Birla Institute of Technology & Science, Pilani
Hyderabad Campus

SECOND SEMESTER 2017-2018
CS F415: DATA MINING
Assignment 1

Submission Date & Time: 19/02/18, 2359 hrs Maximum Marks: 25

The goal of this assignment is to generate frequent item sets and interesting association rules using Apriori algorithm. Support and confidence values are not fixed, experiment with different values of support and Confidence to generate frequent item sets and association rules. **You are also expected to experimented with at least one alternative measure for support and confidence which is given in Table 6.11 and 6.12 of the text book.** The output must strictly be stored in a file in the following format:

- File1: **Freq_Items_sup:s,conf:c** (experiment with different s and c values)
  Format: Freq Itemset (count)
  Example: safety=med, persons=4, class=acc, (90)
- File2: **Assn_Rules_sup:s,conf:c**
  Format: LHS (item set (count)) ---> RHS (item set (count)) - confidence value
  Example: maint=vhigh, persons=4 (144) ----> CLASS=unacc (1210) - conf(0.75)

Datasets:

- Groceries Market Basket Data
- Car Evaluation Database
- Mushroom Database
  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: 3

Report:

- Name and ID of team members.
- Dataset used.
- Pre-processing done on the data.
- Formulas used.
- Number of frequent item sets & association rules for different values of support & confidence (atleast 3 values).

Submission Files:

- Source code files

- Frequent Item sets & Association Rules for different support & confidence values
- Report in PDF format
- README

Remarks:

- Strictly follow the format provided for the output files. Any other format will result in deduction of marks.
- 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 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)
- Frequent Itemsets & Association rules (5 Marks)
- Report (5 marks)
- Viva (5 marks)

Please contact following teaching assistants for any queries:

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