Module Code	Examiner	Department	Tel
INT402	****	Intelligent Science	****

1st SEMESTER 22-23 FINAL EXAMINATION

Postgraduate

Data Mining and Big Data Analytics

TIME ALLOWED: *2 hours*

INSTRUCTIONS TO CANDIDATES

1. This is a blended open-book exam and the duration is 2 hours.
2. Total marks available are 100. This accounts for 70% of the final mark.
3. Answer all questions. Relevant and clear steps should be included in the answers.
4. Only English solutions are accepted. For online students, answers need to be handwritten and fully and clearly scanned or photographed for submission as one single PDF file via LEARNING MALL.
5. Online students should use the format "Module Code-Student ID.filetype" to name their files before submitting to Learning Mall. For example, "INT402-18181881.pdf".

Question 1

[25 points in total]

(1) Design a map-reduce algorithm to count the numbers of different words that show up in a large input document.

[6 points]

(2) Explain what the association rule is in data mining and what is the interest of an association rule $I \rightarrow j$.

[6 points]

(3) For frequent item pairs, if we use a triangular matrix A to store pairs, we count pair of items $\{i, j\}$ only for $i < j$, and we keep pair counts in lexicographic order. If the total number of items n is 20, what is the index (position) k for a pair $\{i, j\}$? If $k = 50$, what are the possible pairs $\{i, j\}$?

[6 points]

(4) What is the monotonicity of itemsets? Describe the A-Priori algorithm, and explain why the A-Priori algorithm works.

[7 points]

Question 2

[25 points in total]

The following figure gives a matrix with seven rows.

Element	S1	S2	S3	S4
0	1	0	0	1
1	1	1	1	0
2	0	0	1	1
3	0	1	1	0
4	1	1	0	1
5	0	0	1	0
6	1	1	1	1

(1) Compute the permutations of indices of elements using the following three hash functions respectively

$$\begin{aligned} &\text{i. } h_1(x) = (3x+1) \bmod 7 \\ &\text{ii. } h_2(x) = (5x+3) \bmod 7 \\ &\text{iii. } h_3(x) = (6x+2) \bmod 7 \end{aligned}$$

[6 points]

(2) Compute the signature matrix using the three permutations from the hash functions in (1)

[6 points]

(3) What are the true Jaccard similarities for all pairs among $\{S1, S2, S3, S4\}$? What are the Jaccard similarities of the signature matrix for all pairs among $\{S1, S2, S3, S4\}$? How to decrease the difference of similarities between the true similarities and similarities of the signature matrix?

[8 points]

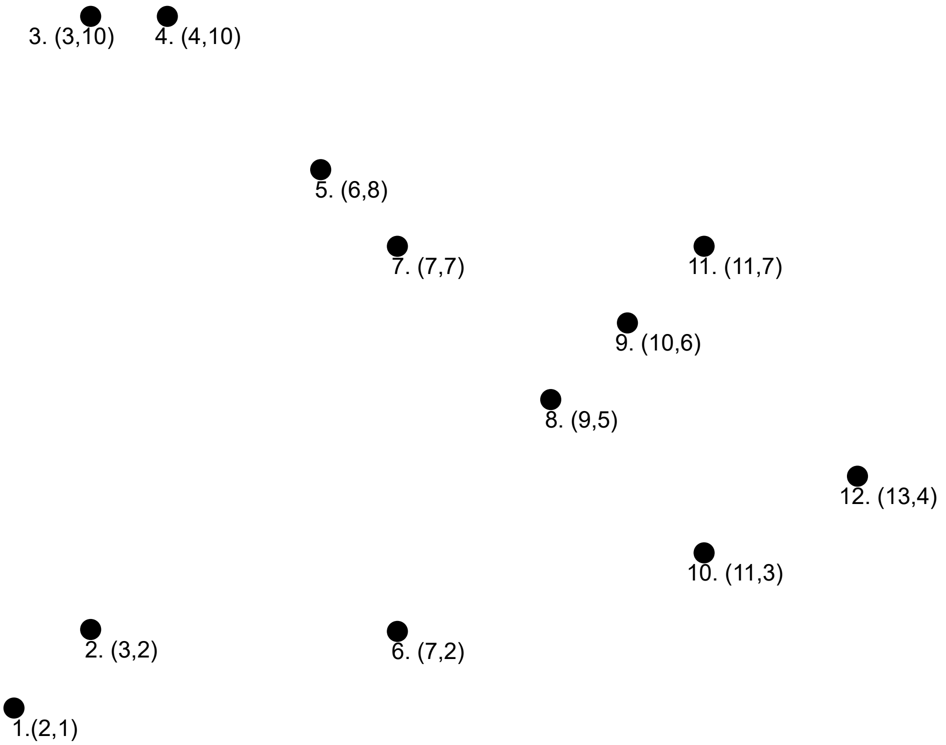
(4) Suppose we use 80 hash functions to generate 80 rows in total for the signature matrix. For locality-sensitive hashing, if we split the signature matrix into 20 bands, say if columns S_4 and S_5 have a similarity of 70%, what is the probability that S_4 and S_5 are hashed to at least 1 common bucket? If we set the similarity threshold $s = 0.7$, what is the probability that S_4 and S_5 being similar pairs are false negatives?

[5 points]

Question 3

[28 points in total]

The following figure shows 12 two-dimensional points in Euclidean space, e.g. ***1. (2,1)*** means the 1st point whose coordinates are (2,1), ***8. (9,5)*** means the 8th point whose coordinates are (9,5). Use a basic hierarchical clustering algorithm with Euclidean distance to cluster these twelve points.



(1) What is the centroid for a cluster that contains points $\{5, 7, 8, 9, 11\}$ and why? What is the clustroid for a cluster that contains points $\{1, 2, 3, 4\}$ and why?

[8 points]

(2) Take the nearness of two clusters to be the **largest** distance between any two points, one from each cluster. Plot a dendrogram showing the clustering result. Note that the height of the dendrogram does not need to be accurate.

[10 points]

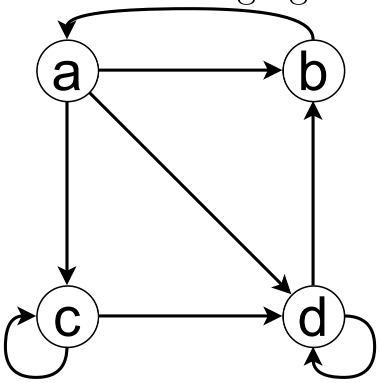
(3) Take the nearness of two clusters to be the **average** distance between all pairs of points, one from each cluster. Plot a dendrogram showing the clustering result. Note that the height of the dendrogram does not need to be accurate.

[10 points]

Question 4

[22 points in total]

Given a web graph shown in the following figure:



(1) Calculate the transition matrix of this graph.

[6 points]

(2) Assuming no taxation. Show the results of four iterations using the power iteration method.

[8 points]

(3) Assuming $\beta = 0.8$. Show the results of four iterations using the power iteration method.

[8 points]

THE END OF EXAM