Beer Search Engine

Luca Di Bello, Enrico Benedettini

December 9, 2023

1 Dataset construction

1.1 Data sources

For our Search Engine, we looked up the most possible complete websites. We considered the best option to opt for websites that were not too complicated to scrap, but that were at the same time able to give us enough data to build an effective search engine.

So, the decision for us to pick the websites to pull the data was based on **two main parameters**:

1. **Structure of the content**: How easy it is to access the content according to the structure, whether the website is loading the content dynamically or statically, and how easy it is to parse the content.
2. **Data Quality**: How precise and detailed the information provided by the website is and how many records we could take out of it to build a good information retrieval system.

Among all the candidates, we have chosen the following websites to pull the data from:

- **WineVybe**: database with thousands of records of beers with Producer and Beer's name, type, alcohol bv, tasting notes, closure, and packaging.
- **RateBeer**: website that contains more than 100 thousand different beers with complete description, taste notes, alcohol bv, price, packaging, critic score, and brewer details.
- **BeerMe**: a database containing 11 thousand records of beers with brewer and beer's name, style, score, production date, and location.

Initially, other websites such as (BeerAdvocate and Untappd) were considered, but they were discarded as applying scraping techniques was not possible due to the dynamic nature of the content and the authentication required to access the data loaded dynamically via APIs.

1.2 Data scraping

The data scraping process was done using the **Scrapy** framework, which is a Python library that allows us to create spiders to crawl websites and extract data from them. The framework is very flexible and allows us to create a spider for each website we want to crawl. The spiders are composed of two main parts: