CSCI 4181/6802: Assignment 4

The majority of instructions as well as all questions and answers for this assignment will be provided via a jupyter notebook. These are a useful form of interative notebooks that can facilitate exploration and data analysis.

To start this assignment, first download the zip archive from the course website (https://maguire-lab.github.io/bioinformatics_algorithms/static_files/assignments/CSCI4181_6802_Assignment4.zip). Place it in your a new folder/directory and extract the files.

Then to access and run the jupyter notebook (the file ending in .ipynb) you'll need to install some software as follows.

Software Pre-requisites

Python 3.6 & the Jupyter Notebook (https://test-jupyter.readthedocs.io/en/latest/install.html), seaborn, pandas, and scikit-learn are required for this tutorial. If already have these installed or know how to install them already then feel free to skip the follow sections!

For people new to this toolchain use the following guidance as appropriate for your system. If you get stuck there are a lot of tutorials and guides online for this, failing that get in contact before the assignment is due.

Windows/Mac

If you are using **Windows** or **Mac**, I highly recommend anaconda: https://www.anaconda.com/download/. It is easy to install and gives you all the bits you will need for these interactive tutorials. Anaconda installs Python, the Jupyter Notebook, and other commonly used packages for data science.

Linux

If you are using **Linux**, I recommend using anaconda/miniconda to set up an environment for Python 3.6, especially if your system interpreter is Python 2.

We require the Python library scikit-learn (http://scikit-learn.org/stable/), a machine learning library for the Python programming language. If using conda/miniconda, this can be installed with:

```
$ conda install scikit-learn
```

Or otherwise with pip using the command:

```
$ pip install scikit-learn
```

We will manage some of the data using the Pandas package. It is included with anaconda but can be installed with **$ conda install pandas** or **$ pip install pandas**.

Finally, we need Seaborn which is a Python data visualization library based on matplotlib. The library can be installed with **$ conda install seaborn** or **$ pip install seaborn**.

Alternative Approach (Not Recommended)

Alternatively, you can try and use the jupyterhub instance on timberlea (https://timberlea.cs.dal.ca:8000/hub/login) but this hasn't been tested and you might find it a bit challenging to access the files you need on this system if you are less experienced.

Starting the assignment

Once you have all those tools installed and working, move into the folder into which you unzipped all the files and start jupyter. Then open the notebook ".ipynb file" in jupyter. The rest of the instructions for the assignment (and some code to check if the installation worked) can be found in this notebook. Please submit your answers via brightspace as an exported PDF of this notebook.