# Assignment 1: SOEN 6611

# Get Raw Data

Worth 5% of your grade, plus a max of 2% bonus

Due: **Jan 28th at 11:59pm** (1 mark off per day late)

You must use one of the following languages: Perl, Python, or JavaScript. If you use a library it must be included in the zipped file you submit to the TA (i.e., make sure it’ll run on a **Linux machine in the lab**). Using an HTML library will make part 1 easier.

## Part 1: Get git [2.5 marks]

The TA will run your script, which must be named get\_git and take a config file.

# ./get\_git git\_conf

You must parse html to find git repositories. Then clone the repo into a directory. In this directory, each repo must have it’s own name. You must also print out the repo name as you are cloning it. Make sure to try it out on a number of different urls (e.g., github and fdroid).

The config file is a plaintext file that will contain the following configuration variables:

url = https://f-droid.org/repository/browse/?fdid=com.grarak.kerneladiutor

root\_directory=/home/pcr/get\_git/

## Part 2: Get bugs [2.5 marks]

The TA will run your script, which must be named get\_atlassian\_bugs and take a config file.

# ./get\_atlassian\_bugs bugs\_conf

Look at the issue tracker for hibernate: <https://hibernate.atlassian.net>

Download the bugs between a particular range. You must print out the bug number of the bug you are downloading. You must also put a random length timeout that has a max\_timeout\_secs. The timeout between each bug should be printed to the screen.

Output: Each bug must be stored in the root\_directory with the bug number as the file name.

The config file is a plaintext file that will contain the following configuration variables:

url = https://hibernate.atlassian.net

bug\_start = 60

bug\_end = 100

max\_timeout\_secs = 60

root\_directory=/home/pcr/bugs/

## Part 3: Bonus [2 marks]. Interact with the API of GitHub.

Use the GitHub API to read the names of the people who have contributed to <https://github.com/poise/python>

Output a list of the contributors names read using the API.

Info about the API can be found here: <https://developer.github.com/v3/>