Information Retrieval **2024**

Students: Matteo Borioli & Stipe Peran

Project Name: Pokémon

---

1. Tools

1. Web Crawling

For scraping, we used the Python library `scrapy`, which allowed us to retrieve data from two websites:

- `https://pokemondb.net/`
- `https://bulbapedia.bulbagarden.net/`

From PokémonDB, we collected most of the data regarding the basic statistics of Pokémon, their names, and their numbers. On the other hand, we primarily retrieved Pokémon images from Bulbapedia. In order to do that, we used two types of spideres: pokemonBulbapedia_listAllPokemon and pokemonDatabase_listAllPokemon. The second one uses a recursive call to scrape the description of each pokémon.

The data was initially saved in two JSON files. Subsequently, using two auxiliary functions, we merged the data and sorted them by ID. The final result can be found in the `bigPokemonData.json` file located in the `data/final_data` folder.

2. Indexing

Our backend is built on the open-source Solr platform, which handles data indexing and processes user queries and requests received from the frontend. To integrate the data obtained through web crawling, we utilized the managed-schema.xml file located in the project root.

3. Frontend

As for the frontend, it is yet to be implemented. It will be developed using React, along with various extensions that we will select as needed throughout the process.

2. Features

Regarding the features we plan to implement, for the simpler ones, we aim to include at least one of the following: Results Presentation or Filtering. As for the more complex features, we are considering implementing Automatic Recommendation.