Milestone 1: Data preparation

Processamento e Recuperação de Dados

Diogo Nunes
up201808546@up.pt
FEUP
Portugal

Jéssica Nascimento
up201806723@up.pt
FEUP
Portugal

João Vítor Fernandes
up201806724@up.pt
FEUP
Portugal

Figure 1. Drivers take the start of the Formula One Russian Grand Prix at the Sochi Autodrom circuit in Sochi on September 26, 2021. (Photo by Alexander NEMENOV / AFP)

Abstract

Data preparation process when creating an information search system about Formula 1.

1 Introduction

This work consists in the creation of an information search system, in our case, directed to formula 1. For that, we resorted to datasets present in Kaggle, more specifically F1 dataset (1950-2021) and Wikipedia.

1.1 Search Scenarios/Examples

With these datasets, our goal is to be able to create results for the most varied researches on Formula 1. Examples:

- Searches on Drivers
  Search: Antonio Giovinazzi
  Answer: Personal data information, in which constructor team is included, important results and important moments in which he was included (incidents, important overtakes, ...) and career description.
- Searches on Constructors
  Search: Ferrari
  Answer: The history of the team in F1, memorable results, best drivers racing for the team, actual drivers on the team (if it is the case).
- Searches on Seasons (year)
  Search: 2020
  Answer: Summary of the full season, crucial highlights, driver and constructor champions, ...
- Searches on Circuits
  Search: Monaco
  Answer: Location, how many times that city/country hosted a race (if it is the case), most successful drivers on that track, description about the track (how many turns, how many DRS zones, ...).
- Searches on specific moments
  Search: Kimi Räikkönen made contact with Antonio Giovinazzi
  Answer: Moments where both drivers crashed each other, maybe crashing with other drivers as well.

Search: Kimi Räikkönen overtook Antonio Giovinazzi
Answer: Moments where both drivers were fighting for position, and a brief description about it, as well as individual information about both drivers.
Search: Kimi Räikkönen started on pole
Answer: Races where a certain driver started on P1, most successful qualifyings from him, best results and driver career description.
Search: Regulations, awards, engines, federations, licenses, F1 cars, parc fermé, tyres, recovery systems, ...
Answer: Information about these F1 topics, independent of any results and structured CSV data. Results will be composed only by paragraphs extracted from Wikipedia.
Search: Who started on grid 5 on 2009 Australia GP?
Answer: The driver that started on grid 5 on Australia's 2009 Grand Prix.
Search: Qualifications for Belgian Grand Prix on 2018
Answer: Report paragraphs on the main qualification events for the Belgian Grand Prix on 2018
Search: Hamilton crash
Answer: Report of the races where the driver Hamilton crashed

2 Data Collection

2.1 Datasets

The dataset contains information about all the races since the start of this sport, 1950, until the present year, 2021. It was compiled from Ergast Developer API and it's composed by 14 CSV files: circuits, drivers, constructors, seasons, qualifying, races, status, results, lap times, driver standings, constructor standings, constructor results and fan ratings. However, we believe we have already so much information using only 8 of them: qualifying, results, status, seasons, drivers, constructors, circuits and races.

2.2 Wikipedia

We have URL links to Wikipedia on 5 CSV files: races, seasons, circuits, constructors and drivers. We created 1 additional CSV file to include information about other F1 pages, related with regulations, awards, licenses, and other technical topics. We also generated 2 CSV files from qualifying.csv and results.csv, to convert all the structured data we had to paragraphs that can be searched for on the next milestones.

3 Data Preparation

Once we have chosen the CSV files to use, it's time to make data cleaning. In our case, only results and status files were cleaned using OpenRefine software. As there were 2 columns with the same essence, we decided to eliminate the column "time/duration" and keep only "milliseconds", renaming it to "time (ms)". The same was done for "position" and "positionText".

With the help of 8 CSVs and python, 2 new files were generated with plain text, gen_result_text and gen_qualifying_text.

Since qualifying, seasons, drivers, constructors, circuits and races files had URLs to Wikipedia pages, scraping was done using BeautifulSoup4, in order to get more large textual information. With the same intention, however for pages that we didn't have information about on the datasets, 19 Wikipedia pages (about F1 History, regulations, awards, evolution, ...) were crawled. Thus, we got all paragraphs existing on each page.

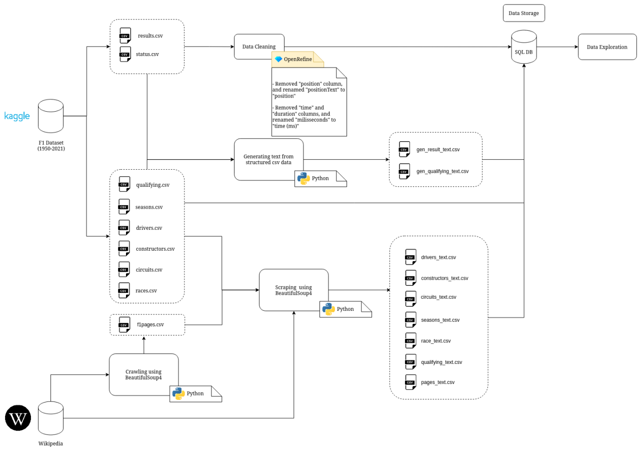

Figure 2. Pipeline

4 Conclusion

We conclude that to create a search system it's necessary to research the data to be used and its treatment. With all the information extracted and processed, we believe that the amount of information we have, from factual data to unstructured data, will be appropriate for our project.