GE2324 Assignment3, 2024/25B

Due: **14-Apr-2025** (Week 13 Monday) 20:00
Each 24 hours late submission halves the score.
(That is, no deduction of marks within the first 24 hours). 0 marks after 3 days.

Question 1 (30 points): We have coordinates for 10 stations:

Station	X	Y
A	0.3	2.3
B	1.1	3.8
C	2.9	0.6
D	3.5	7.6
E	3.2	5.1
F	4.0	8.3
G	6.1	2.1
H	6.7	3.6
I	7.6	6.9
J	7.1	3.8

a) Adopt Euclidean distance, calculate the distance matrix of the 10 stations. (5pts)

b) Adopt Manhattan distance, calculate the distance matrix of the 10 stations. (5pts)

c) According to the matrices obtained in a) and b), perform hierarchical clustering respectively. Write down your detailed steps and draw the hierarchy trees. [Note: There are three commonly used distance metrics to measure distance between each cluster in hierarchical clustering: Single Linkage, Complete Linkage, and Average Linkage. In this question, please use Single Linkage. For two clusters R and S, the single linkage returns the minimum distance between two points i and j such that i belongs to R and j belongs to S.] (20pts)

Question 2 (30 points): We have 6 baskets of items:

Transaction ID	Items
1	A, B, C
2	E, D, A
3	A, E
4	B, C
5	A, B, E
6	C, B, D

a) Based on the table below, use the Apriori algorithm to find the frequent itemsets. Assume that minimum support s=2. (10pts)

b) Find all association rules which have support >= 30% and confidence >= 0.6. Write down your detailed steps. (10pts)

c) Calculate the interest of the rule(s) that you found in b). (10pts) (Note: Calculate the expectation of an item using the % that the item appears among the baskets)

Question 3 (40 points): We have 3 documents:

Document 1: The cat sits on the mat and looks very calm.
Document 2: The cat sits calmly on the mat, while the dog plays nearby.
Document 3: On the grass, the dog plays joyfully as the cat watches quietly.

a) Find the set of 3-shingles for each document, ignore punctuation and case. (5pts)

b) Construct a 3-shingle matrix M to visualize the relationship of the 3 documents, where row is each of the 3-shingles across all documents and column is each of the documents; M_{ij}=1 if document j contains shingle i, otherwise M_{ij}=0. Calculate Jaccard similarity for each pair of documents based on the matrix M. (15pts)

c) For the input matrix (left) and four random permutations (right) shown below:

Item1	Item2	Item3	Item4
1	0	1	0
0	0	1	1
1	1	0	1
0	1	0	0
1	0	1	1
1	0	0	0
0	1	0	1
Input Matrix			

Hash1	Hash2	Hash3	Hash4
4	3	5	2
2	7	1	5
6	1	4	3
1	5	6	7
5	2	2	1
3	6	7	4
7	4	3	6
Random Permutations			

Calculate the item pairwise Jaccard similarity using the input matrix. Apply the Minhashing method to obtain the signature matrix of items and compute the Jaccard similarity according to the signature matrix. Compare the Jaccard similarity of the signature matrix and that of the input matrix, is minhashing a good approximation to the true Jaccard similarity in this case? Why? (20pts)