Web Scraping and Social Media Scraping
Project Rules

Przemysław Kurek & Maciej Wysocki

Organisation

- You can work in groups consisting of minimum two and maximum three students. It is up to you who are you in group with.
- You need to find the topic of the project by yourself. At Campuswire there is a dedicated post in which you declare the websites you will scrape. Posting a domain address == its reservation. Each project group shall scrape a different website, so make sure that the website that you have chosen is unique. Please, post not only the website address, but also the names of the group members.
- The deadline for the project submission is Saturday 10.06.2023 23:59. The deadline is **extremely long**, so there will be no discussion to extend it. The only acceptable reasons for extension are those accepted by the Dean.
- If you need to change the website or make some major changes in your project, please be prepared that no deadline extension will be granted. And such things happen often.

Project Goals

- You need to write three scrapers: one using Beautiful Soup, one using Scrapy, one using Selenium. All of them should scrap the same information from the domain of your choice. The minimum of scraped different links is 100 (no matter how many information is on one page).
- If you choose to scrape a dynamic website (which has to be clearly justified, so be sure you know what you are doing), you can omit Beautiful Soup scraper, but **Scrapy Splash** is required.
- Your goal is to gather the information of your choice, perform some extremely simple analysis of the gathered data and compare performance of all your scrapers.

Expected Submission Files

1. Project description in the `description.pdf` file containing:
    - Names and ID's of all participants.
    - Short description of the topic and the web page.
    - If you omit Beautiful Soup scraper: justification, that the page is dynamic, and scraping can not be done with Beautiful Soup there.
    - Short description of your scraper mechanics - what the program is technically doing.
    - Short technical description of the output you get.
    - Extremely elementary data analysis - you need to prove, that collected data can be used for further analysis, but nothing more (hard limit of data analysis: one page).
    - Detailed description which group participant wrote which part of the project.
2. Presentation:
    - Record short (10 min) screencast presentation.
    - Every student in the group must record a part of it.
    - Every students records short description of the code they have contributed into the project - what is it and how it works.
    - Presentation also should consist of proof of work of all three scrapers - run them for a limited number (for example 100) of pages and present the raw data it gathered.

3. Source files:
    - Three folders: `"soup"`, `"scrapy"`, `"selenium"` containing all files required to run each scraper.
    - Beautiful Soup and Selenium scrapers must contain full programs, that without errors can be evaluated by `python3` interpreter. Scrapy folder should contain a full scrapy project folder, which can be run with `scrapy` command(s). Test them all in command line.
    - At the beginning of your code set up a boolean parameter limiting the number of pages. If the parameter value is `True`, it should limit the number of pages you scrap to 100. Set the default value to `True`.
    - Remember to write clean and understandable code. Write short, informative comments that should help with understanding the logic of the program.

Submissions

- In order to submit your source files create your own Github repository. You can freely work there within your group, and have one repository for all of you. It is a tool for collaboration after all.
- In order to submit your project description and presentation recording use your student's Google Drive. Share those two files in a way, that everyone with link can access them.
- In the end leave just three folders, readme and description file.
- In `README.md` write instruction how to run your scrapers.
- Your repository should be public. Do your best to make it representative to the world.
- Each of participants in the group has to submit your work in your Github Classroom repository. In order to do so create `"project"` folder. In this folder include `"project.txt"` file, which contains Github link to your project, names and ID's of all group members and two links for project description and presentation. **separate those four informations with a new line character**.

Grading

- Not fulfilling above requirements grants 0 points from the project.
- There are 50 points to score in total.
- As long as the scrapers works properly, you are going to get full points. However there are few exceptions:
    - If the description (written or spoken) is not clear, the points for all participants will be reduced.
    - If the codes are messy or not clearly commented, points will be also reduced.
    - If work is severely imbalanced, the points may reduced for some participants.