Report: Web Scraping and Data Collection

Full name: Жомартұлы Бекасыл
Student ID: 051209551500
Subject: Анализ и обработка веб данных

1. Project Objective

The goal of this project was to create a Python script that automatically extracts book data from the website Books to Scrape. The scraper focuses on three categories: *Travel, Mystery,* and *Historical Fiction*.

2. Tools and Libraries

- Python – main programming language.
- requests – for sending HTTP requests.
- BeautifulSoup (bs4) – for parsing HTML.
- csv – for saving the extracted data.
- time, urllib.parse – for handling delays and URLs.

3. Methodology

The script sends requests to category pages, parses book information, and iterates through all available pages. A short delay (time.sleep) was added between requests to avoid overloading the server.

4. Extracted Data

For each book, the following details were collected:

- Title
- Price
- Availability
- Rating (1–5 stars)

5. Implementation Highlights

- get_soup(url) function loads and parses HTML.
- extract_book_info(product) extracts book details.
- get_books_from_category(category_url) iterates through all pages of a category.
- main() function aggregates all results and saves them into books.csv.

6. Results

The script successfully collected book data from the three categories and stored it in a CSV file. The final output is a structured dataset that can be used for further analysis.

7. Conclusion

This project demonstrates practical web scraping with Python. The scraper can be easily extended to other categories or websites with similar structures.