Outbreak investigation using Machine Learning.

You are given a data set consisting of DNA sequences (the file is available here) of the same length. Each DNA sequence is a sequence of characters from the alphabet 'A','C','T','G', and it represents a particular viral strain sampled from an infected individual. Your goal is to write a code that helps to identify so-called transmission clusters that correspond to ongoing viral outbreaks.

The sequences should be considered as feature vectors and characters - as features. The data set is stored as a fasta file, which is essentially a text file that has the following form:

>Name of Sequence1

AAGCACAGGATGTAATGGTGGGGCCGACCGCCTATTATTCTGATGATTACTTGAGGCCCTCGGAGAGGAAGGGG

>Name of Sequence2

AAGCACAGGATGTAATGGTGGGGCCGACCGCCTATTATTCTGATGATTACTTGAGGCCCTCGGAGAGGAAGGGG

>Name of Sequence3

AAGCACAGGATGTAATGGTGGGGCCGACCGCCTATTATTCTGATGATTACTTGAGGCCCTCGGAGAGGAAGGGG

Here each line starting with '>' symbol contains the name of a sequence followed by the sequence itself in the next line.

You may proceed as follows:

1) Read sequences from the file.

2) Calculate pairwise distances between sequences. Use Hamming distance: it is the number of positions at which the sequences are different (see https://en.wikipedia.org/wiki/Hamming_distance)

3) Project the sequences in 2-D space using Multidimensional Scaling (MDS) based on Hamming distance matrix.

4) Plot the obtained 2-D data points. Estimate the number of clusters *K* by visual inspection.

5) Use *k*-means algorithm to cluster the 2-D data points.

You may use library functions to read data from the file and perform MDS. For parsing sequence files, see e.g. http://biopython.org/DIST/docs/tutorial/Tutorial.html#sec11 (it is a big tutorial, but you need only a part about opening fasta files). For multidimensional scaling in python, see e.g. https://scikit-learn.org/stable/modules/generated/sklearn.manifold.MDS.html. For

K-means clustering should be implemented from scratch.

Your submission must contain:

1) Short report that describes how you did it. The report has to include
   - code of your script;
   - visualization plots for MDS with different clusters highlighted in different colors.
2) Your source code written in Python (.py or .ipynb).

Do NOT use archives! No *zip*, *rar*, *7z,* or *tar*-files are allow in the submission!