 Due by 2:30pm, 14^{th} Sept 2018

CMPSC 301
Data Analytics
Fall 2018

Lab 1: Setting Up GitHib Classroom
and Considering Violations Through Data Abuse
7^{th} Sept 2018

Objectives

To learn how to navigate the directories within Ubuntu operating system using command line interface. To set up GitHub for use in the course. You will also consider some of the negative consequences of data abuse according to articles in the general media. You will also learn how to prepare and write a Markdown document.

Reading Assignment

If you have not done so already, please read all of the relevant "GitHub Guides", available at `https://guides.github.com/`, that explain how to use many of the features that GitHub provides. In particular, please make sure that you have read guides such as "Mastering Markdown" and "Documenting Your Projects on GitHub"; each of them will help you to understand how to use both GitHub and GitHub Classroom.

Using Your Computer Science Account

In advance of today's lab you have already received the details about your Alden Hall computer account and learned how to log on. You may use this account on any computer in Alden labs in rooms; 101, 103, or 109. Your files are stored on a central server; and may be reached from any lab machine after logging in to a lab machine.

Hours of lab availability are posted on the bulletin board in each lab and on the following Web site: `http://www.cs.allegheny.edu/`; the on-duty lab monitor is always available in Alden 101.

Navigating using the Terminal: the command-line interface

A command-line interface allows the user to interact with the computer by typing in commands. Computing professionals prefer to use the command line interface, called the "Terminal", built into operating systems like Linux, instead of using the graphical user interface for launching programs and etc. In many situations command line interface tends to be very efficient and effective, for example, it allows you to complete some tasks with a simple one line command instead of using having to click on desktop items using the mouse!

1. Read through the supplemental handout on "Tips on Using Linux and the Command Line Interface'.' Locate the terminal window and open it as explained in the reading handout.

2. Now you will practice using the commands discussed in the handout. Using the terminal window type each of the commands found in Table 1 of the supplemental handout. Make

sure you understand what each command does. You will have to create new files in order to run some commands such as `cp`, `mv`, etc. The most basic method of creating an empty file is with the `touch` command. This will create an text file using the name specified: `touch file1` or multiple files as: `touch file1 file2`. Remember to execute a command, you should press the "Enter" key after typing a command. Check with your neighbors to see if they are able to open the terminal window, and use commands such as `cd, cp, pwd, ..., ls`, etc. If you can help them, please feel free to do so!

3. To avoid confusion and clutter in the future, delete all of the newly created files and directories from the previous practice step.

4. Now create a directory called `cs301F2018` in your home directory, by typing `mkdir cs301F2018` command in your terminal. This is where all of the work you do in this class should reside.

5. From the home directory type the `pwd` command in the terminal.

6. You can now close the terminal window by typing the `exit` command. You can reopen another terminal window and navigate to this directory by using the `cd` (change directory) command, followed by the new location in your current directory. Remember that `cd ..` will move back one directory, while `cd <directoryName>` will move you up to the directory called `<directoryName>`.

Configuring Git and GitHub

During this and the subsequent laboratory assignments, we will securely communicate with the GitHub servers that will host all of the project templates and your submitted deliverables. In this assignment, you will perform all of the steps to configure your account on GitHub You can also learn more about GitHub Classroom by visiting `https://classroom.github.com/`. As you will be required to use Git, an industry standard tool, in all of the laboratory assignments and during the class sessions, you should keep a record of all of the steps that you complete and the challenges that you face to be used in your code documentation. Please ask for help from the course instructor if you have trouble completing certain steps.

Figure 1: GitHub allows students to package and ship their own software to the community.

1. If you do not already have a GitHub account, then please go to the GitHub web site (at `https://github.com/`) and create one, making sure that you use your `allegheny.edu` email address

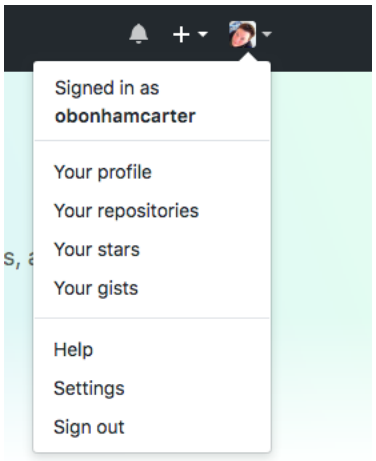

Figure 2: GitHub settings menu.

so that you can join GitHub as a student from Allegheny College who has permission to use various parts of the GitHub service. You are also encouraged to sign up for GitHub's "Student Developer Pack" at `https://education.github.com/pack`, qualifying you to receive free software development tools. Complete the form and once you gain access to the GitHub site, you should be able to see a screen where the icon of your account is shown. On my page, the upper right side is shown in Figure 1 on the GitHub page.

