

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

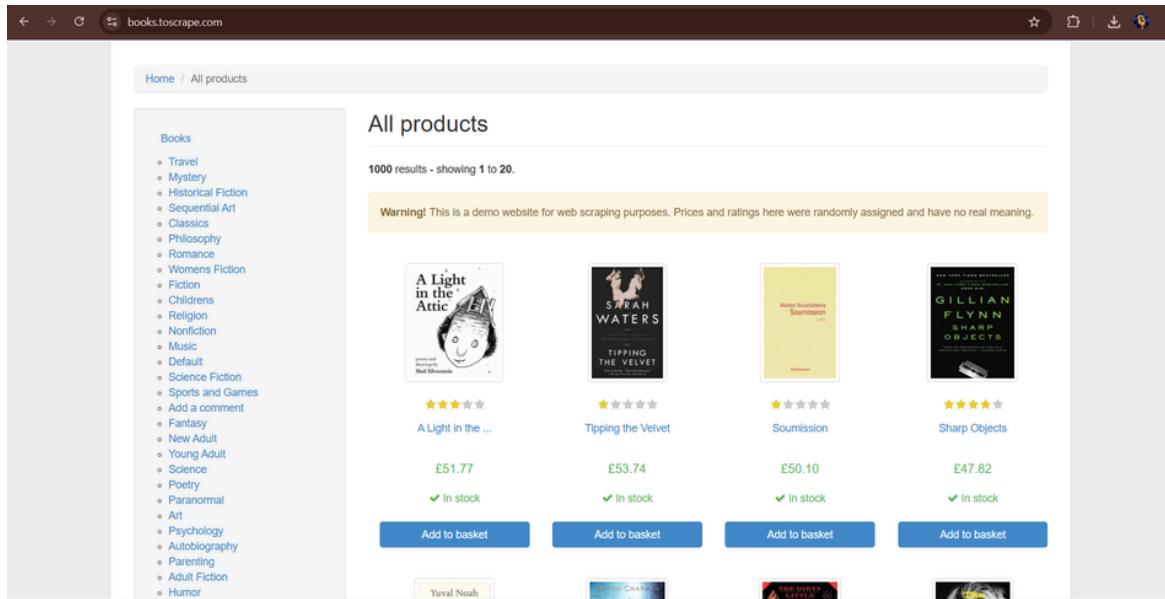


Figure 1: Selected target website for web scraping (books.toscrape.com)

