EE6435 programming homework 5 (clustering)
Out: April 28, 2020
Due: 11:59PM, May 10th, 2020
Full mark: 50 pts

In this homework, you will apply bottom-up hierarchical clustering algorithm to cluster SARS-CoV-2 genomes.

The similarity between all pairs of sequences is provided to you as a matrix (SCOV2_96_matrix.txt). The first line is the name of these sequences. Then all the other lines contain the pairwise similarity following the order of the first line. See a toy example below:

s1 s2 s3 s4 s5

1.0 0.9 0.98 0.89 0.78

0.9 1.0 0.89 0.87 0.65

…

In this example, there are 5 sequences form s1 to s5. The second line is the similarity between s1 to all five sequences: s1 vs s1, s1 vs s2, s1 vs s3, s1 vs s4, and s1 vs s5. Similarly, the third line contains pairwise similarities between s2 and all others.

It is hard for you to directly apply k-means because you need to design your own method about generating the centroid sequence. Usually, the centroid sequence should be the consensus sequence, which will take you extra programming to get. So, we will apply bottom-up hierarchical clustering. The clustering algorithm will stop when you have just one cluster. Then, using this tree, output 2 clusters, 3, 4, and 5 clusters.

Specific requirements about the input and output.

1. The data can be found on Canvas
2. Use python only. Your program should be named as <your student ID>.py. It should take one parameter, which is the full path + name of the input sequence file. For example,
   <program> /documents/hw5/SCOV2_96_matrix.txt
   <program> /Documents/SCOV2_96_matrix.txt
   Etc.
3. Comment the start and end of the clustering implementation in your code. Don't call any existing APIs.
4. For each k=2 to 5, plot the heatmap similar to the one on page 80. Order the sequences based on their clusters and visualize their similarities.
5. Use "average similarity" between clusters.

6. Submit your code and a pdf format report. The report should contain the following parts:
   a. instructions to run your program (similar to readme)
   b. A description about how you generate different number of clusters from the final tree
   c. For each k=2 to 5, show the sequence IDs inside each cluster. For example, when k=2, show the contents of the two clusters. When k=3, show the contents of the three clusters.
   d. The figures for item 4 for k=2 to 5.

7. We will compare the similarity of your codes. Copying others' codes/reports will lead to 0 for this homework or F for this course. To protect yourself and your friends, keep the codes to yourself only.

Don't hardcode the input file's path because it will make our testing very difficult. -10 for hardcoding the input file.