COL 761 - Homework 2

- You need to do the homework in your already formed team of 3. Make sure to upload your code to the GitHub repo you mentioned in HW0.
- Due: **23rd September 11:59PM.**
- Your code must compile and execute on HPC. You may get 0 for compilation or execution errors
- **Do not copy code from your friend or from the internet. Any plagiarized code will result in an F grade for the course.**

Submission Instructions:

- **Submit** a script (**<RollNo>_install.sh), on Moodle (do not zip it). Only one per team. Multiple submissions will be penalized.** Like previous assignment, it should contain one line to clone your repository, one line to cd into it, and load module commands (Nothing else; don't unzip the archive.) Provide github's https link, not the ssh one.

- **Upload** H**W2_<RollNo>.zip** file to your **GitHub repository**. This RollNo should be of one of your team members. Ex. HW2_CSZ198347.zip. Unzipping it should **produce one folder** of the same **name as the zip file.** This folder should contain:
  - **The source code files.**
  - **README.txt** (i) explains all the files you have bundled, (ii) has entry numbers and names of **all team members** (iii) has instructions on how to execute your code and (iv) contribution in percentage of each student.
    There should be minimal manual overhead required to run your code.
  - A **report file** containing all explanations with plots named as <rollnumber.pdf> eg. CSZ198347.pdf .

- Make sure to use the same roll number to give the name of all submission files (HW2_<RollNo>.zip , <RollNo>_install.sh), etc.
- Do not hardcode the dataset and query file names in your code for question 2.
- There should be **only one submission** per group (on github and moodle).
- Latest timestamp of your last pushed commit and Moodle submission of <RollNo>_install.sh will be treated as your submission time.
- Marks will be distributed to individual members based on their percentage contribution. For example, if someone has contributed 20%, then his/her marks will be 20/33*(marks obtained. However, no extra marks will be given for contributing more than 33%. In case of disputes, we will go by majority vote within that team.)
- Please use only the modules available on HPC, do not install any new packages as HPC doesn't allow this. (Using –U flag to install python packages in user mode is also discouraged as it installs a package which can clash with other people's code that we'll be evaluating after yours.)
- Incase you are not faimilier with HPC you can take help from here http://github.com/kanha95/HPC-IIT-Delhi .

1. This question is about familiarizing yourself 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
#nodes
Series of Node Labels
#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. Use elbow plot to determine the correct value of k in k-means clustering on the dataset generated by `generateDataset_d_dim_hpc_compiled`. Write generic script to generate elbow plot which takes 3 arguments, dataset, dimension, and plot name as follows:

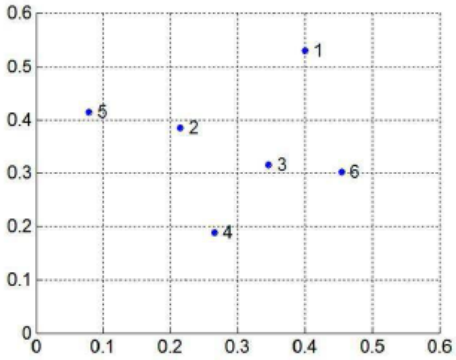
sh elbow_plot.sh <dataset> <dimension> q3_<dimension>_<RollNo>.png

You have to run k-means clustering for different k = {1,2,3,4,.....15} and then plot elbow curve and determine suitable value of k i.e. you have to find out how many clusters are there in the dataset. You can assume Euclidean as a distance measure for the given dataset. The dataset dimension can be 4, 5, 6 or 7. Please mention in the report which dimension you selected. **[20 points]**

Instructions to generate dataset for elbow plot question:

1. Unload any existing module:
   `$ module purge`
2. Load the following module on HPC:
   `$ module load compiler/gcc/9.1.0`
3. Run the file `generateDataset_d_dim_hpc_compiled` (provided on Moodle) to generate dataset as:
   `$ ./generateDataset_d_dim_hpc_compiled <RollNo> <dimension>`
   Egample: `./generateDataset_d_dim_hpc_compiled CSZ198763 7`
4. Make sure the file has execute permissions (running `ls -l` should give `-rwxr-xr-x` if not use command `chmod ug+x dataset_gen_hpc_compiled`). Only 4,5 , 6 and 7 dimensions are allowed.
5. This should generate a file named `generated_dataset_<dimension>D.dat`
6. Different roll numbers will give different datasets, so all teammates use the one that will be used to name the submission file.
7. Use the correct format: CSZ198347 is **correct**. 2019CSZ8347 is not. Also, use capital letters. csz198347 is also not correct.

3. Draw the dendrogram for single linkage clustering on the data below. Show all the steps. **[5 points]:**



Point	x	y
1	0.40	0.53
2	0.22	0.38
3	0.35	0.32
4	0.26	0.19
5	0.08	0.41
6	0.45	0.30

What is the complexity of the fastest possible algorithm? Give your algorithm's pseudocode and complexity analysis. **[10 points]**