ECON860 Data Analysis For Economics(Fall 2023)
Midterm Exam

Due: 2nd November 2023 (11:59 pm)

Name: _____

UID: _____

---

1. This exam is divided into three parts:

   - Obtaining information from list of Github users in a webpage
   - Downloading more information of those Github users
   - Exploring the dataset

   (a) Part 1:

   - The webpage `http://www.charcoalpaper.com/exams/github/user/dataset` contains the information of a list of Github Users. Scrape this webpage to obtain their login names, number of repositories, number of followers, and the date they became a Github user. Save the information in a csv file.
   - Make sure that you check the validity of the login ID downloaded. The information you saved should match what you see when you browse the webpage using a browser. If the information you downloaded does not match what you see in the browser, you may need to change the way you request or parse the webpage.
   - The list may contain identical login IDs. Please remove the duplicates.
   - The list may contain invalid login IDs. Please remove them.
   - At the bottom of the page, there is a section "Bonus GitHub Data". Sometimes some extra github users ids will appear in this section (without the infomation of repo count and follower count). You can choose to record these extra users or not. You will get bonus points if you are able to download the user ids of these users.
   - Save the dataset you downloaded into a csv file. This csv file should contains four columns: "Login ID", "Repo Count", "Follower Count", "Member Since".
   - In the written report, includes the following:
     - Summary statistics of the dataset you download.
     - Sample size of the dataset, number of unique login IDs, number of invalid login IDs, number of login IDs with invalid/missing information.
     - Report the data you got from the section "GitHub Data" separately from those you got from the section "Bonus GitHub Data".

   (b) Part 2:

   - Using the login ID you obtained, download the information of these Github users using the Github API. For each of the Github users, obtain their avatar_url,url,number of following,number of starred, full name, company, blog, location, email ,hireable ,bio, starting time, last update time, and any other information you found interesting.
   - Bonus: You can look into the repository of these Github users and obtain more information. You can get an A in this midterm exam even if you do not attempt this part (given that you complete all other parts flawlessly). You will get bonus points if you are able to get even a small amount of extra information from the users' repositories.
   - Save the dataset you downloaded into a csv file.
   - To use the Github API to obtain user information, use the url `https://api.github.com/users/github_id`.

- If you use the API without login and authorization (just like the GoodRx scrapping program in the lecture), the Github API only allows 60 requests per hour. This is good enough for the midterm exam, but you need to be very careful not to exceed the limit.
- Alternatively, it would be better to use a Github personal access token. You should have a personal access token already, but if you do not, the Github Docs has detailed instructions on how to obtain one (`https://docs.github.com/en/authentication/keeping-your-account-and-data-secure/creating-a-personal-access-token` ; Scroll to the item "Creating a personal access token (classic)"). This increases the limit to 5000 requests per hour, which would be more than enough for the tasks in this exam.
- Treat your Github personal authorization token as a password. Do NOT commit the token into your Git repository. Use the trick we discussed in the TMDB_download lecture and make use of the .gitignore file to avoid committing the token into your repository.
- In the written report, includes the following:
  - Summary statistics of the dataset you download.
  - The information you get in Part 1 is information of the users in 2022, while the information you get in this part is current information of the users. Compare the two datasets.

(c) Part 3:

- Explore the dataset downloaded using scatterplots. Depending on the information you obtained in part 2, you may have different scatterplots. Explain the interesting patterns and economic insights you observe in these scatterplots.
- For example, you can plot the Repo counts and the follower counts in a scatterplot to examine whether there is a relationship between the two. State what you expect (In this case, we probably would expect a positive correlation) and explain why. Report whether you find the expected pattern in the plot, and if not, explain why.
- You can use whatever information you got from the previous part to make the scatterplots. Number of following, number of starred, starting time, and last update time are all good candidates.

Important note:

- In a real world task, usually you are not supposed to commit the files you downloaded into your Git repository. However, for the purpose of the coursework, you should include them (the csv files) in your Git repository so that I can see your work.
- You must hand in your homework via Github. Create a repository named "ECON860_midterm" and invite me as a collaborator. In your repository, you should have the code, the data downloaded, and a .gitignore file. You should also include a file named README, which includes step-by-step instructions on how to run your Python code to collect the data you collected. This is especially important if you have multiple Python files. Also, there should be a written report in pdf format.
- You can commit and push to Github as many times as you like. Only your last commit before the deadline is graded. I can read your previous commits, but they will not be graded.