Columbia Business School

Professor Kriste Krstovski
B9122

Fall 2023
Computing for Business Research

Homework Assignment #2

Due: 11:59 pm, October 9th (Monday), 2023

- Upload your solutions to the Canvas system before the deadline (email Zoe rm4069@columbia.edu, or Kevin kt2942@columbia.edu) if that does not work). If you are uploading your solutions in separate files, you need to package them into a single zip file (e.g., using Winzip) and name it ***b9122_hw2_sol_first_last_name.zip*** where you replace "***first_last_name***" with your full name.
- Homework assignments must be done individually, without using anyone else's solutions. You may obtain tips/tutorials online, but soliciting help from others online or in person is not permitted. Cheating will strictly not be tolerated.

Question 1 (80 points)

Write web crawlers for the following two tasks:

1. Extract at least 10 United Nations press releases containing the word "crisis". Start with the following seed url: https://press.un.org/en. Notice how press release pages have the "PRESS RELEASE" relative link in the top left corner. Here is an example press release:
   https://press.un.org/en/2023/sc15431.doc.htm where the "PRESS RELEASE" has the following relative anchor tag:
   ```
    <a href="/en/press-release" hreflang="en">Press Release</a>
   ```
   Use this information to determine whether the web page is a press release.

2. Crawl the press room of the European Parliament and extract at least 10 press releases that cover the plenary sessions and contain the word "crisis". Start with the following seed url:
   https://www.europarl.europa.eu/news/en/press-room
   Notice how press releases related to plenary sessions contain the text "PLENARY SESSIONS" with the following html:
   ```
   <span class="ep_name">Plenary session</span>
   ```
   Here is an example:
   https://www.europarl.europa.eu/news/en/press-room/20220620IPR33417/national-recovery-plans-meps-assess-the-performance-of-crisis-funding

Question 2: Git (20 points)

Create a Git repository on the GitHub platform. Name the repository `b9122_homework2`, make it publicly available and perform the following:

- Populate the repository with the webcrawler code that we covered in class and the webcrawler code files that you created in Question 1.
- Create a `README.md` file where you'll provide information about the repository, including author information and a description of the code files.
- Make changes to at least one of the added files (whatever changes you prefer).
- Update the repository with the edited file/s.
- For those of you who will be doing the interaction with the github repository using git commands perform the following:

    - The `git log` command displays the commit logs. Use output redirection (“>”) to store the output of this command in a file named `gitlog.txt`.Submit the `gitlog.txt` file and the `url` of your repository
- For those of you who will be using the GitHub Desktop application, perform the following:
    - The “History” tab displays the repository activities. Open this tab and take a screenshot. Submit the screenshot image and the `url` of your repository.