Homework 1
Getting Started & Web Scraping

CMU 15-388/688: Practical Data Science (Spring 2018)
http://www.datasciencecourse.org/
OUT: January 24, 2018
DUE: February 7, 2018, 11:59 PM

Guidelines

- The homework is due at 11:59 pm on Wednesday February 7, 2018. Each student will given 5 late days that can be spent on any homeworks, but at most 2 late days per homework.
- Homework problems are distributed in the form of Jupyter notebooks, which will be turned in to the autograder via the CMU Autolab system (https://autolab.andrew.cmu.edu/). In the notebook, you will fill in your solution code in marked cells, which will be validated by a set of tests on Autolab. Do not modify any given function signatures to ensure proper autograding. You can resubmit as many times as you'd like until the deadline.
- Do not import any additional modules into the Notebook, as they may not be found on the Autolab image. All necessary imports will be in the notebook.
- You are encouraged to refer to online references (such as Stack Overflow, specific library documentation) when completing the homework.
- If you get stuck, you are encouraged to come to office hours, or to ask your question on the class Piazza (https://piazza.com/cmu/spring2018/15388688/home).

1 Getting Started

1.1 Python Installation

All homework consists of notebooks of programming problems in Python 3.X, and will require several scientific computing packages. The simplest way to obtain a Python installation with most of the necessary dependencies (and the method in which we officially support for the class) is to install the Anaconda platform (https://www.anaconda.com/download), which contains both Python and a suite of scientific computing packages. If you don't use Anaconda, then you may need to install additional Python packages as needed in order to run your code locally (such as Pandas, SciPy, Numpy, etc.). We will do our best to help if you run into problems without using Anaconda, but we cannot make any guarantees. Although you are free to use Python 3.6, avoid using any features specific to the version (such as fstrings) as they will break on Autolab.

1.2 Jupyter Notebook

In order to view and edit the notebook, you will need to use Jupyter notebook. The notebook package is included in Anaconda, so if you installed Anaconda you are all set. Otherwise, you will need to install the Jupyter package (http://jupyter.org/). Once you have Jupyter notebook installed, you can start an instance in a shell with the command `jupyter notebook`.

With the default settings, you can view the notebook interface in your web browser at http://localhost:8888/tree, where you can see files in the startup directory. For example, if you run `jupyter notebook` in the directory `/home/`, then the notebook interface will display all files in your `home` directory. For more help on running a Jupyter notebook, you can view the Jupyter quick start guide (http://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/execute.html).

1.3 Download and Submission

In order to start the assignment, go to the course website (http://http://www.datasciencecourse.org/) and download the assignment files, which are tarred together. Untar it in a child directory of the startup directory for the Jupyter notebook. For example, if you started the notebook interface in `/home/`, then put your notebooks in `/home/` or a child directory thereof. Then, via a web browser interface, go to http://localhost:8888/tree, navigate to the downloaded notebook and click on it to view the assignment. You can now edit the notebook and complete the programming assignment.

When completing the homework, it is vital to not change the signatures of the functions we ask you to fill out, as the signatures are required by the autograder. Complete question 1 of the notebook, which asks you to implement some simple functions involving lists and dictionaries. To turn in your work, do the following:

1. **Save your work in the notebook.** Jupyter notebooks save at various intervals, but you want to make sure your latest work is saved before submitting.
2. Run the `make` command to tar your notebooks. Use the same directory structure that you received the notebooks in. (There should be a Makefile in the directory from which you call `make`.)
3. Upload the tar ball to Autolab (https://autolab.andrew.cmu.edu/).
4. Wait for the tests to run (minute or two), then view your feedback.

If you implemented the functions correctly, then you should now have credit for the first portion of the assignment. If you did not receive credit, then there is a problem with your implementation; click on your scores to view the autograder feedback, make the corresponding corrections in your notebook, and resubmit. Once you're familiar with this process, you can complete the rest of the assignment.