Web Exercise 1: Introduction of GitHub and Online Data Science Resources

DUE Date: September 03 (Thursday) , 5:30pm (on Blackboard).
Grade: 10 points

You can complete this exercise using your own notebook computers or desktops. (Estimated total hours: two hours).

GitHub Introduction (70 mins)

GitHub is a very popular software programming collaboration platform (for source control / version control). Many great software projects, start-ups, and open source software packages utilize GitHub to develop, share, and exchange their programming codes and software packages. It is also a very important resource for learning Data Science.

1. **What is "Git"? https://en.wikipedia.org/wiki/Git What is "GitHub"? https://en.wikipedia.org/wiki/GitHub (10 mins to read these definitions).**

2. **Watch TWO YouTube Video**
   a. **"What is Github?" https://www.youtube.com/watch?v=w3jLJU7DT5E**
   b. **"GitHub Tutorial For Beginners" (20 minutes) https://www.youtube.com/watch?v=0fKg7e37bQE**

   (You may not understand all content in this video when you watch it first time. That's fine. Please go ahead to do the step 2 (after watching this video). You can re-watch this video again after you complete the actual tutorial.

3. **Please go to this website and follow their instructions for the GitHub Tutorial (40 minutes): http://product.hubspot.com/blog/git-and-GitHub-tutorial-for-beginners**
   (When you install "git" in your local machine, please use all the default setting or selection during the installation).

Note#1: Before you make the **Git commit**, you may need to set up your user name and email according your **GitHub registration**. Please follow the two commands below to set up the global configuration.

```
git config --global user.email "you@example.com"
git config --global user.name "Your Name"
```

Note#2: One example for creating a new branch: **git checkout -b ming-new-branch**

Note#3: When you create a new repo on the GitHub website, please use "**GEOG594-yournickname" as the name of repo**. DO NOT check out the box for "initialize this repository with a README" in this exercise. After creating the new repository, copy

the command lines from "push an existing repository from the commend line" section: for example:

```
git remote add origin https://GitHub.com/mingtsou/GEOG594.git
git push -u origin master
```

Key concepts in these tutorials:

- **What is a "repo"?**
- **What are the popular commands in terminal modes (cd .., cp, mv, ls, rm)?**
- **What are the differences between "pull" and "push"? What will "clone" do?**

Additional YouTube Video to learn GitHub:

- **"An Introduction to Git and GitHub by Brian Yu" on YouTube (40 mins, very nice introduction of some concepts): https://www.youtube.com/watch?v=MJUJ4wbFm_A**

Markdown language (15 mins)

4. **Now you need to learn how to create the README.md file in your GitHub Repo (BDA594-yourname) and the "Markdown" language in GitHub. In your newly created GitHub Repo. Click the "Add a README" green button. Now you will be in the Edit mode for your README file. (the README file is written in Markdown language). Please read the introduction about Markdown in this link: https://guides.GitHub.com/features/mastering-markdown/**

   After learning the Markdown style, please customize your README.md file to describe the following content:
   1. **Your First name and Last name,**
   2. **The URL to the BDA/GEOG594 class**
   3. **Your own definition of Big Data.**

   Print out this README.md (in the regular display mode) into a PDF file (as one of lab exercise submission items).

Create a GitHub Website (15 mins)

5. **Every user can create a free website using GitHub. Please follow the instruction HERE: https://guides.GitHub.com/features/pages/ (Create a new repo using "yourname.GitHub.io") Follow the instructions, you will create your own website hosted by GitHub. Copy the link of your GitHub website for the lab report.**

Resources for Learning Data Science (20 mins)

6. **Read the article "Top Data Scientists to Follow & Best Data Science Tutorials on GitHub" first. https://www.analyticsvidhya.com/blog/2015/07/GitHub-special-data-scientists-to-follow-best-tutorials/**
7. **Go to the awesome-datascience GitHub https://GitHub.com/bulutyazilim/awesome-datascience In the "Data Sets" list, select one data set from the list and introduce the selected one (with URL) and its potential applications.**
8. **Got to the free-data-science-book GitHub: https://GitHub.com/chaconnewu/free-data-science-books Pick up one book from the list and introduce the selected one with URL and the author/institute information. Explain briefly about why you like to select this book.**

Additional Resources for Git Code Sharing and Management

- **GitLab:** https://about.gitlab.com **GitLab is a tool to build a web-based Git repo systems from a centralized server. (You can build your own "Github" site).**
- **Bitbucket:** https://bitbucket.org **Similar to GitHub. Bitbucket focus on services to professional developers with private proprietary software code. Free private repos with a small team (up to 5 users).**

After finishing these exercises, Please use your own words to answer the following questions (next page): (DO NOT COPY any web resources or Wikipedia texts. We will check your answers with Blackboard tools to verify that your responses are uniquely yours.) By submitting your answers (paper) to Blackboard, you agree: (1) that you are submitting your paper to be used and stored as part of the SafeAssign™ services in accordance with the Blackboard Privacy Policy; (2) that your institution may use your paper in accordance with your institution's policies; and (3) that your use of SafeAssign will be without recourse against Blackboard Inc. and its affiliates.

SafeAssign accepts files in .doc, .docx, .docm, .ppt, .pptx, .odt, .txt, .rtf, .pdf, and .html file formats only. Files of any other format will not be checked through SafeAssign.

LAB-1 Questions:

1. **What are the key functions of GitHub? Who are the users? (Please use your own words to describe them in 200 words – 300 words).**

2. **What are the differences between "clone" and "pull request" in GitHub?**
3. **Introduce the selected "Data Set" from the [awesome-datascience] GitHub (with the data source URL) and describe its potential applications and values.**
4. **Introduce ONE selected "free-data-science-book" with URL and the author/institute information. Explain briefly about why you like this book.**
5. **Print out the README.md (in a regular display mode) as a PDF file for attachment #1.**
6. **Print out your GitHub Website from a Web Browser (Chrome or others) as a PDF file for attachment#2.**

Please submit your LAB-1 Answers (in a MS Word or a PDF file format with additional attachments) to the Blackboard System BEFORE the DUE DATE/TIME.

NOTE: You can attach multiple PDF files into the Assign Submission in the Blackboard. Please make sure to check the agreement for using Global Reference Database before clicking on the "Submit" button.