Homework 2

CS 6720

Due: 9th Feb by the time class starts. No late submissions will be allowed. Please start early.

1. This question is on frequent itemset mining. Implement the Apriori Algorithm for frequent itemset mining. Apply it on the Dataset: http://fimi.ua.ac.be/data/retail.dat. Details about the dataset can be found at http://fimi.ua.ac.be/data/retail.pdf.
    a. Report the frequent itemsets at 10% support threshold. (30 points)

    Please follow the following format strictly while submitting results for this question as evaluation will be automated.

    i. Each frequent itemset must be on a new line. The items must be space separated and in ascending order.

    ii. The itemsets must be grouped based on cardinality and then ordered within lexicographically. For example, if (2),(4),(5),(7),(12),(2,5),(2,7),(5,7),(5,12),(2,5,7) are the frequent itemsets, your output should be as given in this file .

    iii. Please name your file RollNo_1a.txt. For example, if CS10B061 is your roll number your file should be named CS10B061_1a.txt.

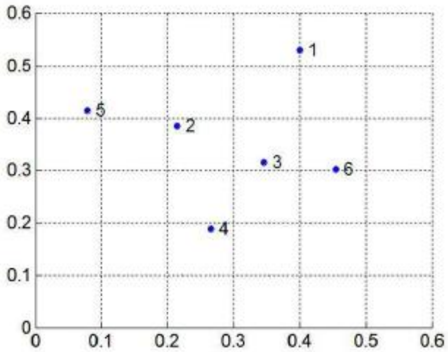
    b. Compare the performance of your implementation with FP-tree (you are free to use an available implementation here. But the FP-tree implementation and your Apriori implementation must be in the same language) you vary the
        i. Support threshold. Explain the results that you observe. (10 points)
        ii. Dataset size in terms of number of transactions at 10%, 25%, 50% and 100% of the entire dataset. You can randomly sample X% of the transactions. Explain the results that you observe. (10 points)
        iii. Number of items. Sort the items in descending order of their frequencies. Plot the running times against the top 10%, 25%, 50% and 100% of the items. In X%, for any transactions, you only consider items that are in that top X%. (10 points)

2. In the given dataset, find the optimum number of clusters by analyzing the knee/elbow point. To get your dataset use the jar file available in this link. To generate the dataset type the following command on Terminal

    java -jar genData.jar <last 2 digits of your roll number>.

    Produce the plot and clearly explain how you arrived at the number of clusters in the dataset. (20 points)

3.
    a. Draw the dendrogram for single linkage clustering on the data below (5 points):



Point	x	y
1	0.40	0.53
2	0.22	0.38
3	0.35	0.32
4	0.26	0.19
5	0.08	0.41
6	0.45	0.30

b. What is the computation complexity of clustering n points without using the support of any other data structure? (5 points)

c. Propose an O(n^2logn) algorithm. Explain your reasoning clearly. (10 points)