University of Dhaka
Institute of Information Technology
Bachelor of Science in Software Engineering
Final Examination, 2020
CSE 504 : Database Management System-II
Marks: 30 Time: 1 Hour 15 Mins

Professionalism	Excellence	Respect

Answer all the questions. The weight of each question is mentioned at the right side. When answering a question, please answer all the subsections of it at once

1. Consider the following table named *person* which contains the age and weight of individuals.

person ID	**age (years)**	**weight (kg)**
A	25	60
B	35	82
C	53	70
D	47	77
E	25	60

(a) Given the above two columns (age and weight), construct the *K-D tree* representation of this data. **6**

(b) Explain briefly how you can use this *K-D tree* to perform the following range query. **4**

*Select * from person where age > 35 and weight < 80*

2. **(a)** Consider the following database about word occurrences in Webpages: **5**

Webpage(url, author)
Occurs(url, wid)
Word(wid, text, language)

Where, Webpage.url and Word.wid are keys.
Occurs.url and Occurs.wid are foreign keys to Webpage and Word respectively.

Assume the following statistics

T(Webpage) = V (Occurs; url) = 10^9
T(Occurs) = 10^{12}
T(Word) = V (Occurs; wid) = 10^6
V (Webpage; author) = 10^7
V (Word; language) = 100

Assume ten records can be fit in one block, hence B(Webage) = T(Webpage)=10 and similarly for all other tables.

(σindex-lookup author='John'(Webpage) ⋈index-join url=url Occurs) ⋈ main-memory-hash-join wid=wid σindex-lookup language='French'(Word)

Compute the cost of the plan for the following case:

Webpage.url = primary index
Webpage.author = secondary index
Occurs.url = secondary index
Occurs.wid = primary index
Word.wid = primary index
Word.language = secondary index

(b) Consider two relations R(A, B, C, D) and S (D, E) with the following statistics: **5**
T (R) = 100, V (R, A) = 100, V (R, B) = 10, V (R, C) = 1, V (R, D) = 50; T (S) = 500, V (S, D) = 30, V (S, E) = 100.
(i) Estimate the number of tuples in $\sigma_{B=25}(R)$
(ii) Estimate the number of tuples in $\sigma_{B=25 \text{ AND } (C=30)}(R)$
(iii) Estimate the number of tuples in $\sigma_{B>25}(R)$
(iv) Estimate the number of tuples in $\sigma_{B>25 \text{ AND } (B=15)}(R)$
(v) Estimate the number of tuples in R |X| S

3. Suppose you are given the following data:

yoochoose-clicks.dat - Click events. Each record/line in the file has the following fields:
Session ID – the id of the session. In one session there are one or many clicks.
Timestamp – the time when the click occurred.
Item ID – the unique identifier of the item.
Category – the category of the item.

yoochoose-buys.dat - Buy events. Each record/line in the file has the following fields:
Session ID - the id of the session. In one session there are one or many buying events.
Timestamp - the time when the buy occurred.
Item ID – the unique identifier of item.
Price – the price of the item.
Quantity – how many of this item were bought.

The test data also contain the same information as click data and you have to answer whether something will be buy for a particular session. Now answer the followings:

(a) How one can calculate P(buy|no_of_click)? **2**

(b) Extract three important features from the aforementioned data that are suitable to design a Bayesian classifier. Justify your answer. **3**

(c) Design a Bayesian classifier to answer whether a session in the test data is related with buy. Justify your answer. **5**