Xi'an Jiaotong-Liverpool University

西交利物浦大學

PAPER CODE	EXAMINER	DEPARTMENT	TEL
CSE313	[redacted]	Computer Science and Software Engineering	[redacted]

1st SEMESTER 2019/20 FINAL EXAMINATION

Undergraduate – Year 4

Big Data Analytics

TIME ALLOWED: 2 Hours

INSTRUCTIONS TO CANDIDATES

1. This is a closed-book examination, which is to be written without books or notes.
2. Total marks available are 100. Marks for this examination account for 70% of the total credit.
3. The numbers on the right indicate the marks available.
4. Answer ALL questions in all sections.
5. Answers should be written in the answer booklet(s) provided and clearly mark question numbers and write "Answer:".
6. The university approved calculator – CASIO FS82ES/83ES can be used.
7. Only answers in English are accepted.

Part A. Basic Textbook Based Questions (40 marks)

A1. Explain the term "Big Data" and the reason why big data is a relative term? [5]

A2. One of the Big Data type is ordered data. Provide a definition of ordered data and give three different types ordered data examples. [5]

A3. Discuss the statement that "BDA is NOT a simple integration of the fastest processor and the largest memory together to solve big data analyses problem." Provide your opinion of either for or against with reasons. [5]

A4. [Distributed Computing] What does it mean when a system is loosely coupled? Give one example of loosely coupled systems that you have experience with. [5]

A5. [Apache Hadoop] What is the general difference between Hadoop and RDBMS from BDA point of view? [5]

A6. [MapReduce] Explain at a high-level how MapReduce works. [5]

A7. [Apache Spark] How Does a Spark Application Work? [5]

A8. [MongoDB] MongoDB does not support *Join* as in RDBMS. Explain from data modelling point of view how RDBMS' join can be transformed from the logical data model (LDM) to a physical data model (PDM)? [5]

Part B. Data Analysis Methods and Algorithms (60 marks)

B9. [Association Analysis] Consider the market basket transactions shown in Table 1. **[20]**

Table 1. Market Basket Transactions.

Transactions ID	Items Bought
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

(a) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?

(b) What is the maximum size of frequent itemsets that can be extracted (assuming *minsup* > 0)?

(c) Write an expression for the maximum number of size-3 itemsets that can be derived from this data set.

(d) Find an itemset (of size 2 or larger) that has the largest support.

(e) Find a pair of items, a and b, such that the rules {a} → {b} and {b} → {a} have the same confidence.

B10. [Classification] Consider the training examples shown in Table 2 for a binary classification problem. [20]

Table 2. Binary Classification.

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

(a) What is the entropy of this collection of training examples with respect to the positive class?

(b) What are the information gains of a1 and a2 relative to these training examples?

(c) What is the best split (between a1 and a2) according to the information gain?

(d) What is the best split (between a1 and a2) according to the classification error rate?

(e) What is the best split (between a1 and a2) according to the Gini index?

B11. [Cluster] Use the similarity matrix in Table 3 to perform complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged. [20]

Table 3. Similarity matrix for C11.

	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

END OF PAPER

This page is empty.