MTech in Data Science
IIT Palakkad
DS5003A : Data Engineering

0930-1230 **Final Exam (26 Nov 2024)** **Marks : 20**

Instructions

1. Write your answers neatly in Blue/ Black ink. Do not use pencil / Red ink. If your answer is not legible, you will not get any marks for that.

2. Doubts and questions will not be answered during the exam. If you have to make any assumption about unspecified things, write the assumption clearly with justification.

3. Answer all parts of a question together. If the parts of a single question are not together, then only the first part will be evaluated. Other parts will not get any marks.

1. There is a collection of documents where each document is a collection of words. Answer the following: (2+2+2)

   a. Write an algorithm to compute co-occurrence count for each pair of words. Where co-occurrence means appearing in a document together.

   b. Extend the above algorithm to compute co-occurrence count for k words.

   c. Use the above algorithm to cluster words such that each cluster will contain k number of co-occurring words if the co-occurrence of k words is larger than a predefined threshold. One word can appear in multiple clusters.

2. There is a collection of images. Each image is represented by a matrix of pixels, where each pixel is a vector of length 3. Answer the following: (2+4)

   a. Write an algorithm to compute distance between any pair of images. Justify the distance measure so that two similar images should have less distance than two dissimilar images.

   b. Use the above algorithm to create an algorithm to group the images into k groups, where images within a group are more similar than images across groups. You should use only one iteration and k need not be predefined.

3. Consider the following schema of banking management system. (4+4)

branch(branch_name, branch_city, assets)
customer(customer_name, customer_street, customer_city)
loan(loan_number, branch_name, amount)
borrower(customer_name, loan_number)
account(account_number, branch_name, balance)
depositor(customer_name, account_number)
executive(employee_name, branch_name, customer_name)

Executive table contains information about the employee of the bank (called executive) and customers who are managed by the executive. Every executive has a sales-score which is the total amount of loans borrowed by customers managed by the executive. Answer the following without creating any additional views, functions, procedures etc.

a. Write a trigger which will check the balance if a customer applies for a loan, and will discard the loan if the loan amount is more than 5 times the balance.

b. Each branch has one leader who has the highest sales-score among all executives in that branch. Create a view with the following structure:
manager(leader_name,employee_name)