Assignment 2: Clustering (25%)

Deadline: 11:59pm, Sunday, 24th May, 2020

The main task of this assignment is to **apply different clustering methods on the two datasets and to find the best model** with optimal parameters. In this assignment, you should use **at least two different clustering evaluation/validation metrics** for each clustering method and present and explain the evaluation results.

Important Notes

1) *Please refer to **lecture notes, lab materials, and resources of week 9 and week 10** for different clustering methods and clustering evaluation methods. We have also attached some useful resource links in Canvas in week10 for clustering evaluation metrics.*
2) *Please **proceed your own way** if we do not specify it in the assignment specification.*
3) *Please come to our **live QA session** (Thursday to Sunday, 8pm to 9pm) if you have any questions. We are always there for you!*
4) *You can use **packages or code from the tutorial**. If you use any other package or code, please put the reference at the bottom of the code. Otherwise, **it will be considered as plagiarism and the relevant section will not be marked***

Assignment Details

In this assignment, you will be given the following **two datasets**:

1) **Travel Reviews Dataset:** This dataset contains reviews by travellers in 10 categories on destinations across East Asia. Each rating is mapped as *Excellent(4), Very Good(3), Average(2), Poor(1), and Terrible(0)* and the average rating per traveller is reported. There is no gold cluster for this dataset.
2) **ICMLA 2014 Accepted Papers Dataset**: This dataset comprises the metadata for the 2014 ICMLA conference's accepted papers, including ID, paper titles, author's keywords, and abstracts. The column 'session' would represent the gold cluster.

You are to submit an ipynb file that contains **the following (section 0, 1, 2, and 3) contents.** The ipynb template can be found in the **A2_ipynb_template**

Section 0. Data Setup

- Load the given two datasets
- ***[Optional]*** Preprocess the data if you think it is necessary.

You can apply different data preprocessing techniques in different data and clustering models if you need.

Section 1. K-means Clustering (4 marks)

- Use the data that you loaded and preprocessed
- Train a k-Means model on the data (*The optimal k should be selected as mentioned in Lecture 10. The optimal k can be selected based on the evaluation.*)
  - Try at least two different similarity measures (e.g. Euclidean) as mentioned in Lecture 9
- Evaluate the model performance by at least two different evaluation metrics and visualise using graphs and explain details

  In the report (***sec 4***), explain the result and justify your answer (with graphs or tables):
  - Similarity measure selection
  - The optimal k selection
  - Clustering Validation/Evaluation (also justify why you used the specific evaluation metrics for each dataset)

Section 2. Hierarchical Clustering (4 marks)

- Use the data that you loaded and preprocessed
- Train hierarchical clustering (*either Agglomerative clustering or Divisive clustering)* on the dataset. (*The best distance metrics (from single-link, complete-link, and average link as mentioned in Lecture 9) will be selected based on the evaluation.*)
  - Try at least three different similarity measure (e.g. Euclidean) as mentioned in Lecture 9
- Evaluate the model performance by at least two different evaluation metrics and visualise using graphs and explain details

  In the report (***sec 4***), explain the result and justify your answer (with graphs or tables):
  - Similarity measure selection
  - The best distance metrics (single link, complete link and average link) for each dataset
  - Clustering Validation/Evaluation (also justify why you used the specific evaluation metrics for each dataset)

Section 3. DBSCAN Clustering (4 marks):

- Use the data that you loaded and preprocessed
- Train DBSCAN clustering with the optimal min_points value and epsilon for each dataset (*The optimal parameters will be selected based on the evaluation*)

    - Try at least three different similarity measures (e.g. Euclidean) as mentioned in Lecture 9
- Evaluate the model performance by at least two different evaluation metrics and visualise using graphs and explain details

  In the report (***sec 4***), explain the result and justify your answer (with graphs or tables):
    - The optimal values of min_points, and epsilon
    - Similarity measure selection
    - Clustering Validation/Evaluation (also justify why you used the specific evaluation metrics for each dataset)

Section 4. Report (10 marks)

The report should be no more than 5 pages (appendix and reference excluded). Please download the **Assignment 2 report (5 pages maximum) template**. You should submit a pdf version of the assignment 2 report.

Section 5. Programming styles (3 marks)

Your program needs to be easily readable and well commented (like the COMP5318 Labs). The following are expected to be satisfied:

- **Readability & Consistency**: Easy to read, and consistent in style
- **Coding Comments**: Comments clarify meaning where needed
- **Robustness**: Handles erroneous or unexpected input

Assignment 2 Submission Method:

You need to submit **a report (pdf file)** and **a code file (ipynb file)**.

1. **Report**: pdf file with a maximum of 5 pages (appendix and reference excluded)
2. **Code**: ipynb file that can be successfully run on Colab.

Please upload both datasets (any file format is ok) to your Google drive, and load the datasets through Google drive file id in your code. *If you want to use Jupyter, please submit all the dataset in an additional zip file.*

Due date: 11:59pm, Sunday, 24th May, 2020

Submission: Canvas Assignment 2 Submission Box

Assignment 2 Late Submission:

Late submissions are allowed **up to 3 days late**. *A penalty of 5% per day late will apply.*