# **BIG DATA –Spring 2019**

# **ASSIGNMENT 4**

**Due Date: 7th April 2019 till 10:00 PM on google classroom.**

|  |
| --- |
| **Submission Details:**  For Question 1: Upload the Word Document containing Association rule and their explanation  For Question 2: Upload properly intended and commented Source code of Map Reduce program  Upload on Google Classroom with your roll number.  **Data Set:**  Consider the same Bank dataset that was provided in the Assignment 2.  **You can work alone or in a group of two. Plagiarism from other students or internet will result in -5% absolute.** |

**QUESTION 1: Association Rule Mining using WEKA**

1. **Data Preprocessing**
   1. Explore the dataset and understand the types of each attribute
   2. Identify issues in data quality like missing value, inconsistently, noise etc
   3. Apply different data preprocessing techniques on different attributes. Techniques such as discretization, normalization, data smoothing, and data reduction **as required**
   4. Association Rule Mining can be applied only on a binarized dataset. Convert the dataset into form suitable for Association Rule Mining.
2. **Association Rule Mining**

After preprocessing the given dataset your task is to find pattern in the data using association rule mining. Experiment with different parameters so that you get at least 30-40 strong rules (e.g., rules with high lift and confidence which at the same time have relatively good support). Select the best 10 most "interesting" rules and for each specify the following:

* an explanation of the pattern and why you believe it is interesting and how can it be helpful
* any recommendations based on the discovered rule that might help the user

**Note**: The top 5 most interesting rules are most likely not the top 5 in the result set of the Apriori algorithm. They are rules that, in addition to having high support, lift, and confidence, also gives some non-trivial, useful information based on the underlying business objectives.

**Note: You will find the Weka and Data Mining tutorial at**

<http://facweb.cs.depaul.edu/mobasher/classes/ect584/WEKA/index.html> **website quite helpful.**

**QUESTION 2: Frequent Pattern Mining in Map Reduce**

Input a **preprocessed** bank dataset file and write an **efficient Map Reduce** Job to generate all frequent patterns.

Use the idea of Map Reduce version of SON Rule Mining Algorithm discussed in class. For details consult Book Mining of Massive dataset chapter 6 (pg 229).

Use your knowledge on Map Reduce and add in details to the above algorithm to make it efficient. Details such as Combiner, Partitioner (if required), Comparator (if required), Associative memory HashMaps or Hashtree etc.

**Note:** You can control the size of the input file given to each Mapper by modifying the Input-Split size of a file in Map Reduce. If the maximum input-split size is less than hdfs block size, then each split will be less than hdfs block. Consult Book Hadoop Definitive guide Pg 228-229