Name: _____ USC ID: _____

Quiz 4: Frequent itemset discovery (10 points), 15 minutes

1. [3 points] Explain what a maximally frequent itemset is. Given an example of such an itemset. Suppose that {A,B,C,D} are all the unique items.

   I is maximally frequent if I is frequent and none of I's supersets (if any) is frequent. [1 point]
   E.g.,

A,B
A,B,C
B,C
A,B,C,D
A,C,D

   Suppose that the support threshold is 2
   Maximally frequent itemset: (A,B,C), (A,C,D) [2 points]

2. [3 points] Explain what an association rule is. Give a real-world example (i.e., using meaningful items such as bread, milk, and coke) of a rule which has high confidence but is not interesting. Explain your answer.

   Suppose that I -> j, I is a set of items and j is an item. If all of items in I appear in some basket, then j "likely" appears in that basket too. [1.5 points]

   E.g., conf({milk, bread} -> coke) = 0.7, sup_ratio({coke}) = 0.8 [1.5 points]

3. [2 points] Discuss the tradeoff between two types of storage method for pairs and their counts: triple-based and triangular matrix.

   triangular matrix: avoid storing counts twice [1 point]
   store as an array with n(n-1)/2 counts
   require space for n(n-1)/2 integers
   triple-based matrix: more economical if matrix is sparse [1 point]
   space for hash table
   p = # of item-pairs that actually occur in baskets
   require 3p integers
   Triples method is better when 3p < n(n-1)/2

4. [2 points] Derive the formula for converting the (integer) item pairs, e.g., (i, j), into the index into a 1-D (triangular) array that actually stores the counts of pairs. Assume that there are n unique items.

   x = (i-1) *n +j – (1+2+ ... + i) [2 points]
   = (i-1)*(n-i/2) +j - i