2. Additionally, please add a description of yourself and an appropriate professional photograph to your GitHub profile. Unless your username is taken, you should also pick your GitHub username to be the same as Allegheny's Google-based email account.

3. If you have never done so before, you must use the `ssh-keygen` program to create secure-shell keys that you can use to support your communication with GitHub. These keys enable you to send your files to GitHub without having to type in a password each time. Note: your ssh keys serve as the password and so you have to be using the machine of these keys.

   Open the terminal as you have done in the previous step. Alternatively, you can search for it by starting to type the word "terminal", and then select that program. Another way to open a terminal window involves typing the key combination `<Ctrl> +<Alt> + t`.

4. Now that you have started the terminal, you will now need to type the `ssh-keygen` command in it. Follow the prompts to create your keys and allow the generator program to save them in the default directory (`.ssh` of your root directory). Press "Enter" when you are prompted to `Enter file in which to save the key ...` and then push enter for your selected passphrase whenever you are prompted to do so in order to avoid typing it each time you communicate with GitHub. Verify that the generator has stored its key files in the `ssh-keygen` root directory (`.ssh`). Note, more information about ssh keys can be found at the following link: `https://www.ssh.com/ssh/keygen/`.

5. Once you have created your ssh keys, log in (again) into GitHub and look in the right corner

for your avatar information. Click on this link (similar to that featured in Figure 2) and then select the "Settings" option. Now, scroll down until you find the "SSH and GPG keys" label on the left, click to create a "New SSH key", and then upload your ssh key to GitHub.

You can copy your (PUBLIC!) SSH key to the clipboard by going to the terminal and typing "`cat ~/.ssh/id_rsa.pub`" command and then highlighting this output. When you are completing this step in your terminal window, please make sure that you only highlight the letters and numbers in your key—if you highlight any extra symbols or spaces then this step may not work correctly. Then, paste this into the GitHub text field in your web browser.

6. Again, when you are completing these steps, please make sure that you take careful notes about the inputs, outputs, and behavior of each command. If you have trouble, then please ask the course instructor.

   Since this is your first assignment and you are still learning how to use the appropriate software, do not become frustrated if you make a mistake. Instead, use your mistakes as an opportunity for learning both about the necessary technology and the background and expertise of the other students in the class, and the course instructor.

7. In the class repository that you created earlier, clone the `classDocs` repository (if you have not done so already), which can be found at:
   `https://github.com/Allegheny-Computer-Science-301-F2018/classDocs`.

   Before you can make the clone, you will have to locate the SSH clone address from the green button (labelled, *Clone or download*) at the website. After clicking on the button, you will use the displayed SSH address in the following command to copy the `classDocs` files to your local machine:
   `git clone <ssh address>`.

   This operation will only be used once. To collect all updated documents from here on, you will type the command, `git pull` within the `classDocs/` directory on your machine. Please ask questions if you have trouble.

GitHub starter link

`https://classroom.github.com/a/wUfHfXsJ`

Make a clone

One you have your account with GitHub Classroom, you are ready to use the service to clone your working directory. Click on the repository link above in red. You will find a submission directory of your repository into which you are to submit your page of work.

Part 2: Data and Data Abuse Case Studies

Data misuse is a very serious problem. Although a data scientist may have the means to process data for particular types of results, such a task may not always be ethically sound. For instance, some of the misuses of data may take the following forms, as shown below.

