# **Web Scrapping**

# In this assignment, you will build several programs that scrape various websites. The following information outlines what you need to do.

## **Before You Begin**

1. Create a new local git repository for this project called web-scraping-{yourname}.
2. Add your code files to this folder.
3. Commit the changes before the end of the assignment.

## **Part 1: Python scrapping**

Complete your initial scraping using Jupyter Notebook, BeautifulSoup, and Requests.

* Create a Jupyter Notebook file called nyt-best-sellers.ipynb. Use this file to complete the scrapping. The following information outlines what you need to scrape.

### The New York Times Best Sellers Books

* Scrape the [NYT Best Sellers Books Site](https://www.nytimes.com/books/best-sellers/) and collect the best sellers' books of the last 2 years (104 weeks). Collect all categories for each week: category name and its books. For each book collect the title, position, and description and image url.
* This list will contain one dictionary for each week.
* **Bonus:** Implement the task using multiprocessing to improve efficiency.

# Example:

best\_seller = [

{"week": 1, "data":[{

"category": "Combined Print & E-Book Fiction",

books: [

{"title": "Cerberus Hemisphere", "img\_url": "...", "position": 1, "description": "..."},

{"title": "Orcolos 2", "img\_url": "...", "position": 2, "description": "..."},

}]],

]

## **Part 2: Puppeteer**

Use Puppeteer to scrape OncoKB website.

* Start by creating a nodejs project.
* Extract the following data from [OncoKB actionable genes](https://www.oncokb.org/actionableGenes) page:
* Gene & Gene URL
* Alterations & Alterations URL
* Navigate to each Alterations URL and create a combined JSON object by[waiting for response](https://pptr.dev/api/puppeteer.page.waitforresponse/) of the following XHR calls:
* https://www.oncokb.org/api/private/utils/numbers/gene/{GENE} - gene field in combined JSON
* https://www.oncokb.org/api/v1/variants/lookup?hugoSymbol={GENE}&variant={VARIANT} - variant field in combined JSON

Hints:

* Ignore alterations with text such as “Exon 19 in-frame deletions and X other alterations”