Project Overview:

Our project involves scraping and indexing data from three websites that specialize in selling football clothing.

Chosen Features:

For now, we are leaning towards implementing the below features.

- Results Presentation — very likely
- Automatic Recommendation — quite likely

Key Milestones:

- Identify the three websites for scraping.
- Choose an effective data structure to consolidate information from the three websites.
- Implement scraping and parsing mechanisms for extracting relevant data.
- Select an optimal format for storing and retrieving the scraped data.
- Establish an indexing system for efficient data access.
- Create a frontend interface for searching the indexed data

Project Progress:

Overview:

We've established a GitHub repository for our project, set up a Python virtual environment (venv), and crafted a convenient .sh script to execute the scraper. Additionally, we've compiled a comprehensive README file that provides clear instructions on running the project and contributing to the codebase.

Websites for scraping:

- www.decathlon.co.uk
- www.adidas.ch
- TODO

Agreed upon data structure:

Below you will find the agreed-upon data structure which best encompasses all the relevant fields of the above sites.

Listing 1: Agreed-upon JSON Schema

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

Scraping and parsing progress:

We have built a pipeline using Scrapy for scraping and parsing the chosen sites. The results are formatted using the above JSON schema.

- www.decathlon.co.uk -¿ 1000 items scraped
- www.adidas.ch -¿ 1474 items scraped

Data storing progress:

Currently, we are storing all the data as JSON; for now, this is sufficient for our needs. Nonetheless, when we start on the indexing part of the project, we may need to begin storing this data using a DBMS. If that is the case, we will most probably use MongoDB as it is very easy to set up and scale, and we have prior experience in using it.

Indexing progress:

We are yet to start on the indexing part of the project. We are considering using OpenSearch for the indexing part of the project. OpenSearch is a search and analytics engine built by Amazon Web Services (AWS) as a fork of Elasticsearch. We have some experience using it and figured it would simplify the indexing process a lot.

Frontend GUI:

We are yet to start on the Fronted part of the project.