DATA MINING (CS F415)

K means and Hierarchical (Agglomerative and Divisive) Clustering

1) Team details:

ROHITH SARANGA (ID: 2017A7PS0034H)

KAUSHIK PERIKA (ID: 2017A7PS0207H)

MAHESH BABU (ID: 2017A7PS0235H)

K Means:-

K-Means clustering intends to partition *n* objects into *k* clusters in which each object belongs to the cluster with the nearest mean. This method produces exactly *k* different clusters of greatest possible distinction.

Hierarchical Clustering:-

i) Agglomerative:

In *agglomerative* or *bottom-up clustering* method, we start with individual points as a cluster. Then, compute the distance between each of the clusters and join the two most similar clusters until we are only left with single cluster consisting of all the points.

ii) Divisive:

In *agglomerative* or *bottom-up clustering* method, we start with individual points as a cluster. Then, compute the distance between each of the clusters and join the two most similar clusters until we are only left with single cluster consisting of all the points.

2) DATASET USED: Amino Acid Sequence

Programming Language used: Python

3) Pre-processing Done:

The data was downloaded from the website and stored in a text file. While reading the data, each line had at its end a new line character that was removed and each line was considered an element of a list named data. At the end of each amino acid sequence was a ‘*’ character which was also removed using list slicing and final data was stored in a dictionary named labels.

In the final dictionary labels, a mapping was created between every unique amino acid label (key in the dictionary) and its corresponding sequence (value of the dictionary). Both the key and the value pairs were in string data type.

4) Formulas used:-

- K means:

Total cost	Sum of distances of all points from their centers.___

- Agglomerative:

	FORMULA
Maximum or complete linkage	Max(d(a, b))
Minimum or single linkage	Min(d(a, b))
Mean or average linkage	$\frac{\text{Sum of all } d(a, b)}{|A| * |B|}$

- Divisive:

Diameter of a cluster	Max(d(a, b))

5) Linkage and distance metric:

Distance Matrix:-

It is an N x N matrix where a point (i, j) denotes the alignment distance between the ith and the jth DNA sequence strings.

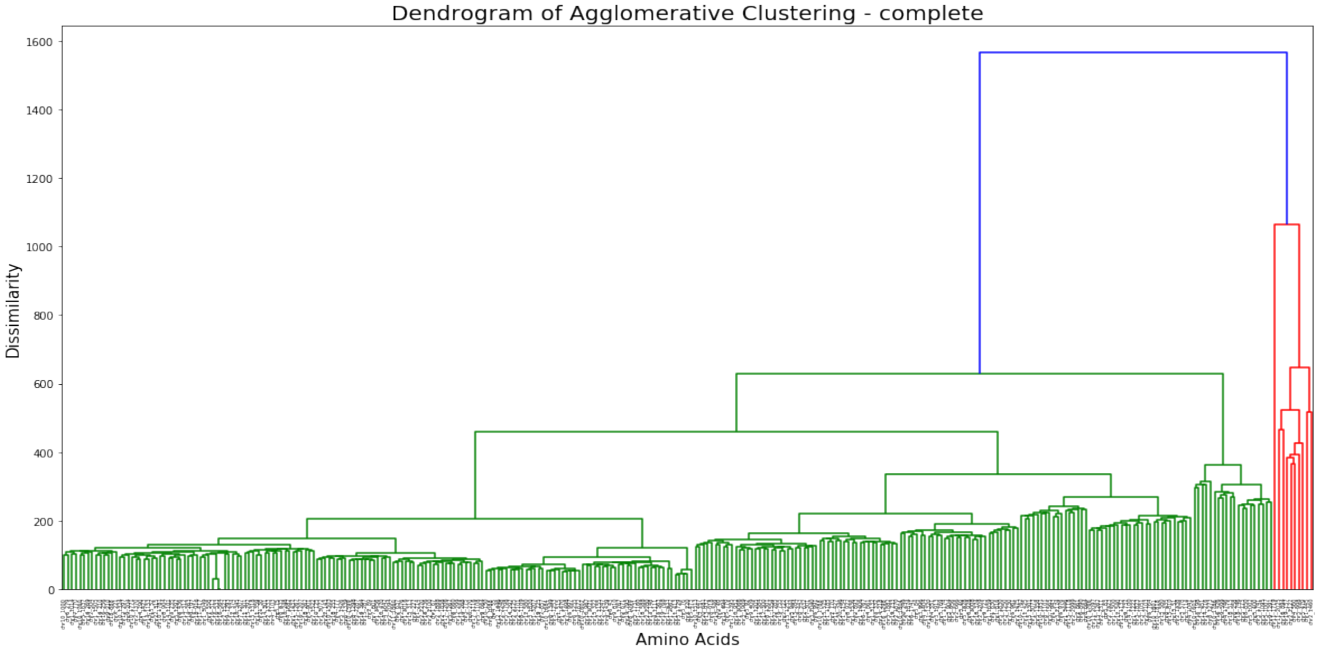
The computation of distance matrix for the entire data set took 18 minutes.

