# SECOND SEMESTER 2017-2018

# CS F415: DATA MINING

# Assignment 2

Submission Date & Time: 26/03/18, 2359 hrs Maximum Marks: 25

A phylogenetic tree or evolutionary tree is a branching diagram or "tree" showing the inferred evolutionary relationships among various biological species based upon similarities and differences in their physical or genetic characteristics. The goal of this assignment is to construct the phylogenetic tree based on DNA/Protein sequences of species given in the dataset using Agglomerative(bottom-up) and Divisive(top-down) Hierarchical Clustering. In Agglomerative, you start with all points as individual clusters and then keep on combining clusters until required number of clusters are not formed using linkages like single, complete, average, ward or centroid whereas in Divisive, you have all points in one cluster initially and break the cluster into required number of clusters. Your task is to compare Agglomerative and Divisive method on any dataset and with any one linkage and plot their phylogenetic trees (dendrograms).

**Datasets:**

- Human Gene DNA Sequences
- Amino Acid Sequence of Human Gene
- Vertebrate DNA Sequences
- Vertebrate Protein Sequences
  Other similar datasets can be used.

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

**Team Size:** 3

**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

**Submission Files:**

- Source code files
- Image files of the dendrogram plot
- 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_ASSN2_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 not be entertained.
- You are expected to demo your code and present your results as per the schedule that will be made available on CMS later.

**Evaluation:**

- Code & comments (10 marks)
- Output files (5 marks)
- Report (5 marks)
- Viva (5 marks)

**References:**

- Phylogenetic Tree
- Divisive Analysis (DIANA) – Section 6.1 page #253 - #259
- Finding Similarity of DNA Sequences
- Dendrogram – python plotly library (SMILE for Java)

Please contact following teaching assistant for any queries:

Keval Morabia (f20150143@hyderabad.bits-pilani.ac.in)