COMPUTER SCIENCE 4373
Assignment #2

Points: 100   **Weight**: 3%

Due: Friday, February 9, 2018 at 2:00 pm in class as well as on BlackBoard

Note: Late assignment will not be accepted without instructor's pre–approval.

Instruction: This assignment should be completed individually. Please make sure your answer is legible (and preferably formatted using MS Words or LATEX/LYX). If a question requires you to follow an algorithm, show a clear trace of the algorithm. If the algorithm is iterative, show the details in the first two iterations. For each of the remaining iterations, show the status of the algorithm at the end of the iteration. Please also submit to the BlackBoard a single your_name_hwk02.zip file that contains a PDF or a Word version of your solutions (**no scanned image please**), and the source code, the input data, and the output of your program.

The following questions are based on this transaction database.

TID	items bought
T100	{M, O, N, K, E}
T200	{D, O, N, K, E, Y}
T300	{ M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, K, I ,E}

1. [**40**] Let $sup_{min} = 60\%$ and $conf_{min} = 80\%$.

   (a) Find all frequent itemsets by following the Apriori Algorithm.

   (b) Identify closed and maximal frequent itemsets.

   (c) Draw the hash tree for candidate 2-itemsets, assuming that the hash function will map A, E, M, Y to the left branch, C, I N to the center branch, and D, K O to the right branch.

   (d) Identify the hash tree notes visited when counting the support of 2-itemsets for transaction T400

2. [**20**] Let $sup_{min} = 60\%$ and $conf_{min} = 80\%$. Find all frequent itemsets by following the FP-growth algorithm.

   (a) Find the F-list and draw the FP-tree.

   (b) For each projected database, list all frequent itemsets that can be enumerated from the database.

3. [**20**] Find all association rules. Each rule should be written as

$$r\# \; buys(item_1), buys(item_2) \Rightarrow buys(item_3) \; [s, c]$$

where $r\#$ is the id of the rule, s is the support, c is the confidence, and $item_i$ is an item (such as, "M", "O",etc.).

   (a) Identify all strong association rules (that is, those rules with support s and confidence c higher than the thresholds).

   (b) Compute the lift measure for each strong association rule.

4. [**20**] Convert the transaction data table into an hwk02.arff file, in which each item is a column with either a value t or ? (that is, with a type { t}). For example:

```
@relation hwk02
@attribute A { t}
...
@data
?,t,?,?,t,...
```

Use Weka Explore to load the data, and run the Aproiri and FPGrowth algorithms to learn interesting association rules. Do it for at least three sets of minimum support and confidence. For each algorithm, report the following items.

   (a) The commands used to run the learning tasks

   (b) The output from the runs