Università della Svizzera italiana

Course Project Report

Martin Lettry, Eduardo Trabattoni

November 13, 2023

1 Project Overview

Our project aims to design, implement, and evaluate a search engine focused on football clothing by scraping and indexing data from three different websites specializing in this domain. To ensure the success of our project, we have carefully considered the implementation of specific features, with an initial focus on:

- **Results Presentation**: We plan to present search results in a tabular format to allow users to view multiple results simultaneously. Each table cell will contain relevant information to enhance the user experience.
- **Automatic Recommendation**: We are also considering implementing a recommendation system to suggest similar products based on various attributes like category, description, and price. These recommendations will enhance the user's shopping experience.

2 Key Milestones

To achieve our project's objectives, we have outlined key milestones:

1. **Website Selection**: We have identified three websites to scrape data from: `www.decathlon.co.uk`, `www.adidas.ch`, and a third website that we will decide upon soon.
2. **Data Structure Definition**: We have established an agreed-upon JSON schema that efficiently captures essential information from the selected websites, including URLs, titles, product data, prices, and images.

Listing 1: Agreed-upon JSON Schema

```
{
  "url": "URL",
  "title": "Title",
  "data": "Data",
  "price": "Price",
  "image": "Image"
}
```

3. **Scraping and Parsing**: We have developed a robust data collection pipeline using Scrapy, which has enabled us to scrape and parse data from `www.decathlon.co.uk` and `www.adidas.ch`. We have successfully collected a substantial number of items from these websites, with 1000 items from Decathlon and 1474 items from Adidas.
4. **Data Storage**: Currently, we are storing the scraped data in JSON format. However, we anticipate the need to transition to a Database Management System (DBMS) as we proceed with the project. MongoDB is a strong candidate for this purpose due to its ease of setup and scalability, combined with our team's prior experience.

5. **Indexing System**: We are in the early planning stages of implementing an indexing system. OpenSearch, a search and analytics engine built by Amazon Web Services, is under consideration for its efficiency in indexing. We aim to streamline the indexing process for efficient data access.

6. **Frontend Interface**: The development of the frontend interface is on our roadmap, and we plan to initiate this phase shortly. The interface will be essential for providing users with a seamless search experience.

3 Project Progress

At this point, we have made significant progress in the following areas:

- **GitHub Repository Setup**: We have established a dedicated repository on GitHub for our project, enabling efficient collaboration and version control.
- **Scraping Infrastructure**: We have created a Python virtual environment (venv) and a convenient shell script to execute the scraper. Additionally, we have compiled a comprehensive README file to guide project contributors and users effectively.
- **Data Collection**: Our scraping efforts have resulted in the successful extraction of data from `www.decathlon.co.uk` and `www.adidas.ch`. We have collected 1000 items from Decathlon and 1474 items from Adidas.
- **Data Schema**: We have finalized the JSON schema for our data, ensuring that it encompasses all essential fields from the selected websites.

4 Future Steps

Our project is progressing on schedule, and we have outlined the following key steps for the future:

1. **Data Storage**: We anticipate transitioning to a DBMS, such as MongoDB, to efficiently store and manage the collected data.
2. **Indexing Implementation**: We plan to begin implementing the indexing system using OpenSearch to optimize data access.
3. **Frontend Development**: The development of the frontend user interface will commence shortly to enable users to search and interact with the indexed data seamlessly.

5 Conclusion

In conclusion, our project is well underway, and we are dedicated to delivering a fully functional search engine for football clothing. Our focus on user experience, with features like results presentation and automatic recommendations, is set to provide a valuable tool for our users. The collaboration between Eduardo Trabattoni and Martin Lettry is going well, and our teamwork is balanced, ensuring efficient progress. We are excited about the project's potential and remain committed to meeting all project milestones with enthusiasm and determination.