

# Tools for Big Data

## Web scraping

[toscrape.com](https://toscrape.com) is a website offering two sites to practice web scraping:

- [Quotes to scrape](#), that lists quotes from famous people. It has many endpoints showing the quotes in many different ways, each of them including new scraping challenges.
- [Books to scrape](#), a fictional bookstore. It's a safe place for beginners learning web scraping and for developers validating their scraping technologies as well.

In this individual project, you will have to collect data from the Books to Scrape website.

## Context

The [Books to scrape](#) website contains books divided by a set of 50 categories. By clicking on a book, we can get more information about it (description, average score, availability, author, etc.).

**Books to Scrape** We love being scraped!

[Home](#) / [All products](#)

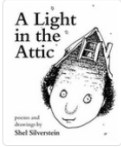
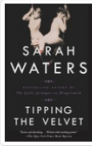

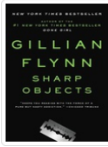
### Books

- Travel
- Mystery
- Historical Fiction
- Sequential Art
- Classics
- Philosophy
- Romance
- Womens Fiction
- Fiction
- Childrens
- Religion
- Nonfiction
- Music
- Default
- Science Fiction
- Sports and Games
- Add a comment
- Fantasy
- New Adult
- Young Adult
- Science
- Poetry
- Paranormal
- Art
- Psychology

### All products

1000 results - showing 1 to 20.

**Warning!** This is a demo website for web scraping purposes. Prices and ratings here were randomly assigned and have no real meaning.

 <p>★★★★★ A Light in the ... £51.77 ✓ In stock <a href="#">Add to basket</a></p>	 <p>★★★★★ Tipping the Velvet £53.74 ✓ In stock <a href="#">Add to basket</a></p>	 <p>★★★★★ Soumission £50.10 ✓ In stock <a href="#">Add to basket</a></p>	 <p>★★★★★ Sharp Objects £47.82 ✓ In stock <a href="#">Add to basket</a></p>
---	---	---	--

## Exam

- ☐ Choose 5 categories
- ☐ For each of these categories, navigate through the books and extract:
  - ☐ the book name
  - ☐ the book description
  - ☐ the book category
  - ☐ the author
  - ☐ the book price
  - ☐ its availability
  - ☐ the URL of the book image (src attribute in img tag)
  - ☐ the number of stars
- ☐ Transform data to a pandas DataFrame and save it as a CSV file.
- ☐ The webscraper should be unique (one webscraper for all books present in the five categories)
- ☐ Provide a **short** PDF document describing and explaining the general structure of a web page of a book (its HTML code and the HTML structure of the elements to be web scrapped)
- ☐ Provide the Jupyter Notebook

The end.