# CSCI 6333 Data Mining & Warehousing

**Module 2: Association Analysis**

**Homework Assignment Two**

# Problem: This assignment has Three parts:

* **Part 1:** I ask you to write a program to implement the classical Apriori algorithm to find the most frequent item sets and then generate association rules from these item sets.

The two parameters are given as follows:

*minsup = 30%*

*minconf = 90%*

The most frequent item sets shall be output and saved a text file called *“mfis.txt,”* and the association rules shall also be output and saved in another text file called *“ar.txt.”*

Note: You may use any reasonable heuristic in your implementation to improve the running time performance of the Apriori Algorithm.

* **Part II:** I ask you to sort the associate rules obtained from Part 1 according to confidence values, and then output the top 10 association rules with highest confidence values. These top ten association rules shall be output and saved in the third text file called *“topar.txt.”*
* **Part III:** In an MS word file, I ask you to evaluate the top ten association rules and briefly explain whether each of these rules is meaningful or not. Save your word file as *“explan.doc”* or *“explan.docx.”*
* Data set: The data set is the 1984 Congressional Voting Record Database archived at

UCI Machine Learning Repository:

|  |  |
| --- | --- |
| **Congressional Voting Records Data Set**  *Download*: [Data Folder](https://archive.ics.uci.edu/ml/machine-learning-databases/voting-records/), [Data Set Description](https://archive.ics.uci.edu/ml/machine-learning-databases/voting-records/house-votes-84.names)  **Abstract**: 1984 United Stated Congressional Voting Records; Classify as Republican or Democrat | https://archive.ics.uci.edu/ml/assets/MLimages/Large105.jpg |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 435 | **Area:** | Social |
| **Attribute Characteristics:** | Categorical | **Number of Attributes:** | 16 | **Date Donated** | 1987-04-27 |
| **Associated Tasks:** | Classification | **Missing Values?** | Yes | **Number of Web Hits:** | 69495 |

**Program Requirement:**

Your program shall be coded in such a streamlined way that it takes the input data file that confines with the same name given at the UCI site and generates the output files aforementioned in the Problem section. The grader shall not be asked to do any intervention with the execution of your program.

**Programming Language:**

C++, or Java, or Python

# Alternative Approaches to Homework Assignment Two

The background of students in this class is diverse. If your programming skill is not strong enough at this point, I suggest you choose one of the following two alternative approaches:

* You can team with one student who is good at programming to complete this assignment.
* You can research for existing data mining tools/software packages and use these to solve Part I and Part II. In this case, you shall cite clearly what tools/software packages are used in your solutions.