Take home Quiz

1. [1 points] Discuss the tradeoff between two types of storage method for pairs and their counts: triple-based and triangular matrix.
   Answer - triangular matrix: avoid storing counts twice [0.5 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 [0.5 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
2. [1 points] How can you apply Market Basket analysis on plagiarism detection based on document similarity? Explain.
   Answer - Baskets = sentences [0.5 points]
   Items = documents containing those sentences [0.5 points]
   Item/document is in a basket if sentence is in the document. Look for items that appear together in several baskets. Items (documents) that appear together too often could represent plagiarism.
3. [5 points] Here is a collection of twelve baskets. Each contains three of the six items 1 through 6. {1, 2, 3} {2, 3, 4} {3, 4, 5} {4, 5, 6} {1, 3, 5} {2, 4, 6} {1, 3, 4} {2, 4, 5} {3, 5, 6} {1, 2, 4} {2, 3, 5} {3, 4, 6} The support threshold is 4. The hash function is i x j mod 11. Using the PCY Algorithm, you need to show 1. frequent singles, 2. frequent buckets, and 3. frequent pairs.
   Suppose we run the **Multistage** Algorithm on the data in (1), with the same support threshold of 4. The first pass is the same as in (1), and for the second pass, we hash pairs to nine buckets, using the hash function i + j mod 9. Determine the counts of each of the 9 buckets. (1.5 pt) Does the second pass reduce the set of candidate pairs, why?
   First Part - [2 points]
   Frequent Singles: 1, 2, 3, 4, 5, 6
   Frequent Buckets - 1, 2, 4, 8
   Frequent Pairs - (2,4), (3,4), (3,5)
   Second Part - [2 points]
   Bucket Count

0	0
1	3
2	2
3	2
4	0
5	2
6	4
7	4

8 5

Yes, the second pass reduces the set of candidate pairs since some pairs will no longer be candidates (e.g. (4,6) has count=3 which is lesser than the support threshold) - [1 point]

4. [2 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.
   Answer - Suppose that I -> j, I is a set of items and j is an item. If all of the items in i appear in some basket, then j "likely" appears in that basket too. [1 points]
   E.g., conf({milk, bread} -> coke) = 0.7, sup_ratio({coke}) = 0.8 [1 points]
5. [0.5 point] Confidence is an indication of how frequently the items appear in the baskets?
   False
6. [0.5 points] h1(x) = 2*x + 14 % 3, h2(x) = 4*x + 7 % 3 are good choices of hash functions for the **multi-hash** algorithm?
   False