UNIVERSITY OF MALAYA

EXAMINATION FOR THE DEGREE OF MASTER OF DATA SCIENCE

ACADEMIC SESSION 2019/2020 : SEMESTER II

WQD7005 : Data Mining

June 2020

INSTRUCTIONS TO CANDIDATES :

Answer **ALL** questions (50 marks).

(This question paper consists of 5 questions on 3 printed pages)

Mini-assignment (50 marks)

Instructions: Work individually, submission via Spectrum.

. 1. You are required to make a user-agent that will crawl the WWW (your familiar domain) to produce dataset of a particular website.

* the web site can be as simple as a list of webpages and what other pages they

link to

* the output does not need to be in XHTML (or HTML) form

a multi-stage approach (e.g. produce the xhtml or html in csv format )

(10 marks)

1. Draw snowflake schema diagram for the above dataset. Justify your attributes to be selected in the respective dimensions.

(10 marks)

1. You are required to write code to create a decision tree (DT) model using the above dataset (Question 1). In order to achieve the task, you are going to cover the following steps:

* Importing required libraries
* Loading Data
* Feature Selection
* Splitting Data
* Building Decision Tree Model
* Evaluating Model
* Visualizing Decision Trees

(10 marks)

1. You are required to write code to find frequent itemsets using the above dataset (Question 1). In order to achieve the task, you are going to cover the following steps:

* Importing required libraries
* Creating a list from dataset (Question 1)
* Convert list to dataframe with boolean values
* Find frequently occurring itemsets using Apriori Algorithm
* Find frequently occurring itemsets using F-P Growth
* Mine the Association Rules

(10 marks)

1. You are required to write code to implement either time-series clustering or density-based clustering model using the above dataset (Question 1). If you select density-based clustering approach to achieve the task, you are going to cover the following steps:

* Importing required libraries
* Load the dataset (Question 1) into a DataFrame object
* Visualize the data, use only two of these attributes at the time
* You may need to normalise the attribute if necessary
* Show positive correlation between attributes if necessary
* Construct a density-based clustering model and extract cluster labels and outliers to plot your results.

(10 marks)

Submissions:

The student is expected to submit answers to each question individually, and submit the document in PDF format. The student can include online materials, screenshots, videos and/or codes (ipynb format) to support your answer