Tuesday Quiz

1. [2 points] For the following question, consider the entire set of items contains: A, B, C, D, E,..., J (a total of 10 items). In the A Priori algorithm, how much memory do you need if you use
    a. Triangular-Matrix method
    b. Triples Method to count the occurrence of each possible pair assuming only 1/4 of the pairs (doublets) have an occurrence > 0?

    (you can assume that a counter uses 4 bytes) **(Just write the number for the answer)**

    Ans a.) 10 * 9/2 * 4 = 180 *bytes*

    b.) 10 * 9/2 * ¼ = 11.25 = 11 pairs

    11*12= 132 bytes

    If you have assumed there are 11.25 pairs, it is wrong as 11.25 pairs has no physical meaning. I have considered the answer to be correct for both ceil(11.25) = 12 and floor(11.25) = 11.

    It is always important to consider the physical meaning of the quantities in consideration and ponder whether these values actually make sense or not.

2. [0.5 point] All high-confidence rules are interesting? False
3. [0.5 point] In Apriori, If item i does not appear in s baskets, then no pair including can appear in s baskets? True
4. [1 point] In the **PCY** algorithm, what should be the conditions for a pair {i, j} for being a candidate pair?
   Both i and j are frequent items [0.5 points]
   The pair {i, j} hashes to a bucket whose bit in the bit vector is 1 [0.5 points]
5. [1 points] How can you apply Market Basket analysis on the 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.
6. [4 points] Consider the following basket data and a support threshold s = 2, answer the following questions.

$B_1 = \{m, c, b\}$ $B_2 = \{m, p, j\}$

$B_3 = \{m, c, b, n\}$ $B_4 = \{c, j\}$

$B_5 = \{m, p, b\}$ $B_6 = \{m, c, b, j\}$

$B_7 = \{c, b, j\}$ $B_8 = \{b, c\}$

Find all frequent itemsets with set size <= 3. Write down two association rules and their confidence and interest numbers. One of your association rule should be derived from a frequent pair (i.e., X->Y), and the other one should be derived from a frequent triplet (i.e., X, Y->Z)

{m}, {c}, {b}, {p}, {j},
{m, c}, {m, b}, {m, p}, {m, j}, {c, b}, {c, j}, {b, j},
{m, c, b}, {c, b, j}.
Total 2 points for all correct pairs. -0.25 for every 2 missing / incorrect

0.5 points for confidence, 0.5 points for interest for 2 examples

7. [1 point] Let h1(x) = 2*x + 14 % 3, h2(x) = 4*x + 7 % 3. In the **multi-hash** algorithm, are these 2 hash functions a good choice to use? Explain your answer in detail.
   No, they are not a good choice to use for the multi-hash algorithm. These 2 hash functions are DEPENDENT, hence they do not satisfy the purpose of the multi-hash algorithm.
   0.5 points for right answer
   0.5 points for right explanation