Information Retrieval

Status report

Deidda Paolo, Lämmle Christian

November 13, 2023

1 Introduction

The scope of our project is to build a system to retrieve, elaborate, store, and provide information about space explorations to users based on their information needs. In order to retrieve and save those informations we are using Scrapy[1], an open source Python framework. To index and query, we are using PyTerrier[2], another Python framework inspired by Terrier for Java.

Our idea was to find and save the name of the project, a (small) description, the date of launch, and the organization responsible for each space mission.

2 Websites

2.1 ESA[3]

This is the site that we were given to start our project; unfortunately, this site has a very good anti-scraping system that keeps banning our little spider. So we had to make a lot of test and change a lot of settings to make it work, we finally managed to achieve this today (13.11.2023). Now we can start to save the information which are the name and a little description of the mission.

2.2 NASA[4]

We chose this website because we couldn't overlook the agency that successfully landed the first human on the moon. Unfortunately, scraping the mission list page is not as easy as it may seem, as it seems to generate the page on the fly using JavaScript, and Scrapy doesn't seem to handle it very well. We have been working on it for a while now, with the help of both TAs and another experienced student. However, it appears to be more challenging than we initially expected. We will definitely investigate it further since we find it very useful for learning. Nevertheless, we might consider changing our data source if the difficulties end up requiring more time than it's actually worth.

2.3 Wikipedia[5]

Wikipedia offers a nice list of space missions, each with a lot of information. This is why we decided to add it as one of our sources; at the same time, it is not an official website of a space agency, and we have to take more care while working with this information as it can be incorrect. Based on our initial scouting of the site, this crawl is going to require more attention than anticipated since the pages have a lot of information on them, and we have to decide if everything is worth scraping or not. Moreover, on this site, we only find the first mission to accomplish something, so the list of missions is not complete at all.

References

[1] https://scrapy.org/

[2] https://pyterrier.readthedocs.io/en/latest/

[3] https://www.esa.int/ESA/Our_Missions

[4] https://www.nasa.gov/missions/

[5] https://en.wikipedia.org/wiki/Timeline_of_space_exploration#1970%E2%80%931979