Web Scraper Project Algorithmic Analysis **Pech Aque, Damaris Esther**

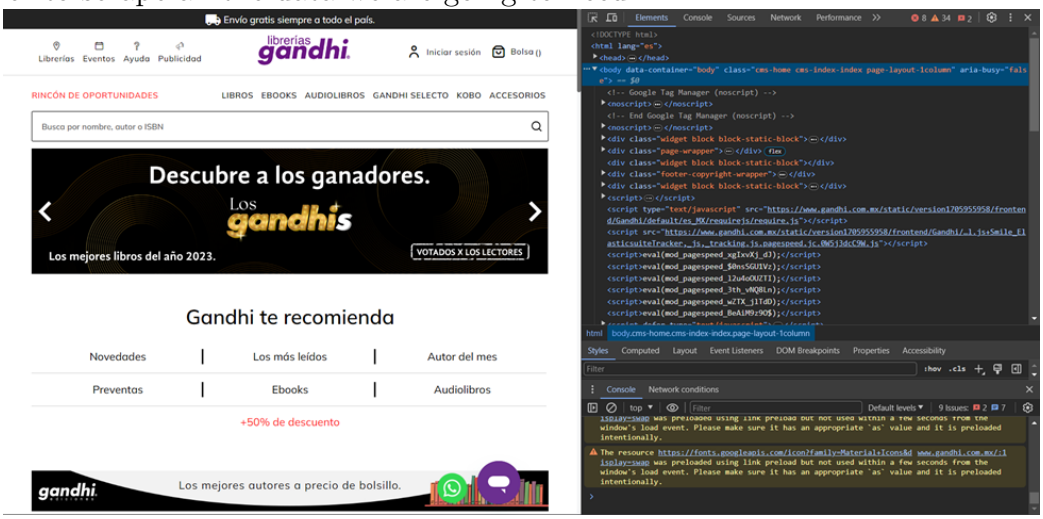
Lopez Tzec, Isaias de Jesus
Herrera Balam, Daniel Fernando
Paredes Dzib, Valeria de los Angeles
Velasco Martin, Jonathan Abisai

4th Year A - Data Engineering
Polytechnic University of Yucatán
Professor: Adrian Roberto Carmona Rodriguez
Date: January 22, 2024

Web Scraper

Preliminary Analysis of the Website

We start analyzing the website looking for what kind of data we are interested in, how the URL works for the research, and how the HTML is organized. Knowing all this information is going to make the development of the script easier to scrape all the data we are going to need.



Planning the Scraper

After analyzing the website, when you write a keyword for the research, a lot of information appears for books like name, author, price, kind of object (type of cover of the book).

Legal and Compliance Review

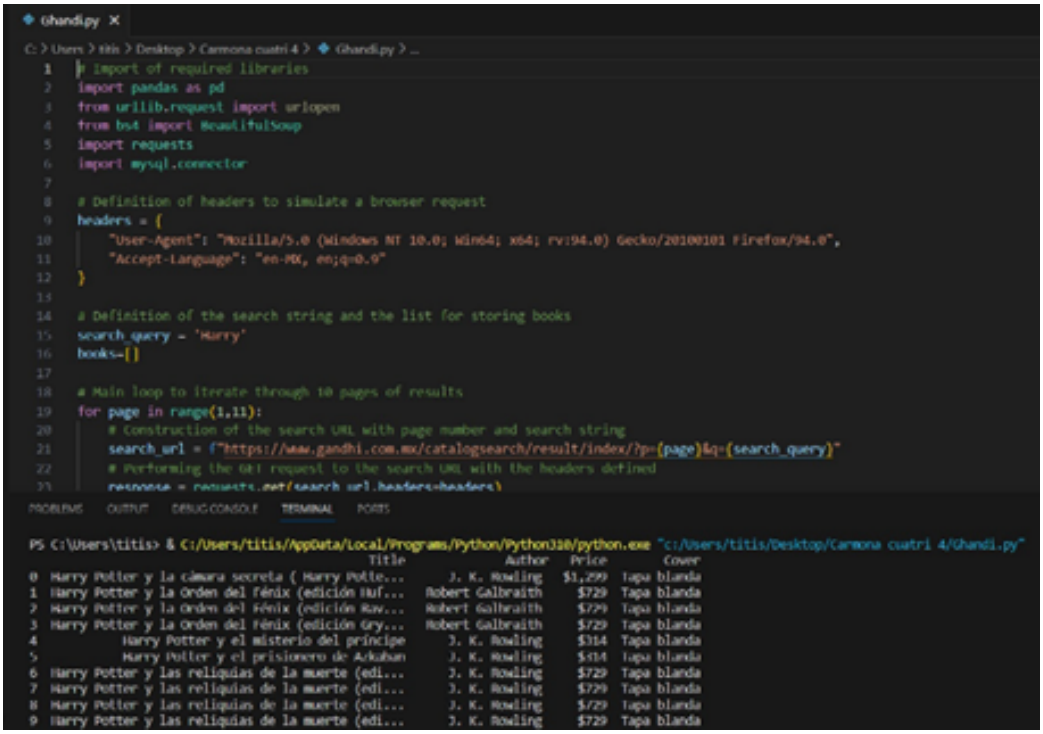
The web scraping project for the bookstore page underwent a comprehensive Legal and Compliance Review, ensuring adherence to ethical and legal standards.

Selection of Tools and Languages

We decided to use Python for the scraper because it is the base for our career and also one of the easiest languages for tasks like this.

Development of the Scraper

As mentioned, we use Python and a few libraries for our scraper. BS4 lets us work with the HTML of the page and manipulate it for what we want.

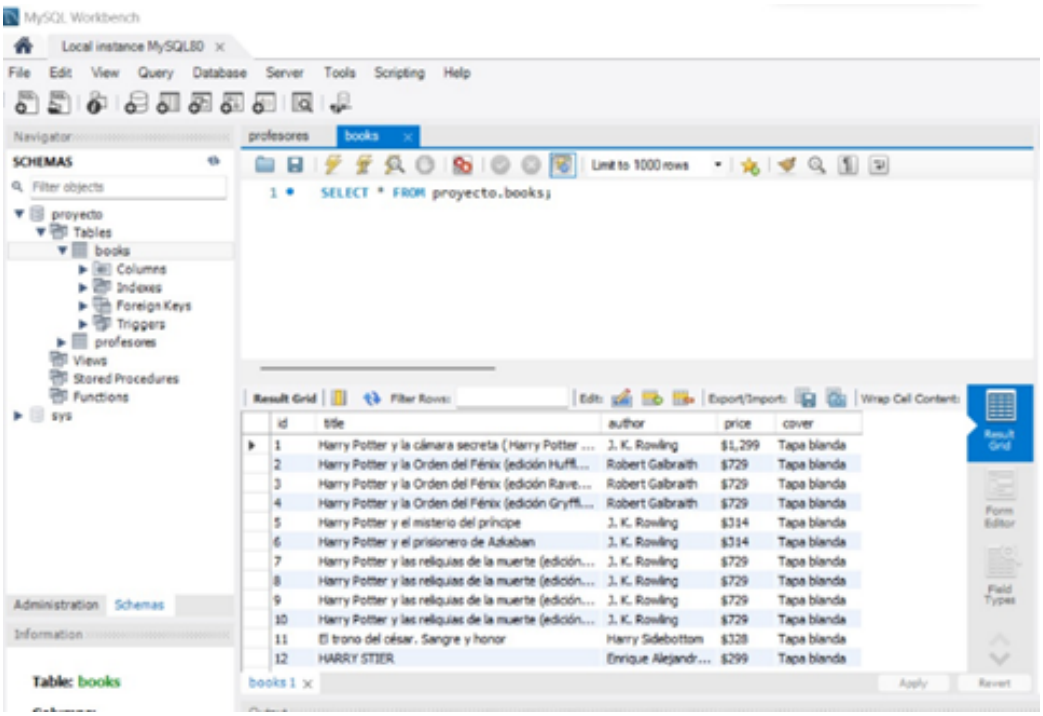


Efficiency and Performance Management

To enhance the efficiency and performance of the web scraping process, several measures were implemented.

Data Storage

For effective data storage, a structured approach was taken by utilizing a MySQL database.



Testing and Debugging

Thorough testing procedures were incorporated to ensure the correct collection of data during the scraping process.

Documentation and Maintenance

Comprehensive documentation practices were followed to provide clarity on both the code and the scraping process.

Implementation and Monitoring

For efficient deployment and monitoring, the web scraper is planned to be encapsulated within a Docker container.


docker