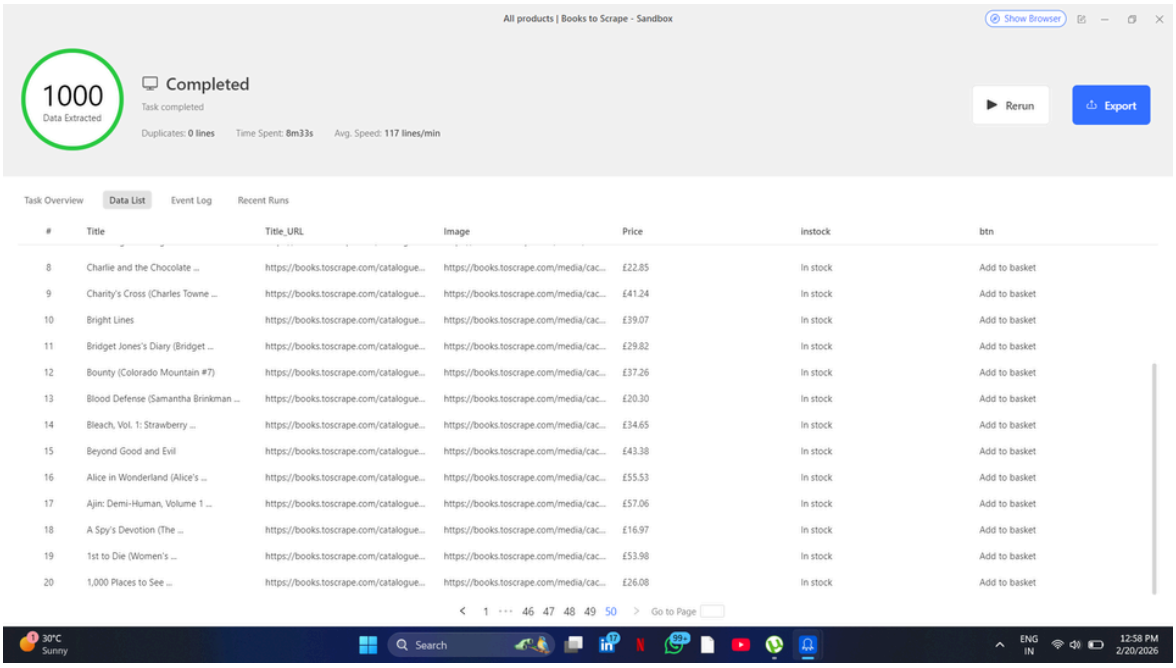

Figure 6: Successful execution of web scraping task with 1000 records extracted

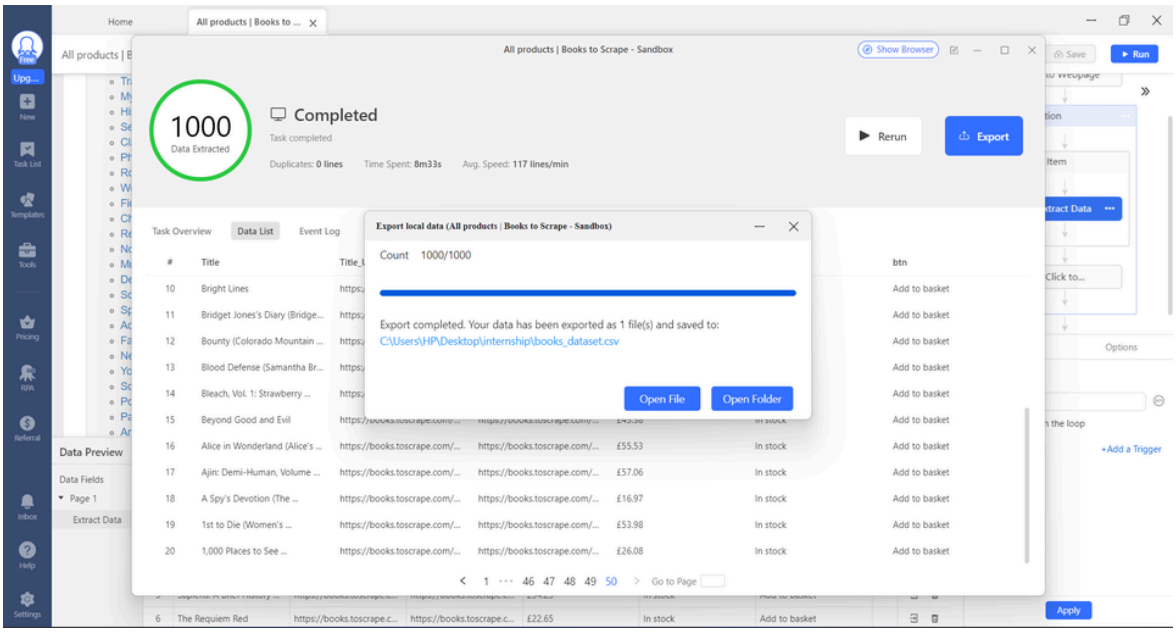

Figure 7: Exporting extracted data into CSV file format

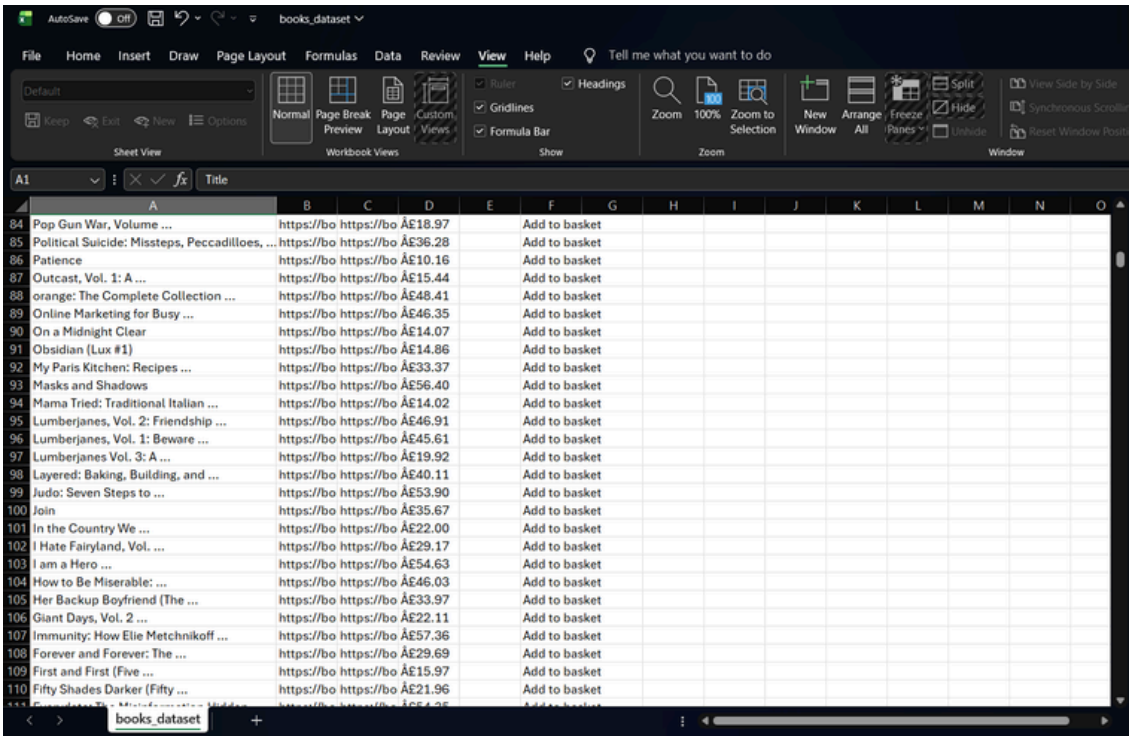

Figure 8: Confirmation of successful dataset export to local system

Dataset Description

The final dataset consists of 1000 records with the following fields:

Column Name	Description
Title	Name of the book
Title_URL	Link to the book page
Image	URL of the book image
Price	Price of the book
instock	Availability status
btn	Action button text

File Format: CSV
Duplicates: 0

Results and Observations

- Total records extracted: 1000
- Pagination handled successfully
- No duplicate data found
- The output dataset is clean, structured, and ready for further analysis