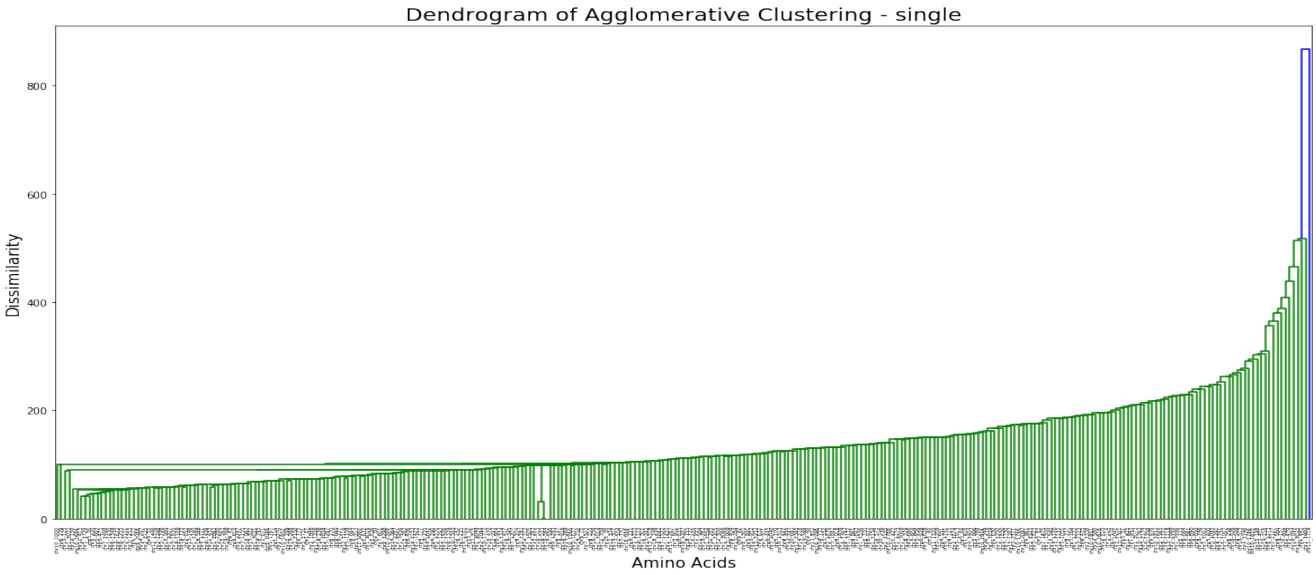
Linkage Matrix:-

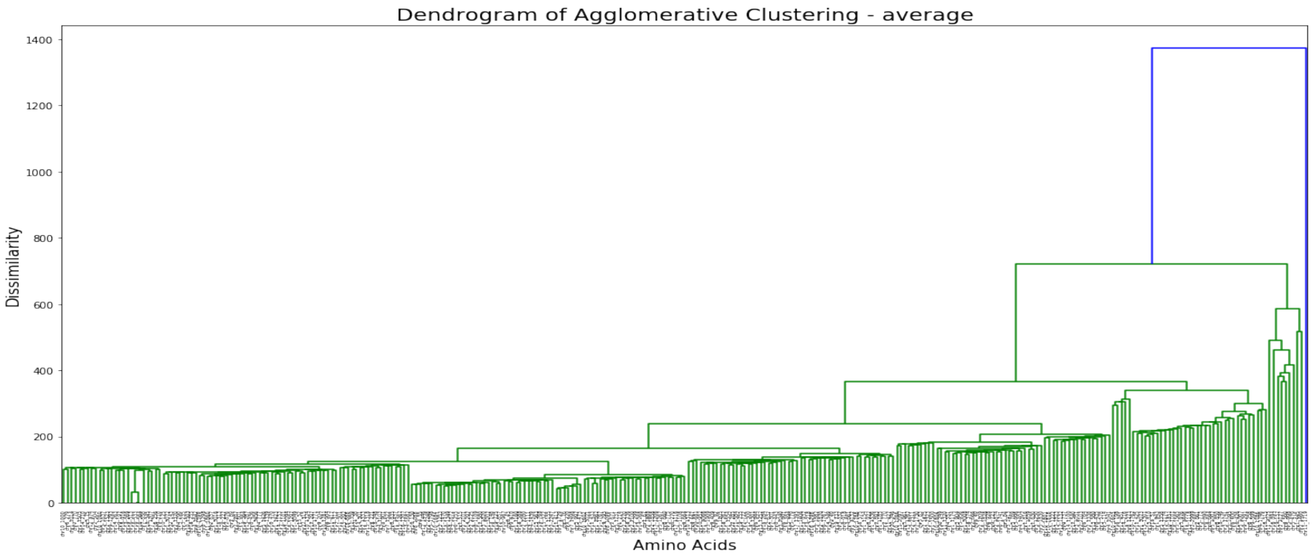
Scipy uses a special matrix called linkage matrix to draw dendrograms. The shape of the matrix is (N-1) x 4, where the ith row represents the merging of two clusters to form the (n+i)th cluster. The first and second columns of the matrix contain the clusters being merged, the third column contains the distance between the two clusters being merged, and the fourth column contains the number of elements in the merged cluster.

