Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

AUGUST/RESIT EXAMINATIONS 2018/2019

MODULE: CA4010 - Data Warehousing & Data Mining

PROGRAMME(S):

CASE	BSc in Computer Applications (Sft.Eng.)
ECSAO	Study Abroad (Engineering & Computing)
ECSA	Study Abroad (Engineering & Computing)

YEAR OF STUDY: 4,O,X

EXAMINER(S):

Mark Roantree	(Internal)	(Ext:5636)
Dr. Hitesh Tewari	(External)	External

TIME ALLOWED: 3 Hours

INSTRUCTIONS: **Answer 4 questions. All questions carry equal marks.**

PLEASE DO NOT TURN OVER THIS PAGE UNTIL YOU ARE INSTRUCTED TO DO SO.
The use of programmable or text storing calculators is expressly forbidden.
Please note that where a candidate answers more than the required number of questions, the examiner will mark all questions attempted and then select the highest scoring ones.

There are no additional requirements for this paper.

QUESTION 1 **[TOTAL MARKS: 25]**
(Cube Computation)
This question uses **figure 1** in **Appendix 1**.

Q 1(a) **[6 Marks]**
A data scientist working at Dublin airport has compiled a dataset with the data captured when boarding cards are scanned. This includes airline, destination and date. Your lowest level of granularity is flight and thus, your measure is the number of passengers per flight. Your goal is to analyse passenger numbers by airline, destination and season (dates).

Draw the lattice for the data Cube. Also, separately list all of the cuboids from your lattice.

Q 1(b) **[6 Marks]**
What is meant by an *ancestor cell*? As part of your answer, describe a *descendant cell* in relation to an ancestor cell.
Provide examples using the data in **figure 1**, which clearly illustrate the difference between both types of cells.

Q 1(c) **[13 Marks]**
Assume a subset of the dataset looks at 3 destinations in France: Toulouse, Perpignan and Lyon. We have 3 airlines: Aer Lingus (EI), Air France (AF) and Ryanair (FR). The dates for the study are (May,Jun,Jul,Aug). All 3 airlines can fly to all 3 destinations.
In total, we have 200 flights in the sample with a partially complete dataset shown in **figure 1**, containing all *1-D* cuboids.

i. Write the dimensional *values* for all 3-D cuboids (Hint: there will be 36 in total). Your answer should *not* include any measures.

ii. Provide sample measures for all 3-D cuboids for Toulouse. Hint: You must ensure that the aggregates are valid.

[End of Question1]

QUESTION 2 ***[TOTAL MARKS: 25]***

(Association Rule Mining)
This question does **not** use the Appendix.

Q 2(a) **[6 Marks]**

Table 1 shows 4 transactions, each with a set of items in a shopping basket. Assume that minimum support, **minsup** = 50% and minimum confidence, **minconf** = 60%.

T001	A,C,H
T004	A,B,E,F,H
T005	A,B,C,D
T008	A,B,C,E

Table 1. Shopping Basket Transactions

List all frequent itemsets together with their support.

Q 2(b) **[9 Marks]**

i. List those itemsets from part a) that are **closed**.
ii. List those itemsets that are **maximal**.
iii. For all frequent itemsets of maximal length, list all corresponding association rules (ie. including subsets) satisfying the requirements for *minimum support* and *minimum confidence* together with their confidence. In other words, list **each rule** and **confidence measure**.

Q 2(c) **[10 Marks]**

Compute lift for *every* association rule from Q2(b) part iii.

[End of Question2]

QUESTION 3 **[TOTAL MARKS: 25]**

(Data Warehousing)
This question does **not** use the Appendix.

Q 3(a) **[8 Marks]**

Discuss the differences between traditional databases and OLAP-oriented data warehouse systems under the following headings:

i. orientation
ii. database design
iii. unit of work
iv. summarisation

In each case, use a real world example of both types of data management systems.

Q 3(b) **[7 Marks]**

Provide an illustration of an ETL architecture. Ensure you include the **description** and **goal** for each layer (or component), and describe what takes place.

Q 3(c) **[10 Marks]**

i. Explain the difference between an *independent* data mart and a *dependent* data mart? Give an example of each from the real world.

ii. Explain the design concept that is used to manage and control the development of multiple data marts by (possibly separate) teams.

iii. Draw a sample Bus Architecture. Use the organisation/company of your choice and list 4 requirements and show how they appear in the bus architecture and how their overlap is represented.

[End of Question3]

QUESTION 4 ***[TOTAL MARKS: 25]***

(Clustering)
This question uses **figure 2** in **Appendix 2**.

