BIN 500 – Assignment 4

Due date: December 26, 2019

Late policy: For each assignment, 20 points deduction will be applied for one day late, and 10 points additional deduction for each extra day.

Use python to implement the required methods to solve the problems below. Please submit your python program codes (.py files) to ODTUclass. Please do not forget adding comments to your code. Remind that the instructor has the right to request a demo of the codes at any point and can determine the final score based on the demo performance.

1. In the tab-delimited file "data.txt", there is a dataset of 2000 UniProt proteins with their different features such as their unique identifiers, gene names, lengths, and sequences. This dataset is known to have missing values that are marked as 'NA' where you can use the help of exception handling while processing. Be careful with the mass values as they contain commas which might need to be removed. Use this file to create appropriate data structures that will be used to answer the questions below.

   Find the answers of the questions using your code and print out the results for each one.

   a. What is the longest protein in this dataset? What is the average length of all proteins?
   b. How many proteins contain the sequence "PRSTQ"? Print their entry names and the position of the given sequence.
   c. What is the percentage of the proteins having at least one 3D structure? What is the ratio of the total "X-ray crystallography" structures to "NMR spectroscopy" structures? Is there any other method for this column rather than those two stated above?
   d. What is the average mass to length ratio for this dataset?