6) Comparison of Dendrograms (Top down and Bottom up):-

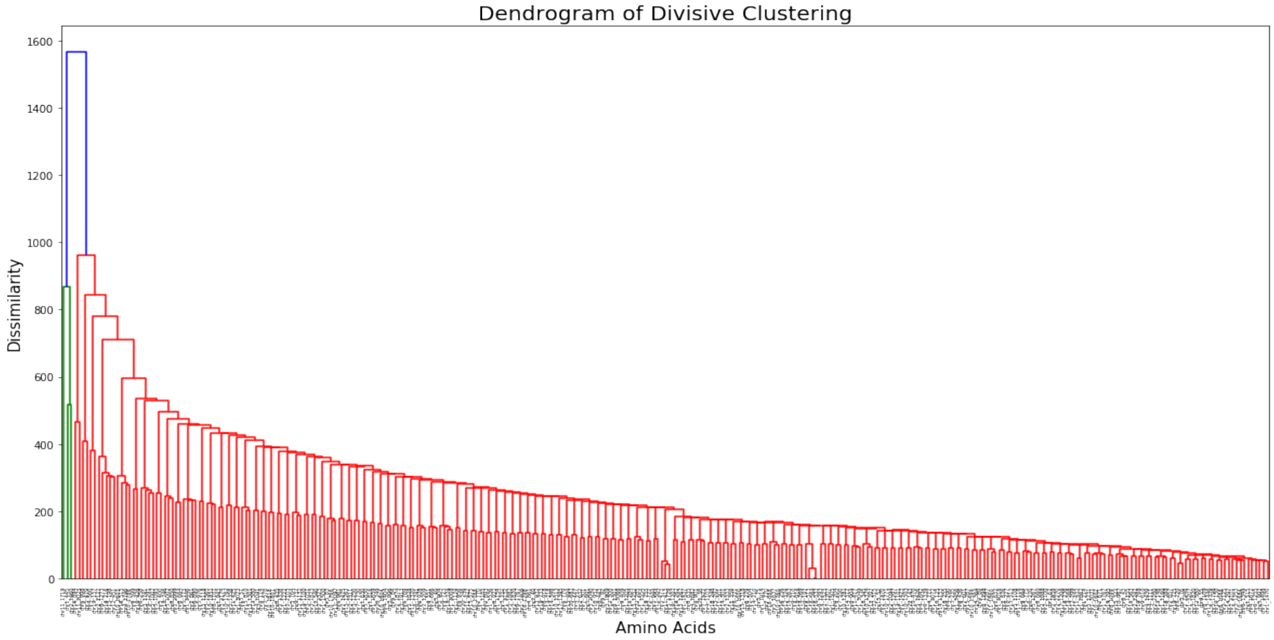
Agglomerative (Bottom up):-







Divisive (Top down) : -



7) Comparison of Dendrograms (K-Means with hierarchical clustering):-

K-Means runs faster than both the bottom-up and top-down approaches of hierarchical clustering.

Agglomerative clustering completes execution in polynomial time. Divisive clustering requires exponential time