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

Module Code	Examiner	Department	Tel
INT402		INT	

1st SEMESTER 2020/21 Remote Open-Book Exam

Postgraduate

Data Mining and Big Data Analytics

Exam Duration: ***2 Hours***

Crash Time Allowed: ***15 Minutes*** *(for online exam)*

INSTRUCTIONS TO CANDIDATES

1、 **This is a blended open-book exam. Please tick the integrity disclaimer *when uploading your answers on LEARNING MALL* and complete the assessment independently and honestly.**

2、 **Total marks available are 100. This accounts for 60% of the final mark.**

3、 **Answer all questions. Relevant and clear steps should be included in the answers.**

4、 **Only English solutions are accepted. Answers need to be handwritten and fully and clearly scanned or photographed for submission as one single PDF file via LEARNING MALL.**

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

6、 **The duration is 2 hours, and an additional 15-minute crash time beyond the exam duration will be allowed for you to report and resolve minor technical issues which may be encountered during the exam. Where there are any major problems preventing you from continuing the exam or submitting your answers in time, please do not hesitate to email the Module Examiner or Assessment Team of Registry (assessment@xjtlu.edu.cn).**

Notes:

- To obtain full marks for each question, relevant and clear steps should be included in the answers.
- Partial marks may be awarded depending on the degree of completeness and clarity.

Question 1

(1) 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]

(2) If we use a triangular matrix A to count pairs of items $\{\boldsymbol{i}, \boldsymbol{j}\}$, and the number of total items $\boldsymbol{n}$ is 17, what pair's count is in $\boldsymbol{A}[\boldsymbol{k}]$, where position $\boldsymbol{k}$ is 60?

[5 points]

(3) Confidence of association rule is the probability of j given $\mathrm{I} = \{i_1, i_2, \ldots, i_k\}$.

Consider the following expression：

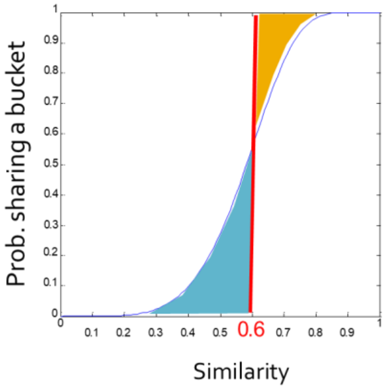
$$\mathrm{conf}(\mathrm{I}\rightarrow\mathrm{j}) = \frac{\mathrm{support}(\mathrm{I} \cup \mathrm{j})}{\mathrm{support}(j)}.$$

Is it true or false? Justify your answer.

[7 points]

(4) Picking r and b to get the best S-curve with 50 hash-functions $(r = 5, b = 10)$. The threshold $s = 0.6$, which is shown as the following figure. What are the upper (yellow) and lower (blue) regions? Why do we want to be more concerned about how to minimize the upper (yellow) region than the lower (blue) region? How to minimize the upper (yellow) region?

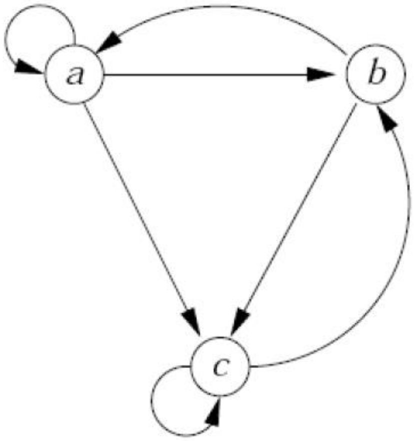
[8 points]



[Total 25 points]

Question 2

Compute the PageRank of each vertex in the following figure:



(1) Assuming no taxation

[10 points]

(2) Assuming β= 0.8

[10 points]

[Total 20 points]

Question 3

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

(1) 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]

(2) Compute the Minhash signature for each column if we use the three permutations in **(1)**. Indicate which of these permutations is true permutation and explain why?

[8 points]

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

[6 points]

[Total 20 points]

Question 4

Suppose there are 120 items, numbered 1 to 120, and also 120 baskets, also numbered 1 to 120. Item i is in basket b if and only if i divides b with no remainder. Thus, item 1 is in all the baskets, item 2 is in all sixty of the even-numbered baskets, and so on. Basket 12 consists of items $\{1, 2, 3, 4, 6, 12\}$, since these are all the integers that divide 12. Answer the following questions:

(1) If the support threshold is 6, which items are frequent?

[3 points]

(2) What is the confidence of the following association rules?

a) $\{3, 4\} \rightarrow 5$

b) $\{2, 3, 6\} \rightarrow 7$

[6 points]

(3) Apply the A-Priori Algorithm with support threshold 6 to compute:

a) the candidate itemsets of size 1 and truly frequent itemsets of size 1.

b) the candidate itemsets of size 6 and truly frequent itemsets of size 6.

[6 points]

[Total 15 points]

Question 5

Three computers, A, B, and C, have the numerical features listed below:

Feature	A	B	C
Processor Speed (GHz)	1.75	2.68	3.56
Disk Size (GB)	750	480	330
Main-Memory Size (GB)	6	4	6

We may imagine these values as defining a vector for each computer; for instance, A's vector is $[1.75, 750, 6]$.

(1) If a certain user has an expectation vector of $U = [2.43, 370, 5]$, calculate the cosine similarity between the user vector and each computer's vector. Explain which computer will be recommended to the user. Briefly justify yourself in one or two sentences.

[7 points]

(2) If we compute the cosine similarity between any two of the vectors, the disk size will dominate and make differences in the other components essentially invisible. Design a data preprocessing method to cope with this problem. Briefly describe the steps of your method.

[6 points]

(3) Calculate the cosine similarity between the user vector and each computer's vector after using your method and explain why your solution is effective.

[7 points]

[Total 20 points]

END OF RESIT EXAM