Web Scraping and Social Media Scraping
Project Rules

Anna Lewczuk & Przemysław Kurek

Organisation

- You can work in groups consisting of up to 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 scrap. Posting a domain address == its reservation. Each project group shall scrap 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 09.05.2020 23:59. The deadline is **extremely long**, so there will be no discussion to extend it. The only acceptable reason for extension is confirmed by Students Office sick live.
- 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.
- If you choose to scrap dynamic website (which has to be clearly justified, so be sure you know what you are doing), you can omit Beautiful Soup scraper.
- Your goal is to gather the information of your choice, perform some extremely simple analysis of gathered data and compare performance of all your scrapers.

Expected Submission Files

1. Project description in the `description.pdf` file consisting of:
    - 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 BS there.
    - Short description of your scraper mechanics.
    - 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 participant wrote which part of the project.
2. 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 only spider files, which after copying to relevant scrapy project folder can be run with `scrapy crawl` command(s). Test them all in command line.
    - At the beginning of your code you set up a boolean parameter that if it is `True`, it will limit the number of pages you scrap to 100. Set default as `True`.

Submissions

- In order to submit all your 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 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.
- Also, each of you 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, and names and ID's of all group members.

Grading

- Not fulfilling any of above requirements grants 0 points from the project.
- There are 50 points to score in total.
- As long the scrapers works properly, you are going to get full points. However there are few exceptions:
  - If the description 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.