CIS 8392
Topics in Big Data Analytics

#Assignment 2

Yu-Kai Lin

Assignment 2

Step 1. Obtain your GitHub token from https://github.com/settings/tokens

Step 2. Choose a GitHub user who has at least 20 repositories and 10 followers (Read this page, which gives some pointers on how to find such a GitHub user)

- The user can be either an individual or organization (e.g., google)
- Each student will find a unique user. No two students can use the same user.
- Once you find a GitHub user, double check that the user has not been taken by another student : CIS 8392 Assignment 2 Data Singup Sheet
- Once you are certain that no other student uses the same GitHub user, sign up yours in the data signup sheet. After you signed up your data, take a screenshot (preferably include the date/time) of the sign-up sheet for your own record in case others modify your input.

Assignment 2

Step 3. Use R Markdown to achieve the following:

1. Specify author, date, and title in the YAML metadata of your document
2. Describe the GitHub user and provide the URL to the GitHub page of the user
3. Use the `gh` package to collect data
4. Create the following tables:
    - A table showing the user's `login`, `name`, `public_repos`, `followers`
    - A table summarizing the followers' `login`, `name`, `public_repos`, `followers`
    - A table summarizing the repositories' `name`, `size`, `forks_count`, `open_issues_count`, `closed_issue_count`
5. Use `ggplot2` to provide at least 2 different and meaningful data visualization from the data. Provide detailed discussion for the each of the plots. (It is not enough to just change one variable in the axis.)

Assignment 2

Here are some additional notes about writing a RMarkdown report. Violating these rules may lead to a lower grade.

- Put the data in the same folder as your Rmd file. Whenever we run/knit an RMarkdown file, it uses the folder with the Rmd file as the working directory.
- Read the data in your Rmd code chunk using relative path. If you use an absolute path, I will not be able to knit the Rmd file to an html file from my end.
- You will lose 5 points if for any reason (input path, error in code, etc.) the Rmd file cannot be knitted to an html file.
- All tables (any output of a data frame) must be formatted using kable in your R Markdown report.
- Distinguish headings (## heading) and normal text. We should not put all the text in headings.
- Do not put your discussions/explanations in code chunk. Write them as normal text.
- Do not use `include=FALSE` or `echo=FALSE` in your code chunk. I need to read your code. You may use `message=FALSE`, `warning=FALSE` to suppress messages/warnings.
- Do not write an excessively long line of code. Break it into multiple lines to improve readability.

Assignment 2

Step 4. Knit the R Markdown file (.Rmd) to an HTML file

Step 5. The Rmd and HTML files must follow the naming rule below:

- `Assignment2-YourLastName-GitHubUserLogin.FileExtension`
  - For example:
  - Assignment2-Lin-aaronj1335.Rmd
  - Assignment2-Lin-aaronj1335.html

Step 6. Submit the two files (individually) to iCollege

Assignment 2

Due by the beginning of next class

Extra credit: the student who has the best report (determined by the instructor) will be given 5 extra points towards the final grade

- Submissions that are too similar would not be considered for the extra credit

Grading is based on the following:

- Grading is based on the submitted files on iCollege. Do not wait till the last minutes before the deadline. You will lose 10 points for late submission. You will receive 0 point if you submit your assignment via email.
- Whether all required files were submitted to iCollege on time, following the naming rule
- Whether the Rmd file is syntactically correct and can render the html file
- Whether the report has a professional format and style (succinct and yet provides adequate and clear discussions about the data and the plots)
- Whether the report meets the requirements specified in Step 3
- Whether the R file can successfully collect the desired data
- Whether the submission matches the record in the data sign-up sheet
- Whether the chosen GitHub user has more than 30 repositories