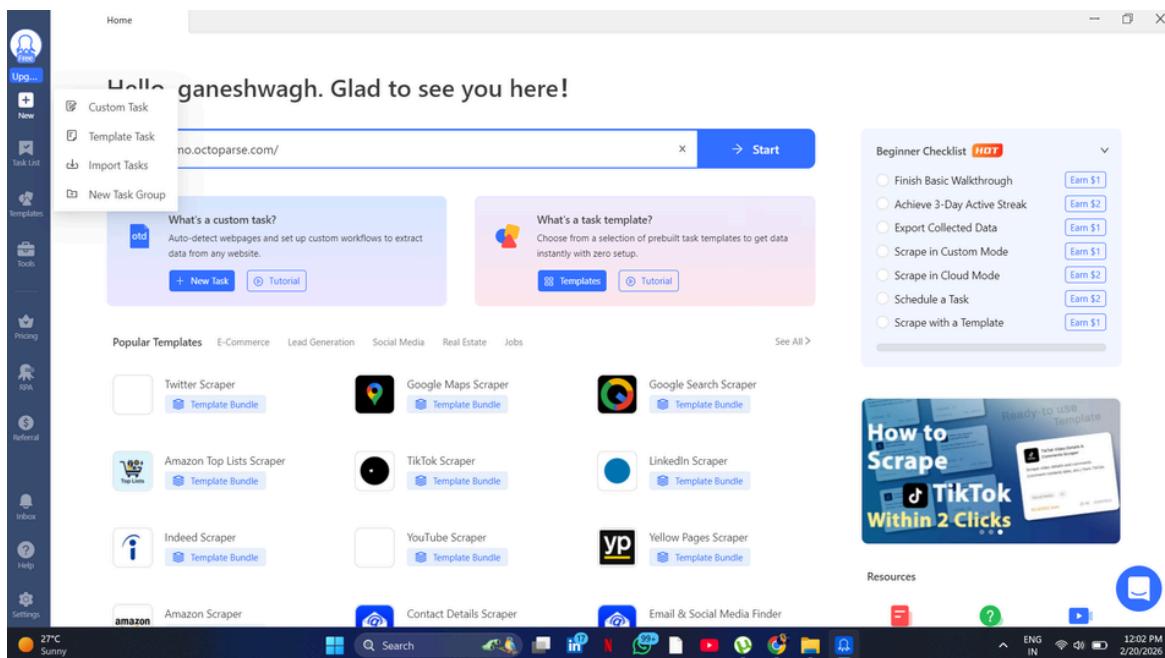


Figure 2: Creation of new web scraping task in Octoparse

No.	Title	Title_URL	Image	Price	instock	btn
1	A Light in the ...	https://books.toscrape.com/cata...		£51.77	In stock	Add to basket
2	Tipping the Velvet	https://books.toscrape.com/cata...		£53.74	In stock	Add to basket
3	Soumission	https://books.toscrape.com/cata...		£50.10	In stock	Add to basket
4	Sharp Objects	https://books.toscrape.com/cata...		£47.82	In stock	Add to basket
5	Sapiens: A Brief History ...	https://books.toscrape.com/cata...		£54.23	In stock	Add to basket
6	The Requiem Red	https://books.toscrape.com/cata...		£22.65	In stock	Add to basket

