MCIS6273 Data Mining (Prof. Maull) / Fall 2022 / HW0

Points Possible	Due Date	Time Commitment (estimated)
5	Monday, Aug 29 @ Midnight	*up to* 4 hours

- **GRADING:** Grading will be aligned with the completeness of the objectives.
- **INDEPENDENT WORK:** Copying, cheating, plagiarism and academic dishonesty *are not tolerated* by University or course policy. Please see the syllabus for the full departmental and University statement on the academic code of honor.

OBJECTIVES

- Familiarize yourself with the JupyterLab environment, Markdown and Python
- Familiarize yourself with Github and basic git
- Explore JupyterHub Linux console integrating what you learned in the prior parts of this homework
- Listen to the Talk Python['Podcast'] from July 6, 2018: 1M Jupyter Notebooks analyzed

WHAT TO TURN IN

You are being encouraged to turn the assignment in using the provided Jupyter Notebook. To do so, make a directory in your Lab environment called `homework/hw0`. Put all of your files in that directory. Then zip that directory, rename it with your name as the first part of the filename (e.g. `maull_hw0_files.zip`), then download it to your local machine, then upload the `.zip` to Blackboard.

If you do not know how to do this, please ask, or visit one of the many tutorials out there on the basics of using zip in Linux.

If you choose not to use the provided notebook, you will still need to turn in a `.ipynb` Jupyter Notebook and corresponding files according to the instructions in this homework.

ASSIGNMENT TASKS

(20%) Familiarize yourself with the JupyterLab environment, Markdown and Python

As stated in the course announcement Jupyter (https://jupyter.org) is the core platform we will be using in this course and is a popular platform for data scientists around the world. We have a JupyterLab setup for this course so that we can operate in a cloud-hosted environment, free from some of the resource constraints of running Jupyter on your local machine (though you are free to set it up on your own and seek my advice if you desire).

You have been given the information about the Jupyter environment we have setup for our course, and the underlying Python environment will be using is the Anaconda (https://anaconda.com) distribution. It is not necessary for this assignment, but you are free to look at the multitude of packages installed with Anaconda, though we will not use the majority of them explicitly.

As you will soon find out, Notebooks are an incredibly effective way to mix code with narrative and you can create cells that are entirely code or entirely Markdown. Markdown (MD or `md`) is a highly readable text format that allows for easy documentation of text files, while allowing for HTML-based rendering of the text in a way that is style-independent.

We will be using Markdown frequently in this course, and you will learn that there are many different "flavors" or Markdown. We will only be using the basic flavor, but you will benefit from exploring the "Github flavored" Markdown, though you will not be responsible for using it in this course – only the "basic" flavor. Please refer to the original course announcement about Markdown.

§ **THERE IS NOTHING TO TURN IN FOR THIS PART.** Play with and become familiar with the basic functions of the Lab environment given to you online in the course Blackboard.

§ **PLEASE *CREATE A MARKDOWN DOCUMENT* CALLED `semester_goals.md` WITH 3 SENTENCES/FRAGMENTS THAT ANSWER THE FOLLOWING QUESTION:**

- **What do you wish to accomplish this semester in Data Mining?**

Read the documentation for basic Markdown here. Turn in the text `.md` file *not* the processed `.html`. In whatever you turn in, you must show the use of *ALL* the following:

- headings (one level is fine),
- bullets,
- bold and italics

Again, the content of your document needs to address the question above and it should live in the top level directory of your assignment submission. This part will be graded but no points are awarded for your answer.

(0%) Familiarize yourself with Github and basic git

Github (https://github.com) is the *de facto* platform for open source software in the world based on the very popular git (https://git-scm.org) version control system. Git has a sophisticated set of tools for version control based on the concept of local repositories for fast commits and remote repositories only when collaboration and remote synchronization is necessary. Github enhances git by providing tools and online hosting of public and private repositories to encourage and promote sharing and collaboration. Github hosts some of the world's most widely used open source software.

If you are already familiar with git and Github, then this part will be very easy!

§ CREATE A PUBLIC GITHUB REPO NAMED "`mcis6273-F22-datamining`" AND PLACE A README.MD FILE IN IT. Create your first file called `README.md` at the top level of the repository. You can put whatever text you like in the file (If you like, use something like lorem ipsum to generate random sentences to place in the file.). Please include the link to **your** Github repository that now includes the minimal `README.md`. You don't have to have anything elaborate in that file or the repo.

(0%) Explore JupyterHub Linux console integrating what you learned in the prior parts of this homework

The Linux console in JupyterLab is a great way to perform command-line tasks and is an essential tool for basic scripting that is part of a data scientist's toolkit. Open a console in the lab environment and familiarize yourself with your files and basic commands using git as indicated below.

1. In a new JupyterLab command line console, run the `git clone` command to clone the new repository you created in the prior part. You will want to read the documentation on this command (try here https://www.git-scm.com/docs/git-clone to get a good start).
2. Within the same console, modify your `README.md` file, check it in and push it back to your repository, using `git push`. Read the documentation about `git push`.
3. The commands `wget` and `curl` are useful for grabbing data and files from remote resources off the web. Read the documentation on each of these commands by typing `man wget` or `man curl` in the terminal. Make sure you pipe the output to a file or use the proper flags to do so.

§ THERE IS NOTHING TO TURN IN FOR THIS PART.

(80%) Listen to the Talk Python['Podcast'] from July 6, 2018: 1M Jupyter Notebooks analyzed

Data science is one of the most important and "hot" disciplines today and there is a lot going on from data engineering to modeling and analysis. Jupyter notebooks are essential to the basic data science toolkit, but they are interesting in their own right.

There are millions of notebooks on Github alone, many of which contain many useful things in them, from examples on how to use a particular library to datasets and beyond, in a variety of disciplines from economics to biology to engineering and even the social sciences.

Adam Rule is a researcher in computer science HCI (Human Computer Interaction) and his research from a few years ago was very interesting as it centered around examining nearly one million Jupyter notebooks on Github. It was truly fascinating!

You will be using Jupyter extensively in this course and this warm up podcast will expose many interesting things about them.

Please listen to this one hour podcast and answer some of the questions below. You can listen to it from one of the two links below:

- Talk Python['Podcast'] Show #171: 1M Jupyter notebooks analyzed

- direct link to mp3 file 1m-jupyter-notebooks-analyzed.mp3

§ PLEASE ANSWER THE FOLLOWING QUESTIONS AFTER LISTENING TO THE PODCAST:

1. List 3 things that you learned from this podcast?
2. What is your reaction to the podcast? Pick at least one point Adam brought up in the interview that you agree with and list your reason why.
3. After listening to the podcast, do you think you are more interested or less interested in learning from Jupyter notebooks on Github?