Student Number _____________ Last Name _______________ First Name ________________

3804ICT/3031ICT/7031ICT Quiz 1 – Answer Short Questions

Contents: Introduction to Data Mining, Data Exploration and Data Pre-processing, and Data warehouse and OLAP.

Instructions: Please list all the detailed calculation steps which help you to get the final solutions.

Question Set 1. Introduction to Data Mining (10 points)

Please answer the following questions:

a) What is data mining? (2 points)

b) What are functions of data mining and what are the typical applications for each task? (4 points)

c) Please briefly describe major issues in data mining. (2 points)

d) What is your expectation from this data mining course? (2 points)

Question Set 2. Data Exploration (12 points)

age	Gender	postcode	weight	admission date
23	M	4222	55 kg	01/08/2018
45	M	4232	60 kg	29/07/2018
21	F	4201	45 kg	26/06/2018
67	M	4309	85 kg	02/04/2018

a) Give the patient records, please categorize the types of the following attributes: age, gender, postcode, weight, and admission date. (2 points)

b) Which are discrete and continuous attributes in the given data? (2 points)

c) What do basic statistical descriptions include? (2 point)

d) What is the five-number summary of the given body weight data: (41,41,42,43,45,46,49,50,55,58,61,66,73,79,80,85,87,89,92,98) (3 point).

e) If the body weight data can be divided into three groups: a). ≤55, b). [56,80] c). ≥81, what is the approximate median value? Please calculate the frequency of each interval and use equation $median = L_1 + \left(\frac{\frac{N}{2}-(\sum freq)_l}{freq_{median}}\right) width$ to compute the approximation of median value. (3 points)

Question Set 3. Data Pre-processing (8 points)

Please answer the following questions:

a) Data quality can be assessed in terms of accuracy and completeness. Propose three other dimensions of data quality. (2 points)

b) Please clarify the differences between “incomplete data” and “noisy data”. What are the corresponding dealing strategies? (3 points)

c) Please calculate the cosine similarity between two documents represented by two bag-of-word vectors: d_1=(4,1,5,6,8,2,4,6,1) and d_2=(0,1,3,3,9,0,4,2,5). (3 points)

Question Set 4. Data Pre-processing (12 points)

Please answer the following questions:

a) What is “data reduction” and what are its strategies, giving examples? (2 points)

b) Use the Min-max normalization and z-score normalization to normalize the following group of data: (200,300,400,600,1000). (4 points)

c) Suppose a group of 20 weight records (in Kg) has been sorted as follows: (41,41,42,43,45,46,49,50,55,58,61,66,73,79,80,85,87,89,92,98). Partition them into four bins by a). equal-depth partitioning method and b). equal-width partitioning method. (6 points)

Question Set 5. Data Warehouse (OLAP operations) (8 points)

Please answer the following questions:

a) What are four characteristics of Data Warehouse? (2 points)

b) What are the main components in the bottom tier of a multi-tiered architecture? (1 point)

c) Briefly compare star schema, snowflake schema, and fact constellation schema. (2 points)

d) How many OLAP operations? (2 points)

e) If a data cube has 5 dimensions, each of which has 3 levels, what is the total number of cuboids can be generated? (1 point)