Q 4(a) **[15 Marks]**

Using an Agglomerative Hierarchical Clustering approach, cluster the 6 points (A,B,C,D,E,F) in **figure 2**. At each step, show the current state of the graph and the new similarity matrix.

Q 4(b) [10 Marks]

You are testing a speed camera which records the speed of cars outside the university. Your job is to detect outlier values recorded by the device. The machine is considered to have 95% accuracy in its readings.

i. In terms of normal distribution, describe what working hypothesis you use. Briefly discuss an appropriate discordancy test and significance probability function.

ii. Describe the different approaches taken in block procedures and sequential procedures.
Use examples from the speed dataset in giving your answer.

In general, which approach is better?

[End of Question4]

QUESTION 5 [TOTAL MARKS: 25]

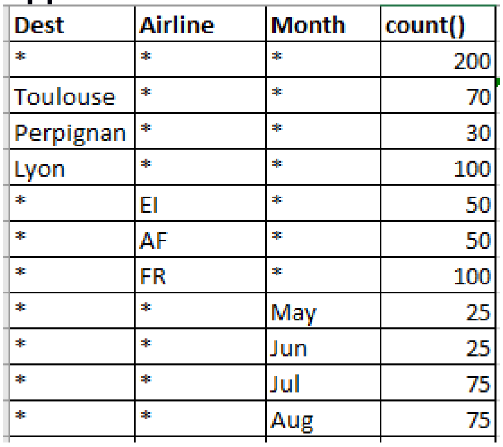
(Classification)
This question uses **figure 3** in **Appendix 3**.

Q 5(a) [7 Marks]
Write the equation *Info(D)* for the initial Information Gain for a dataset D. Explain what the equation does.

Q 5(b) [18 Marks]
Using Information Gain, rank the four attributes in figure 3 (age, specRx, astig, tears) for selection for the *first split only*, from best to worst. In other words, you are required to provide the initial classification only.

[End of Question5]

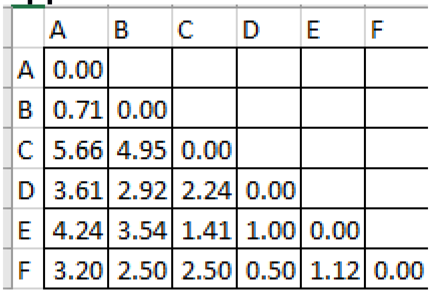
Appendix 1



Dest	Airline	Month	count()
*	*	*	200
Toulouse	*	*	70
Perpignan	*	*	30
Lyon	*	*	100
*	EI	*	50
*	AF	*	50
*	FR	*	100
*	*	May	25
*	*	Jun	25
*	*	Jul	75
*	*	Aug	75

Figure 1: Boarding Card Dataset

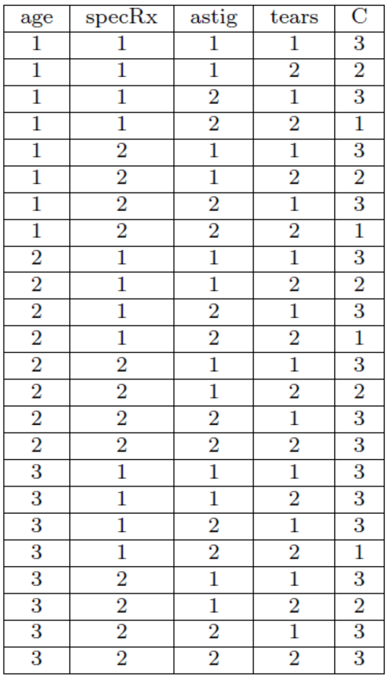
Appendix 2



	A	B	C	D	E	F
A	0.00					
B	0.71	0.00				
C	5.66	4.95	0.00			
D	3.61	2.92	2.24	0.00		
E	4.24	3.54	1.41	1.00	0.00	
F	3.20	2.50	2.50	0.50	1.12	0.00

Figure 2: Dissimilarity matrix for (A,B,C,D,E,F)

Appendix 3



age	specRx	astig	tears	C
1	1	1	1	3
1	1	1	2	2
1	1	2	1	3
1	1	2	2	1
1	2	1	1	3
1	2	1	2	2
1	2	2	1	3
1	2	2	2	1
2	1	1	1	3
2	1	1	2	2
2	1	2	1	3
2	1	2	2	1
2	2	1	1	3
2	2	1	2	2
2	2	2	1	3
2	2	2	2	3
3	1	1	1	3
3	1	1	2	3
3	1	2	1	3
3	1	2	2	1
3	2	1	1	3
3	2	1	2	2
3	2	2	1	3
3	2	2	2	3

Figure 3. Lens Dataset ***[END OF EXAM]***