Xi'an Jiaotong-Liverpool University
西交利物浦大学

Module Code	Examiner	Department	Tel
INT402	Xi Yang	INT	1506

1st SEMESTER 2021/22 RESIT 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.**

5、 **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. Online students should use the format "Module Code-Student ID.filetype" to name their files. For example, "INT402-18181881.pdf".**

Question 1

1. The following figure gives a matrix with six rows.

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

a) Compute the permutations by using the following three hash functions respectively:

h1(x) = (2x + 1) mod 6
h2(x) = (3x + 2) mod 6
h3(x) = (5x + 2) mod 6

[6 points]

b) Compute the Minhash signature for each column if we use the three permutations in **(**a**)**. Indicate which of these permutations is true permutation and explain why?

[8 points]

c) How close are the estimated Jaccard similarities for the six pairs of columns to the true Jaccard similarities?

[6 points]

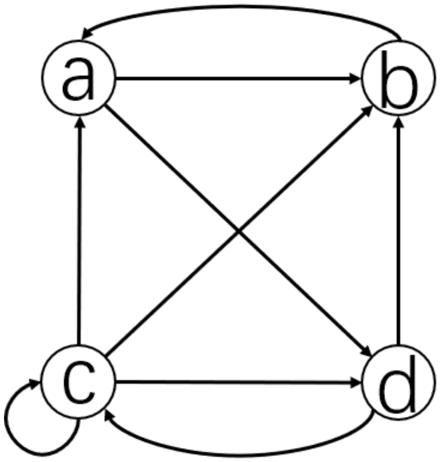
2. Design a map-reduce algorithm to take a very large file of integers and produce as output: The same set of integers, but with each integer appearing only once.

[5 points]

[Total 25 points]

Question 2

1. Compute the PageRank of each vertex in the following figure:



a) Assuming no taxation. Show the results of two iterations at least.

[10 points]

b) Assuming β= 0.9. Show the results of two iterations at least.

[10 points]

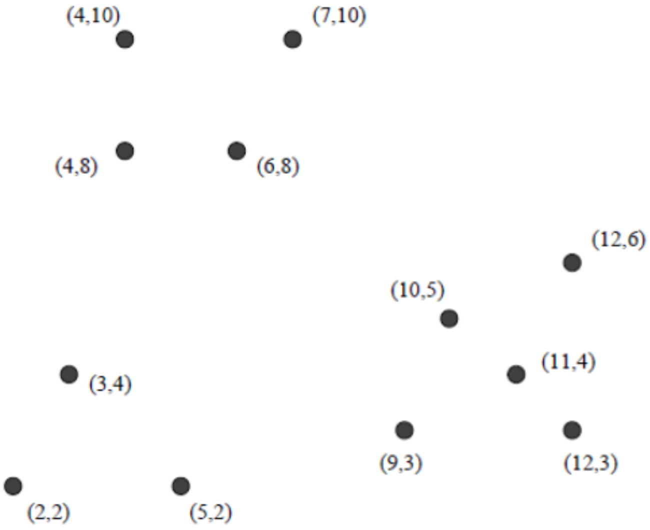
2. Compute the Jaccard similarities of each pair of the following three sets: $\{1, 2, 3, 4\}$, $\{2, 3, 5, 7\}$, and $\{2, 4, 6\}$.

[5 points]

[Total 25 points]

Question 3

1. Suppose that we use the basic hierarchical clustering algorithm to cluster twelve points in a two dimensional Euclidean space (see the following figure).



Each point is named by its (x, y) coordinates.

a) Show the clustering result by plotting the hierarchical tree.

[10 points]

b) How would the clustering of the following figure change if we used for the distance between two clusters by

i. the minimum of the distances between any two points, one from each cluster, and

[10 points]

ii. the average of the distances between pairs of points, one from each of the two clusters.

[10 points]

[Total 30 points]

Question 4

1. Given Ratings of movies by users and SVD for the matrix M shown below: Suppose Leslie assigns rating 3 to Alien and rating 4 to Titanic, giving us a representation of Leslie in "movie space" of [0, 3, 0, 0, 4].

$$\begin{array}{r|ccccc|}
 & \text{Matrix} & \text{Alien} & \text{Star Wars} & \text{Casablanca} & \text{Titanic} \\
\hline
\text{Joe} & 1 & 1 & 1 & 0 & 0 \\
\text{Jim} & 3 & 3 & 3 & 0 & 0 \\
\text{John} & 4 & 4 & 4 & 0 & 0 \\
\text{Jack} & 5 & 5 & 5 & 0 & 0 \\
\text{Jill} & 0 & 0 & 0 & 4 & 4 \\
\text{Jenny} & 0 & 0 & 0 & 5 & 5 \\
\text{Jane} & 0 & 0 & 0 & 2 & 2 \\
\hline
\end{array}$$

$$\underset{M}{\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 0 & 0 & 2 & 2 \end{bmatrix}} = \underset{U}{\begin{bmatrix} .14 & 0 \\ .42 & 0 \\ .56 & 0 \\ .70 & 0 \\ 0 & .60 \\ 0 & .75 \\ 0 & .30 \end{bmatrix}} \underset{\Sigma}{\begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix}} \underset{V^{\mathrm{T}}}{\begin{bmatrix} .58 & .58 & .58 & 0 & 0 \\ 0 & 0 & 0 & .71 & .71 \end{bmatrix}}$$

a) Find the representation of Leslie in concept space.

[4 points]

b) What does that representation predict about how well Leslie would like the other movies appearing in our example data?

[4 points]

2. What are the first ten 3-shingles in the following sentence in the two cases:

[5 points]

i. consider the case of using characters as tokens, and

ii. consider the case of using words as tokens.

> *The most effective way to represent documents as sets, for the purpose of identifying lexically similar documents is to construct from the document the set of short strings that appear within it.*

3. Answer the following questions:

a) What is the difference between following terms: "Data Reduction", "Dimension Reduction", "Feature Selection", "Feature Extraction" and "Sampling"?

[4 points]

b) Explain the role of feature dimensionality in the problem of model overfitting?

[3 points]

[Total 20 points]

END OF RESIT EXAM