Figure 3: Automatic detection of data fields by Octoparse

Figure 4: Creation of scraping workflow and pagination configuration

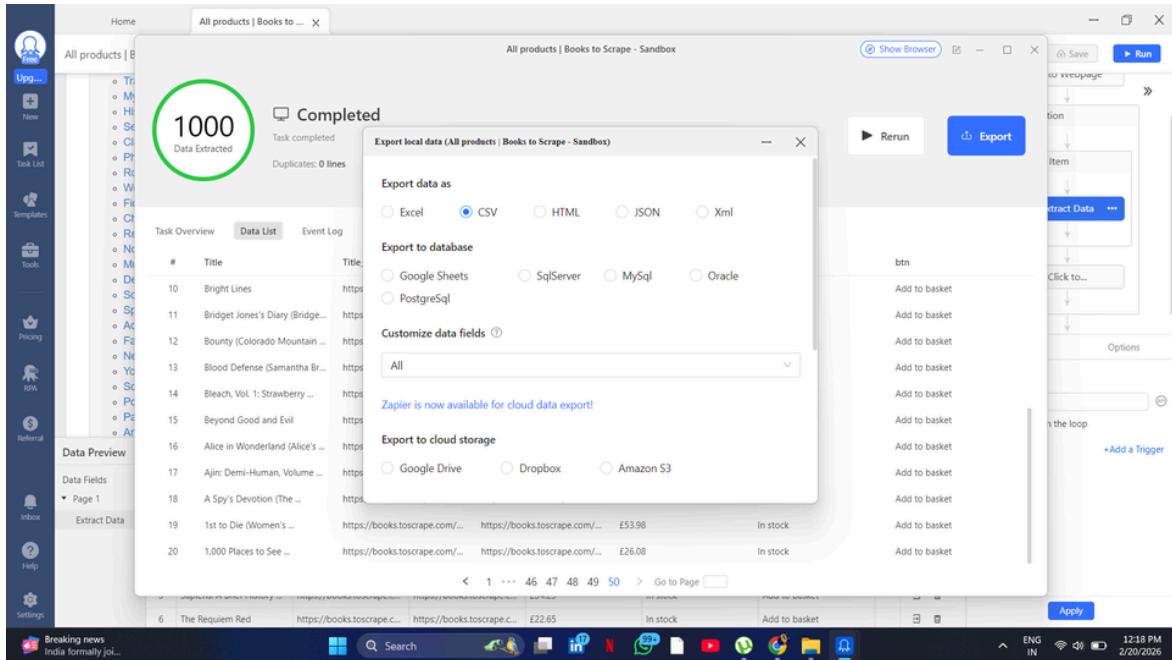


Figure 5: Preview of extracted data fields in Octoparse

#	Title	Title_URL	Image	Price	Instock	btn
8	Charlie and the Chocolate ...	https://books.toscrape.com/catalogue...	https://books.toscrape.com/media/cac...	£22.85	In stock	Add to basket
9	Charyn's Cross (Charles Towne ...	https://books.toscrape.com/catalogue...	https://books.toscrape.com/media/cac...	£41.24	In stock	Add to basket
10	Bright Lines	https://books.toscrape.com/catalogue...	https://books.toscrape.com/media/cac...	£39.07	In stock	Add to basket
11	Bridget Jones's Diary (Bridget ...	https://books.toscrape.com/catalogue...	https://books.toscrape.com/media/cac...	£29.82	In stock	Add to basket
12	Bounty (Colorado Mountain #7)	https://books.toscrape.com/catalogue...	https://books.toscrape.com/media/cac...	£37.26	In stock	Add to basket
13	Blood Defense (Samantha Brinkman ...	https://books.toscrape.com/catalogue...	https://books.toscrape.com/media/cac...	£20.30	In stock	Add to basket
14	Bleach, Vol. 1: Strawberry ...	https://books.toscrape.com/catalogue...	https://books.toscrape.com/media/cac...	£34.65	In stock	Add to basket
15	Beyond Good and Evil	https://books.toscrape.com/catalogue...	https://books.toscrape.com/media/cac...	£43.38	In stock	Add to basket
16	Alice in Wonderland (Alice's ...	https://books.toscrape.com/catalogue...	https://books.toscrape.com/media/cac...	£55.53	In stock	Add to basket
17	Ajir: Demi-Human, Volume 1 ...	https://books.toscrape.com/catalogue...	https://books.toscrape.com/media/cac...	£57.06	In stock	Add to basket
18	A Spy's Devotion (The ...	https://books.toscrape.com/catalogue...	https://books.toscrape.com/media/cac...	£16.97	In stock	Add to basket
19	1st to Die (Women's ...	https://books.toscrape.com/catalogue...	https://books.toscrape.com/media/cac...	£53.98	In stock	Add to basket
20	1,000 Places to See ...	https://books.toscrape.com/catalogue...	https://books.toscrape.com/media/cac...	£26.08	In stock	Add to basket

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

Figure 7: Exporting extracted data into CSV file format

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
84	Pop Gun War, Volume ...	https://bo	https://bo	£18.97	Add to basket										
85	Political Suicide: Missteps, Peccadilloes, ...	https://bo	https://bo	£36.28	Add to basket										
86	Patience	https://bo	https://bo	£10.16	Add to basket										
87	Outcast, Vol. 1: A ...	https://bo	https://bo	£15.44	Add to basket										
88	orange: The Complete Collection ...	https://bo	https://bo	£48.41	Add to basket										
89	Online Marketing for Busy ...	https://bo	https://bo	£46.35	Add to basket										
90	On a Midnight Clear	https://bo	https://bo	£14.07	Add to basket										
91	Obsidian (Lux #1)	https://bo	https://bo	£14.86	Add to basket										
92	My Paris Kitchen: Recipes ...	https://bo	https://bo	£33.37	Add to basket										
93	Masks and Shadows	https://bo	https://bo	£56.40	Add to basket										
94	Mama Tried: Traditional Italian ...	https://bo	https://bo	£14.02	Add to basket										
95	Lumberjanes, Vol. 2: Friendship ...	https://bo	https://bo	£46.91	Add to basket										
96	Lumberjanes, Vol. 1: Beware ...	https://bo	https://bo	£45.61	Add to basket										
97	Lumberjanes Vol. 3: A ...	https://bo	https://bo	£19.92	Add to basket										
98	Layered: Baking, Building, and ...	https://bo	https://bo	£40.11	Add to basket										
99	Judo: Seven Steps to ...	https://bo	https://bo	£53.90	Add to basket										
100	Join	https://bo	https://bo	£35.67	Add to basket										
101	In the Country We ...	https://bo	https://bo	£22.00	Add to basket										
102	I Hate Fairyland, Vol. ...	https://bo	https://bo	£29.17	Add to basket										
103	I am a Hero ...	https://bo	https://bo	£54.63	Add to basket										
104	How to Be Miserable: ...	https://bo	https://bo	£46.03	Add to basket										
105	Her Backup Boyfriend (The ...	https://bo	https://bo	£33.97	Add to basket										
106	Giant Days, Vol. 2 ...	https://bo	https://bo	£22.11	Add to basket										
107	Immunity: How Elie Metchnikoff ...	https://bo	https://bo	£57.36	Add to basket										
108	Forever and Forever: The ...	https://bo	https://bo	£29.69	Add to basket										
109	First and First (Five ...	https://bo	https://bo	£15.97	Add to basket										
110	Fifty Shades Darker (Fifty ...	https://bo	https://bo	£21.96	Add to basket										

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