CS4.301: Data and Applications **Due: 11:55PM, 7th October 2020**

Assignment #5

Instructor: Kamal Karlapalem *Released:* 30th September 2020

In this assignment, you have to type SQL queries that satisfy the questions posed below. Try not to overthink things. You are allowed to use intermediate views for these questions. You do not need to submit queries for the creation and population of tables unless the question asks for it.

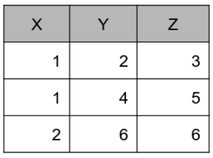
Problem: Point Clouds

You are a researcher working on 3D computer vision problems. You've been given a dataset comprising of 3D point cloud data - (x, y, z) coordinates of the world. Solve the questions below using SQL statements to better understand your data.

The dataset is defined as:

A table POINT with attributes X, Y, Z where each row is a single point (x, y, z).

You can assume that height is in the z axis. Here's a sample dataset to make things clearer:



X	Y	Z
1	2	3
1	4	5
2	6	6

Figure 1: Sample data for POINTS

1 Questions

1. Display all the points in increasing order of x, then y, then z.
2. Find the highest and lowest points as per their elevation.
3. Find the k nearest points to the origin, using euclidian distance.
4. Find the mean of the point cloud.
5. Create a 3x3 rotation matrix in a table ROTATE that rotates anticlockwise about the **Y** axis with variable angle α in radians as input.

   In two dimensions, the rotation would look like this,

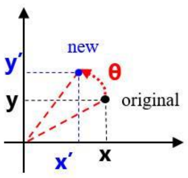


Figure 2: Two dimensional rotation

Updated coordinates would be,

$$x^{'} = x\cos\theta - y\sin\theta$$
$$y^{'} = x\sin\theta + y\cos\theta$$

And hence the rotation using a matrix would be,

$$R\mathbf{v} = \begin{bmatrix}\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta\end{bmatrix} \cdot \begin{bmatrix}x \\ y\end{bmatrix} = \begin{bmatrix}x^{'} \\ y^{'}\end{bmatrix}$$

where R is the rotation matrix $\mathbf{v}$ represents the initial point and $R\mathbf{v}$ represents the rotated point.

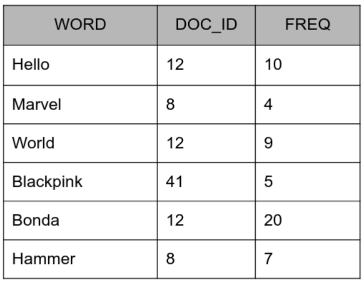
6. Use the ROTATE table to rotate all the points in POINTS anticlockwise by alpha! (Hint: You can write three separate queries for updating the X, Y, Z coordinates of the points in POINTS. However, do not write queries for each point)

Problem: Information Extraction

While building search engines and similar applications, it is important to have an index which tells you the frequency of words in documents. Each record in the table given below tells us that a word X occurred in the document Y, n times.

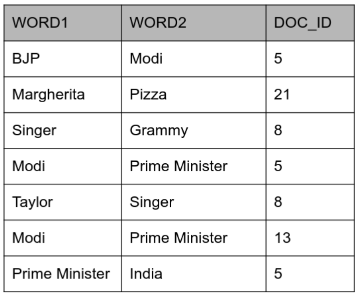
The dataset is defined as follows:

1. A table INVINDEX with attributes 'WORD', 'DOC_ID', 'FREQ' where each row represents that the word W occurred in document D with frequency N.
2. A table ENTITY with attributes 'WORD1', 'WORD2' and 'DOC_ID' where each row represents that there's a link between Word1 W1 and Word2 W2 obtained from document D.



WORD	DOC_ID	FREQ
Hello	12	10
Marvel	8	4
World	12	9
Blackpink	41	5
Bonda	12	20
Hammer	8	7

(a) Sample data for INVINDEX



WORD1	WORD2	DOC_ID
BJP	Modi	5
Margherita	Pizza	21
Singer	Grammy	8
Modi	Prime Minister	5
Taylor	Singer	8
Modi	Prime Minister	13
Prime Minister	India	5

(b) Sample data for ENTITY

Answer the following questions with a similar dataset.

2 Questions

1. For the phrase "Hello World", print all the document IDs and the corresponding scores for the documents (only the non-zero scores). The score of a word W in a document D is:

$$score = \frac{\text{freq of W in D}}{\text{total number of words in D}} \tag{2.1}$$

   For any phrase P, the score of a document D is therefore the *"sum of scores of all words in P"*.

2. For a given entity “BJP” print all entities which occur at a distance 3 from “BJP” and all the 3 links are retrieved from the same document. Also print the corresponding document ids. Here, the links between different words is analogous to a graph, where a directed edge exists between the different pairs of words in the ENTITY relation. A node Y is at a distance k from a node X if there exist k edges linking node X to node Y. That is, Y is at a k-hop distance from X.

   For example, the sample output for the above query would be :
   India 5

3. For the documents retrieved in question 2, print the frequency of the word ”Lotus”.

4. The average score of a word is given by the sum of scores of the word across all documents (where it appears) divided by the number of these documents. Display the words in decreasing order of average score.

5. Display ”This is the last assignment!”

3 Bonus

Try solving the questions using a single SQL query each instead of multiple sub queries.

There **will not** be any extra marks for the Bonus. Do not fret if you aren’t able to solve them immediately. This is just for you to explore.

4 Submission Instructions

Submit a single .txt file named **<teamname>.txt** (without the < and >) containing the required SQL commands. Each SQL command must begin on a new line.

—

All the Best!