Columbia Business School

Professor Kriste Krstovski
B9122

Fall 2022
Computing for Business Research

Homework Assignment #2

Due: 11:59 pm, October 10th (Monday), 2022

- Upload your solutions to the Canvas system before the deadline (email Deema daa2182@columbia.edu, Serrana sa4117@columbia.edu, or Suro sl5203@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 are to be done individually, and without the use of anyone else's solutions. You may obtain tips/tutorials from the internet, but soliciting help from others online or in person is not permitted. Cheating will strictly not be tolerated.

Question 1 (80 points)
In this question you are tasked with writing modified versions of the web crawler that we covered in class. More specifically, you asked to create two webcrawlers for the following tasks:

1. Crawl pages whose seed url is the press releases page of the Federal Reserve System: https://www.federalreserve.gov/newsevents/pressreleases.htm and collect pages that contain the word "covid" found within the page. The goal is to collect at least 10 such urls. At the end of the crawling the code should output the urls of the webpages found to contain the word "covid". When checking whether the word is present on the webpage you should consider lower- and upper-case word versions (Covid, COVID, covid). One way to do this is to lowercase the webpage text prior to doing word matching.
2. Crawl pages whose seed url is the press releases page of the Securities and Exchange Commission: https://www.sec.gov/news/pressreleases and collect urls of press releases that contain the word "charges". The code should output the first 20 such links that it finds. For each link output the url and the text. Similar to the previous task, when checking for the presence of the word "charges" you should consider lower- and upper-case versions.

Question 2: Git (20 points)
Create a Git repository on the GitHub platform named `b9122_homework2` 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 description of the two 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 that 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 that will be using the GitHub Desktop application perform the following:
    - The “History” tab displays the repository activities. Open this tab, take a screenshot. Submit the screenshot image and the `url` of your repository.