**CS 536/CS 432 – Data Mining**

**Assignment 2**

**Due: March 9 (Thursday) at 12 midnight**

**Instructions:** (1) You may discuss the assignment with others. However, you MUST do and submit your OWN work. (2) Submit a soft-copy report to the submission folder on LMS. Include report and code needed to reproduce your results.

1. **Apriori and FP-Growth Algorithms’ Practice (20 points)**

Consider the following transactional database:

|  |  |
| --- | --- |
| TID | Items |
| 1 | BD |
| 2 | ABD |
| 3 | AC |
| 4 | EF |
| 5 | CDEF |
| 6 | BE |
| 7 | AE |
| 8 | AEF |
| 9 | ADE |
| 10 | AE |
| 11 | BDF |
| 12 | DE |
| 13 | DFF |
| 14 | CDE |

1. Find all frequent itemsets using the Apriori algorithm. Assume minimum support count is 3.
2. Find all frequent itemsets using the FP-growth algorithm. Assume minimum support count is 3.
3. Identify all closed and max itemsets.
4. Generate all strong association rules from the longest closed pattern(s) found in the database. Assume minimum confidence is 60%.
5. **Frequent Itemset Mining Using RapidMiner (40 points)**

Experiment with RapidMiner’s implementation of Apriori and FP-growth algorithms. Apply these algorithms to the Marketing Campaign dataset provided on LMS. Read the description of the dataset and do the following:

1. For Apriori generate rules and itemsets for (i) default parameter values, (ii) rules = 50, (iii) confidence = 0.7; rules = 50, (iv) minimum support is 0.1.
2. For FP-growth, generate itemsets for (i) default parameter values, (ii) minimum support = 0.1, (iii) find min number of itemsets is unchecked, and (iv) find min number of itemsets is unchecked; minimum support = 0.1.
3. From results in (a), separate out all strong classification rules, i.e., rules that contain the class attribute (survey answer) on the right-hand-side.
4. Provide a brief summary and interpretation of the results.
5. **Sequential Pattern Mining Using SPMF (30 points)**

SPMF is an open-source Java-based tool that implements many frequent pattern and strong rule mining algorithms. Here, we will experiment with the GSP and PrefixSpan algorithms for sequential pattern mining. Apply this to the data set given in question 1 above. Suppose there are 5 sequences in the data set as follows: transactions 1 to 2, transactions 3 to 7, transactions 8 to 10, transactions 11 to 12, transactions 13 to 14. You will have to convert the data set into the appropriate format for use in SPMF. Find frequent sequences for various minimum support values. For at least one minimum support value, verify that the the computed sequences are correct.

1. **Exercise 6.14 from the book (Han et al., 3rd Ed) (5 points, bonus for undergrads)**
2. **Exercise 7.7 from the book (Han et al., 3rd Ed) (5 points, bonus for undergrads)**