CSE 3024: Web Mining **Slot: L51 + L52**

Online Submission Deadline: 15th October 2020

Web Content Mining, Web Usage Mining

[4 + 3 + 8]

- **Upload your code and result as a single PDF file in VTOP and MOODLE**
- **File should contain**
  - **Question**
  - **Code**
  - **Result / Output screen (including contents of all generated files)**

---

1. Write a program to show the implementation of agglomerative hierarchical clustering (single, complete and average linkage) using the below mentioned dataset. Show the resultant clusters using graph and dendrogram.
   - Consider Euclidean distance as measure
   - Handle missing values, if any
   - Implement cross validation

   https://drive.google.com/open?id=1FGHlK1Ffn6RvxfMFMhBIYr6o7c-JfXCt

   The detailed description of the dataset is given in the below link:
   https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work

2. Write a program to show the implementation of apriori algorithm using web log usage data for web usage mining purpose. (Consider any publicly available web log data to show the implementation.)

3. Consider the COVID-19 dataset for India given in the following link.

   https://www.kaggle.com/sudalairajkumar/covid19-in-india

   Analyze the dataset to cluster various states in India with respect to *number of active cases, death rate, recovery rate and testing-confirm cases ratio*. [Use any **TWO** clustering algorithms of your choice and provide the performance analysis of the techniques used w.r.t. results obtained.]