Birla Institute of Technology & Science, Pilani
Hyderabad Campus

CS F415: DATA MINING
SECOND SEMESTER 2018-19
Assignment-2

INSTRUCTOR: DR. ARUNA MALAPATI

Submission date and time: 10th March, 2019 23:59 hrs

Maximum marks: 30

The goal of this assignment is to implement K-Means algorithm and also to construct the phylogenetic tree(dendrogram) of the dataset using Agglomerative(bottom-up) and Divisive(top-down) hierarchical clustering. For Agglomerative clustering, use **at least three** linkages from single, complete, average, ward, centroid etc. Your task:

- Compare Agglomerative and Divisive method on the dataset and plot their phylogenetic trees (dendrograms).
- Cluster them using K-Means algorithm. (You are free to choose the values of K and experiment with multiple runs with different mean points).
- After obtaining the clusters from K-Means, compare them with clusters produced by the hierarchical clustering techniques(for all the linkages you've chosen) and highlight the differences for the particular K value.

Datasets:

- **Vertebrate DNA sequences**
- **Amino Acid Sequence**

You are free to you any other dataset provided that it is of comparable size as the datasets provided.

Programming Languages: C, C++, Java, Python

Team Size: 4

Report:

- Name and ID of team members.
- Dataset used.
- Pre-processing done on the data(if any).
- Formulas used.
- Linkage and distance metric used and the type of data it can cluster properly.
- Comparison of dendrogram plot of top-down and bottom-up clustering.
- Comparison of K-Means with hierarchical clustering.

Submission Files:

- Source code files
- Image files of the dendrogram plots
- Report in PDF format
- README

Remarks:

- All submission documents should be zipped together and submitted to CMS through one of the group member's account before deadline. Name of the file should be DM_ASSN1_201x0xxx_201x0xxx_201x0xxx.zip
- All source codes will be checked for plagiarism on Moss (for a Measure of Software Similarity). Any kind of plagiarism will lead to severe penalization.
- You are expected to demo your code and present your results as per the schedule that will be made available on CMS later.

References:

- Libraries for plotting the dendrograms: Dendrogram (Python), SMILE (Java).
- How to find similarity between two sequences
- Divisive Analysis (DIANA), Refer to Section 6.1 (page 253-259)

Evaluation:

- Code & comments (20 marks)
- Output files (5 Marks)
- Report (5 marks)
- Viva (5 marks)

Please contact the following teaching assistants for any queries:

1. Harshith Thonupunoori(f20150071@hyderabad.bits-pilani.ac.in)