Quiz #3: Frequent Itemsets Week 1

Name: _____ ID: _____

1. Write a MapReduce program that multiplies two matrices A and B in **one** stage. You can assume that the matrices are provided to you in a file in a sparse matrix format. Each line of the file represents an element in a matrix. For example, a line: ['A', 0, 0, 1] indicates that A[0, 0] = 1. You may assume that both matrices are 5 x 5. (2 pts)

In the map phase: (1 point)
for each element (i,j) of A,
    emit ((i,k), A[i,j]) for k in 1..5
Better: emit ((i,k), ('A', i, j, A[i,j])) for k in 1..5
Or just emit ((i,k), ('A', j, A[i,j])) for k in 1..5
for each element (j,k) of B,
    emit ((i,k), B[j,k]) for i in 1..5
Better: emit ((i,k), ('B', j, k, B[j,k])) for i in 1..5
Or just emit ((i,k), ('B', j, B[j,k])) for i in 1..5

In the reduce phase, (1 point)
    emit key = (i,k) value = Sumj (A[i,j] * B[j,k])

2. Consider the following input file of basket data and a support threshold **s = 2**, answer the following questions.

$\mathbf{B_1 = \{m, c, b\}}$ $\mathbf{B_2 = \{m, p, j\}}$
$\mathbf{B_3 = \{m, c, b, n\}}$ $\mathbf{B_4 = \{c, j\}}$
$\mathbf{B_5 = \{m, p, b\}}$ $\mathbf{B_6 = \{m, c, b, j\}}$
$\mathbf{B_7 = \{c, b, j\}}$ $\mathbf{B_8 = \{b, c\}}$

Find all frequent itemsets with set size <= 3 (1 pt).

{m}, {c}, {b}, {p}, {j}
{m, c}, {m, b}, {m, p}, {b, j}, {m, j}, {c, b}, {c, j}
{m, c, b}, {c, b, j}
* lose 0.5 point for each missing/wrong itemset

Write down one association rule and its confidence and interest numbers. Your association rule should be derived from a frequent pair (1 pt each)
rule 1 point, confidence 1 point, interest 1point
If only writing down the rule formula, lose 3 points
For example
1) {m} -> {c} confidence : 3/5 = 0.6 interest : |3/5 - 6/8| = 0.15
2) {m, c} -> {b} confidence : 1 interest: 1 - 6/8 = 0.25
3) {m} -> {b} c= 4/5 I=0.05
4) …

3. Here is a collection of 6 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 } The support threshold is 2. The hash function is i x j mod 11. Using the **Apriori Algorithm**, you need to show 1. frequent single items and 2. frequent pairs. (1 pt) Using the **PCY Algorithm**, you need to show 1. frequent buckets and 2. frequent pairs. (2 pts)

Apriori:
Frequent single items: {1},{2},{3},{4},{5},{6} (0.5pt)
Frequent Pairs: {1,3}, {2,3}, {2,4}, {3,4}, {3,5}, {4,5} {4,6} (0.5pt)

PCY:
Frequent Buckets: {3,4,6,8,1,9,2} (0.5pt)
Frequent Pairs: {1,3}, {4,6}, {3,5}, {2,3}, {2,4}, {4,5} {3,4} (0.5pt)

4. Given items apple, beer, tea, and coke if we need to count all possible pairs of them, we can use the one-dimensional vector to store the counters (triangular matrix). Show how you 1. assign an index to each item, 2. Use the lexicographic order to order the vector component, 3. an example of how to find the counter for (apple, tea) (e.g., the counter of (apple, beer) is at position 0 of your vector) (2 pts)

For example:

1. (0.5 point) Assign 1,2,3,4 to apple, beer, tea, coke.

2. (0.5 point) Keep pair counts in lexicographic order: {1,2} , {1,3}, {1,4}, {2,3}, {2,4}, {3,4}

3. (1 point) Pair {i, j} is at position $(i-1)(n-i/2)+j-i$, so {apple, tea} ({1,3} is at position 2. （if starts from position 1）