BIOL-GA1007 PROGRAMMING FOR BIOLOGISTS
End of term take-home exam 2018

Due Thursday 20th December, 2018. 100 points total. This assignment must be individual work.

Part A: (40 points)
In the attached python file “quiz_knn_answer_key.py” is a correct implementation of a 1-nearest neighbour model (as seen in the final quiz) using only the Python builtin data types.
Modify function knn_predict to implement a 3-nearest neighbour model. Do not use any functions in your code except functions defined in the file (do not use functions in the standard python library or other libraries)- use only builtin python statements.
Comment the code of the function knn_predict to clearly explain how it works. Include a comment on why sys.float_info.max is used.
Upload your modified python script called "final_part_A_NETID.py"

For this question, do not use Jupyter lab, but instead use a text editor to write this code, and use the git revision control system.

Include in your code a comment showing the python you would add to cause pdb (the python debugger) to break at a position just at the start of the function knn_predict above.

Part B: (30 points)
When writing your code for part A, use the git version control system i.e. as you make edits to your code, commit your changes in git.

Tag the final commit with the name "END_TERM_EXAM_2019".

Upload a bash shell script, named "git_part_B_NETID.sh" showing the steps needed to: (i) initialise and create the git repository, (ii) at least two example commits, (iii) the tagging operation, (iv) reverting to the first commit (see checkout in git), (v) reverting back to the last commit (HEAD).
Save the output of "git log" after you finish the code for part A to a text file called "git_log_part_B_NETID.txt". (The git log output shows a history of all your commits with comments).
Upload both the shell script and the txt file.

Part C: (30 points)
In python, use the pandas, numpy and the seaborn plotting packages to analyse and visualise the following dataset.
The dataset in the file "tetrahymena.tsv" is from a study of growth conditions of Tetrahymena cells. The average cell diameter (in microns) and cell concentration (count per ml) were measured for each culture. The measurements were repeated twice for each culture, giving two technical replicates for each culture.

(1) read in the dataset
(2) filter out excessively small and large cells with diameter <= 19.2 or >= 26.0

(3) use the mean concentration and diameter over the technical replicates. (This is to remove a statistical issue called pseudo-replication.)
(4) create a scatter plot of concentration versus diameter. Save as a pdf.
(5) create new columns "log_concentration" and "log_diameter" that have the natural log of concentration and diameter respectively.
(6) create a scatter plot of log concentration versus log diameter (if this is linear then the variables are related by a power law- is this the case?). Save as a pdf.

Use JupyterLab for your solution.
Upload your .ipynb file as final_part_C_NETID.ipynb, and an html file with embedded plots as "final_part_C_NETID.html".
Marks will be given for documenting the steps of your analysis in JupyterLab to produce a final report that clearly describes your analysis and results, as you might use in a scientific report.