HOMEWORK 2

Note: Please strictly adhere to the submission instructions given at the end of the document

Due date: 9th October, 11:59PM

1. This question is about getting yourself familiar with frequent subgraph mining tools. Run it on the Yeast (https://bit.ly/3EVA5Cc) Dataset. This is a database of molecules. The format is the following:

   ```
   #graphID

   # of nodes

   Series of Node Labels

   # of edges

   Series of "Source node, Destination Node, Edge label"
   ```

   Run gSpan, FSG (also known as PAFI), and Gaston (you should be able to find it online) against frequency threshold in the dataset (you may need to write a script to change the format of the dataset) at minSup = **5%, 10%, 25%, 50%** and **95%**. Plot the running times and explain the trend observed in the running times. Specifically comment on the growth rates and why one technique is faster than the others. You are free to consult the respective papers. **(20 points)**

   Libraries you can potentially use:

   - gSpan : https://sites.cs.ucsb.edu/~xyan/software/gSpan.htm
   - FSG : http://glaros.dtc.umn.edu/gkhome/pafi/download
   - Gaston : https://liacs.leidenuniv.nl/~nijssensgr/gaston/download.html

2. Dataset (https://bit.ly/3odTt7B) . [Marks: 10 points for correct algorithm. 30 points for correct implementation. 10 points for correct explanation.]
   a. Reduce to dimensions 2,4,10,20 using PCA.
   b. Index using KD-tree and M-tree for L2 distance. You can use off-the shelf libraries if you wish to. Maintain the dataset in memory.
   c. Write an algorithm to perform k-NN query on M-tree and KD-tree. Needless to say your algorithm should try to use the index structure and maximize the pruning potential of the search space.
   d. Choose 100 random points as query and plot the average running time of 5-NN query with standard deviation as error bars for KD-tree, M-tree and Sequential scan against dimension. Explain the trends.

Instructions:

- You need to do the homework in your already formed team of 3.

- Only one submission per team will be accepted. Make sure no duplicate submissions are made by another team member.
- Upload all code to GitHub.
- Your GitHub repo must contain all scripts and code you used in this assignment (including preprocessing scripts and code, if any).
- Your submission should be as following:

1. Submit an install.sh file on moodle such that running '**sh install.sh'** should clone your repository and load required modules.
    - After unzipping, we should get a folder of the name 'HW2_kerberosid'. Eg: If submitter's entry number is 2019CSZ8763, your folder name will be "HW2_csz198763".
    - Inside the main folder, there should be two sub-folders titled Q1 and Q2 corresponding to each question. Include all code/scripts for the corresponding questions in these sub-folders. The main folder should also contain two files Q1.sh, Q2.sh. Further it should contain a writeup file and a readme file (refer to the last 3 points for details about the writeup and readme file).
    - Running **'sh Q1.sh <data> <plot_name>'** should generate run all the preprocessing steps, and scripts/code corresponding to Q1 and generate a plot with the name <plot_name> with .png extension. Note that use relative paths (we will provide path and name for saving plot) and do not hardcode the name of the plot (it should be same as <plot_name> given in input).
    - Similarly, running **'sh Q2.sh <data> <plot_name>'** should generate run all the scripts/code corresponding to Q2 and generate a plot with the name <plot_name> with .png extension. Note that use relative paths (we will provide path for saving plot) and do not hardcode the name of the plot (it should be same as <plot_name> given in input).
    - **Please do not submit data files. We will give the relative path of data in place of <data> argument in above instructions.**
    - Include a **single writeup/report** containing the explanation and plots for the questions given above. Writeup file name must be "**report_kerberosid.pdf**" Also mention any assumptions you have made specific to each question.
    - Include a **readme file.** Filename should be **readme_kerborosid.txt**. The file should give all instructions to run your code on hpc. However, instructions for running Q1.sh and Q2.sh should be as specified above. Also include steps on how to run the preprocessing code/scripts (if any). Just in case, any ambiguity arises, we will refer to the preprocessing steps and instructions to resolve.
    - If submitter's entry number is 2019CSZ8763, your readme file name will be "readme_csz198763.txt" and writeup file name will be "report_csz198763.pdf". All letters in file name will be in small. Follow this convention strictly.

Note:

- *Students need to test their code on HPC before submission.* ***Code not running on HPC will be given zero marks.***
- *Submission time will be latter of submission time on moodle and time of last github commit.*
- *Late policy will be as informed at the beginning of the course.*

Anti-Plagiarism Policy for Homework 2

Any detected attempts at plagiarism either from parallel/past submissions or the Internet will risk an F-grade in the course.