Recap – Scraping

What is HTML code?

- HTML is the **structure** of a website (vs. CSS for style and JavaScript for actions).
- The different areas of the page are identified by **tags**. Ex: `<p>` for a paragraph of text.
- To extract content from a page, we need to know in which tag it is, thanks to the **browser inspector**.

The main steps of scraping

1. **HTTP request** (*requests* package): to get the content (i.e. the HTML code), with `requests.get()`.
2. **Parsing** (*Beautiful Soup* package): split the raw HTML code into a **navigable tree**, where you can **search for tags** with `soup.find('tag')`, `soup.find_all('tag')`, etc.
3. **Crawling** (*Selenium* package, optional): if needed, **navigate between pages** of the website, either by **constructing new URLs** and making new HTTP requests, or by simulating a real user with **Selenium** (see correction of J8 Exercices).

HTTP status codes

When you request a page's content with the *requests* package, you get a Response Code:

- **2xx : success codes**. Ex: 200 = success.
- **3xx : redirection**. Ex: 302 = temporary redirection.
- **4xx : client errors**, i.e. missing data on the website. Ex: 404 = page not found.

- **5xx : server errors**, i.e. problem with the website provider. Ex: 503 = server unavailable (overloaded) or in maintenance, 504 = time-out (no answer).

Databird – Thomas de Mareuil