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

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

1st SEMESTER 22-23 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. 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) What are Hash Functions? What are the hash collisions of Hash Functions? How to evaluate if Hash Functions are good?

[6 points]

(2) Confidence of association rule is the probability of j given $I = \{i_1, i_2, ..., i_k\}$. Consider the following expression:

$$\text{conf}(\text{I} \rightarrow \text{j}) = \frac{\text{support}(\text{I} \cup \text{j})}{\text{support}(j)} \tag{1}$$

Is it true or false? Justify your reason.

[6 points]

(3) Given a support threshold s, then sets of items that appear in at least s baskets are called frequent itemsets. How do you choose a support threshold s?

[6 points]

(4) What is the monotonicity of itemsets? Describe the PCY algorithm.

[7 points]

Question 2

[25 points in total]

The following figure gives a matrix with seven rows.

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

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

$$\text{i. } h_1(x) = (5x + 1) \bmod 7$$
$$\text{ii. } h_2(x) = (2x + 3) \bmod 7$$
$$\text{iii. } h_3(x) = (10x + 6) \bmod 7$$

[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]

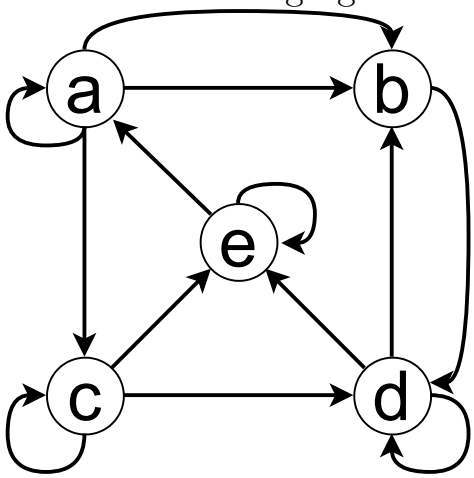
(4) Suppose we use 200 hash functions to generate 200 rows in total for the signature matrix. For locality-sensitive hashing, if we split the signature matrix into 20 bands, say if columns S_4 and S_5 have a similarity of 30%, what is the probability that S_4 and S_5 are hashed to at least 1 common bucket? If we set the similarity threshold $s = 0.9$, what is the probability that S_4 and S_5 being similar pairs are false positives?

[5 points]

Question 3

[24 points in total]

Given a web graph shown in the following figure:



(1) Calculate the transition matrix of this graph.

[10 points]

(2) Assume the teleport set is {e} only with $\beta = 0.9$. Compute the topic-sensitive PageRank using three iterations.

[7 points]

(3) Using Google Matrix with $\beta = 0.8$. Compute the topic-sensitive PageRank using three iterations.

[7 points]

Question 4

[26 points in total]

(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 $[0, 3, 0, 0, 4]$.

	Matrix	Alien	Star Wars	Casablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 0 & 0 & 2 & 2 \end{bmatrix} = \begin{bmatrix} .14 & 0 \\ .42 & 0 \\ .56 & 0 \\ .70 & 0 \\ 0 & .60 \\ 0 & .75 \\ 0 & .30 \end{bmatrix} \begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix} \begin{bmatrix} .58 & .58 & .58 & 0 & 0 \\ 0 & 0 & 0 & .71 & .71 \end{bmatrix}$$

$M \qquad U \qquad \Sigma \qquad V^{\mathrm{T}}$

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) Given a sentence showing below:

> **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.**

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

a) consider the case of using characters as tokens (space character is considered, use _ to express the space character), and

[5 points]

b) consider the case of using words as tokens.

[5 points]

(3) Answer the following questions.

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

[6 points]

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

[2 points]

THE END OF EXAM