



Scraping

We want to scrap HackerNews website. During this exercise you will learn how to scrape a website and how to store data.

Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites. The web scraping software may directly access the World Wide Web using the Hypertext Transfer Protocol or a web browser. (Wikipedia)

So we will download HTML pages with code and then interpret (or read) these pages to get the data we need from there. HTML is a web markup language that can be represented as a tree. All the webpages have an HTML structure.

1. Visit HackerNews and understand the HTML structure of a page
2. Schematise the HTML structure by drawing tree block
3. Get the HTML page in Python using the `requests` library
4. Start parsing the page with Beautiful Soup
 - a. Have a look at the documentation
 - b. Print all the title on the homepage (30 names)
5. Structure your information by creating a `dataclass` named `Post` containing all the relevant information
6. Parse the following informations: `name`, `points`, `author`, `data published`, `number of comments`

7. Write the data in a CSV
8. Write code to parse the 5 first pages of HackerNews
9. Think about a way to run your code every hour