# DSA5101 Project Instructions

## Objective

In this course, you have learnt specialised algorithms which can be applied to different types of data. You are expected to apply at least two types of algorithms (dimensionality reduction, clustering vectors, PageRank, clustering graphs etc) across **up to two** datasets (in-memory algorithms would suffice). You should also analyse and compare the results of the different methods you choose to use.

Some ideas:

- On your favourite graph dataset, you can compute the PageRank scores of its nodes. Try to find interesting relationships, for example the correlation between in-degree of nodes and their PageRank scores. Apply some graph clustering algorithms, and measure the performance using conductance.

- Compare point clustering methods and hierarchical clustering methods. Perhaps also see if dimensionality reduction before clustering helps.

## Project Submission Format

Your submission should consist of two parts, your **code** and a **pdf report**.

- **Code**: It is probably easiest to use Python and associated libraries such as numpy, scikit-learn etc. You may submit a Jupyter Notebook. The code should be runnable as it is. If some parts require long training, then please attach a saved model that can be loaded into the notebook and comment out the training steps (but still include them as comments!) in your submission. Include clear comments (use the markdown capabilities of jupyter!) to explain what you are doing at each step.

If you choose to use another programming language, please indicate it in your report.

- **PDF report**: This report should include an introduction of your dataset, visualizations, outputs of your code and most importantly, discussions and analysis of your results.
  Feel free to use algorithms that were not taught in this course, however you should include a detailed description of the algorithm.

## Mandatory Components

1. An introduction of your dataset(s) (maximum two) with visualizations.
2. A clear outline of the methods (at least two) you plan to apply and what you wish to find out about the dataset(s).
3. Performance metrics to compare the different methods that you employ.

## Grading Criteria

1. Scientific correctness of your methodologies
2. Correct use of libraries
3. Clear documentation of code and findings
4. Creativity and style

## Datasets

You are encouraged to find a dataset that you are genuinely interested in. Some resources for datasets include:

1. UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/index.php]

2. Kaggle [https://www.kaggle.com/datasets]

3. GovTech [https://data.gov.sg/dataset?organization=govtech]