Note: Unless specified otherwise, the included examples were taken from:
`https://www.observeit.com/blog/importance-data-misuse-prevention-and-detection/`

- Cambridge Analytica, a political data firm hired by President Trumps 2016 election campaign, gained access to private information on more than 50 million Facebook users. The firm offered tools that could identify the personalities of American voters and influence their behavior. The data, a portion of which was viewed by The New York Times, included details on users identities, friend networks and likes. The idea was to map personality traits based on what people had liked on Facebook, and then use that information to target audiences with digital ads. (Reference: `https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html`)

- Bullying on Twitter: researcher find that 15000 bully-related tweets are sent daily. The Internet can be a hostile place, and Twitter is no exception. According to a new study, about 15,000 bullying-related tweets are posted every day, meaning more than 100,000 nasty messages taint the digital world each week. To further understand what happens in the virtual world, researchers from the University of Wisconsin in Madison trained a computer to analyze Twitter messages using an algorithm created to point out important words or symbols that may indicate bullying. In 2011, during the time of this study, 250 million public tweets were being sent daily a number almost 10 times the population of the state of Texas. The researches commented that (*ironically, inspite of the online bullying*), "What we found, very importantly, was that quite often the victim and the bully and even bystanders talk about a real-world bullying incident on social media," Zhu told the University of Wisconsin-Madison News. (Reference: `https://www.huffingtonpost.com/2012/08/02/bullying-on-twitter_n_1732952.html`)

- A recent high profile case of data misuse occurred when an employee at one of the world's fastest growing companies, Uber, violated the company's policy by using its *God View* tool to track a journalist who was late for an interview with an Uber exec. If you haven't heard, *God View* allows the company's staff to track both Uber vehicles and customers. It is not open to drivers at all, but it is apparently "widely available" at a corporate level. Tracking the journalist obviously flies in the face of even Uber's latest privacy policy, which states that employees are prohibited to look at customer rider histories except for "legitimate business purposes."

- In 2012, state auditors found that 88 police officers in departments across the state of Minnesota misused their access to personal data in the state driver's license database to look up information on girlfriends, family, friends, or others without authorization or relevance to any official investigation. Auditors said that this is not uncommon and that more than half of the police officers in the state made questionable searches in the database.

- The Florida Supreme Court heard a case in which a lower court found that Broward County, police officers misused data by conducting real-time GPS tracking on the location of a man's cellphone, using undisclosed techniques in collaboration with the cellphone carrier.

- A Chicago police officer responsible for administrating the department's criminal history database used the system to look up his girlfriend's record. Similar cases have shown up in other states, resulting in cases involving stalking, harassment, and identity theft.

- AT&T will pay $25 million to the Federal Communications Commission as a result of an investigation that discovered that employees at international call-centers illegally disclosed the personal information of upwards of 280,000 customers. The workers sold U.S. AT&T customer names and Social Security numbers to third parties who used it to unlock mobile phones so the devices would work on networks other than AT&T's, said Wednesday in a news release by the Federal Communications Commission.
- ... And plenty of other examples exist.

Your Task For Part 2

As part of an introductory lab exercise, you will research a similar example of the misuse of data to those from above. Your task is to find an article from a reputable news source (i.e., Washington Post, New York Times, Los Angeles Times, and etc) where some form of crime (or injustice) has been committed due to the misuse of data.

Explain the article and then write up four main points of argument to explain why the scenario of your news article is in violation of some public trust. You are to submit your response (via your GitHub Classroom repository) to the instructor where your four reasons are clearly displayed and supported by a short discussion. Please the Markdown language to prepare your reflection document concerning your article. Information about the Markdown language can be found above in the Reading Assignment Section. This submission should be about a page where you have provided sufficient discussion to outline why your example is a violation of public trust. Be sure to provide a reference to the article that inspires your thinking.

Important Details

Lab directory structure: You are to create a labs directory (`mkdir labs` in which you are to add the GitHub Classtoom repositories for each of your weekly labs (use this command `mkdir labs/labsxx`, where xx is the two digit lab number). For example, your first and second lab repository should be located in the paths, `labs/lab01` and `labs/lab02`, respectively.

Add you name to your work: Please remember to include your name on everything you submit for the class.

Required Deliverables

This portion of the assignment invites you to submit an electronic version of the following deliverable through your GitHub Classroom lab repository. Note: this repository is the one which you clone from the above link.

1. A document called `reflections.md` fulfilling the Part 2 requirements which will be submitted inside the directory called `writing/`.
2. Share your assignment files with the instructor through your Git repository by correctly using using appropriate Git commands in the following order to send you submission to the GitHub Classroom server.
   - `git add -A`,
   - `git commit -m ''my completed work''`
   - `git push`

Due by 2:30pm, 14^{th} Sept 2018

When you have finished, please ensure that the GitHub web site has your pushed work by visiting your repository at the site. Please see the instructor if you have any questions about assignment submission.