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 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.

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".

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 [Total 25 points]

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

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

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

i. $h1(x) = (2x + 1) \bmod 7$
ii. $h2(x) = (3x + 2) \bmod 7$
iii. $h3(x) = (4x + 2) \bmod 7$

[6 points]

b) Compute the Minhash signature for each column if we use the three permutations in **(a)**.

[4 points]

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

[6 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 13, what pair's count is in $\boldsymbol{A}[\boldsymbol{k}]$, where position $\boldsymbol{k}$ is 48?

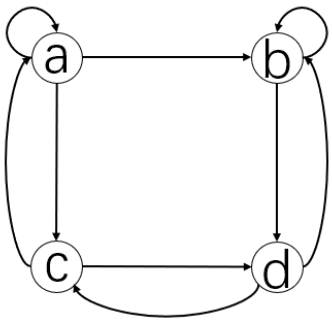
[4 points]

3. We have a matrix which contains $\boldsymbol{n}$ rows. Suppose a column has $\boldsymbol{m}$ **1**'s and therefore $\boldsymbol{n} - \boldsymbol{m}$ **0**'s, and we randomly choose $\boldsymbol{k}$ rows to consider when computing the **minhash**. Prove that the probability of getting "**don't know**"as the minhash value for this column is at most $\left(\frac{\boldsymbol{n}-\boldsymbol{k}}{\boldsymbol{n}}\right)^{\boldsymbol{m}}$.

[5 points]

Question 2 [Total 25 points]

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

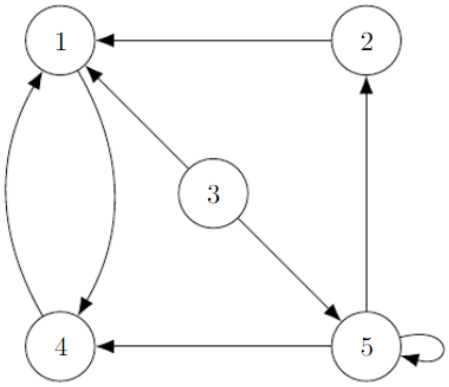


Assuming β= 0.8. Show the results of two iterations at least.

[18 points]

2. Given the graph blow and the topic-specific PageRank vector

$\boldsymbol{r} = [\mathbf{0.3662}, \mathbf{0.0442}, \mathbf{0.0800}, \mathbf{0.3519}, \mathbf{0.1578}]$ using the teleport set $\boldsymbol{S} = \{\mathbf{3}, \mathbf{5}\}$.



What is the value of $\boldsymbol{\beta}$ (corresponding to the teleport probability $\mathbf{1} - \boldsymbol{\beta}$)? Justify your answer.

[7 points]

Question 3 [Total 25 points]

1. Given five 3-dimensional data points shown below,

$P1$	$(3, 1, 2)$
$P2$	$(0, 2, 1)$
$P3$	$(3, 0, 5)$
$P4$	$(1, 1, 1)$
$P5$	$(4, 2, 2)$

Apply K-means clustering method to group them into 2 clusters, using L_1 distance measure.

Suppose that the initial centroids are $C1$: $(1, 0, 0)$ and $C2$: $(3, 0, 0)$. Use the following table **as a template** to show **each step** of clustering clearly. Explain why the final clustering has been achieved (i.e., discuss the stop condition of K-

means).

Cluster	Old Centroids	Cluster Elements	New Centroids
1	$(1, 0, 0)$		
2	$(3, 0, 0)$		

…

…

[7 points]

(Hint: L_1 distance (Manhattan distance) is a distance metric between two points in a N dimensional vector space. It is the sum of the lengths of the projections of the line segment between the points onto the coordinate axes. In simple terms, it is the sum of absolute difference between the measures in all dimensions of two points.)

2. Briefly discuss the major difference between Classification and Clustering. List one real application for each of them respectively.

[6 points]

3. The following figure is a utility matrix, representing the ratings, on a 1-5 star scale, of eight items, through a to h, by three users A, B, and C.
Compute the following from the data of this matrix.

	a	b	c	d	e	f	g	h
A	4	5		5	1		3	2
B		3	4	3	1	2	1	
C	2		1	3		4	5	3

a) Testing the utility matrix as Boolean, compute the Jaccard distance between each pair of users.

[3 points]

b) Compute the Cosine distance between each pair of users.

[4 points]

c) Do the results of (a) and (b) reflect the true similarity of the users? Please elaborate on your reasons. If the answer is no, please give a solution.

[5 points]

Question 4 [Total 25 points]

1. Given a query of "**transfer learning for video tagging**" and a collection of the following three documents:

Document 1	<A survey on transfer learning>
Document 2	<Transfer learning for image tagging>
Document 3	<Transfer learning: from image tagging to video tagging>

Use the Vector Space Model, **TF/IDF weighting scheme**, and **Cosine vector similarity** measure to find the most relevant document(s) to the query. Assume that "a", "on", "for", "from" and "to" are stop words.

The formula of TF/IDF Weighting is: $w_{ij} = t_{ij} \times \log\left(\frac{N}{n_j}\right)$

where:

t_{ij}***: the number of times term j appeared in document i.***

N***: the Total number of document.***

n_j***: the number of documents that term j appears in***

a) If the support threshold is **6 ,** which items are frequent? Calculate "document frequency" and "inverse document frequency" for each word.

Word list	Document Frequency	Inverse Document Frequency

[6 points]

(Hint: $log\ 2 = 0.301,\ log\ 3 = 0.477$***)***

b) Represent each document as a weighted vector by using TF/IDF weight scheme. Length normalization is not required.

[6 points]

c) Represent the query as a weighted vector and find its most relevant document(s) using **Cosine Similarity** measure.

[8 points]

2. Given a large dataset with hundreds of attributes, which is too large to keep in memory and too big to scan over on a single machine. Design an algorithm to build a decision tree.

[5 points]

[Total 100 points]

END OF FINAL EXAM