**Spring 2025 | DATA 266 | Homework -2**
**Deadline – 11.59 PM – 03/23/2025**
**20 Points**

**Question # 1 (7 pts):**

Apply combinations of advanced LLMs and prompt engineering techniques on IMDB movie review dataset (select 200 sample from the dataset). Finally, compare the performance in terms accuracy, precision, recall, and f1-score.

- LLMs: select any three LLMs released in 2024 or 2025
- Prompting techniques:
    - Zero shot prompting
    - 10 shots prompting
    - Chain-of-thought prompting with 10-shots
    - Self-consistency prompting with 10-shots

```
! pip install datasets
from datasets import load_dataset
imdb = load_dataset("imdb")
```

**Question # 2 (7 pts):**

Use the 20 Newsgroups dataset. Apply BERTopic using the three different embeddings and calculate coherence score for each case. Annotate the topics from your understanding of the topics and then make a comparative analysis on them.

a. **Doc2Vec** embeddings
b. **MPNet** embeddings (all-mpnet-base-v2 from SBERT).
c. **advanced SBERT** model (e.g., all-roberta-large-v1 or Instructor-XL).

```
from sklearn.datasets import fetch_20newsgroups
newsgroups_train = fetch_20newsgroups(subset='train', remove=('headers', 'footers', 'quotes'))

```

**Question # 2 (6 pts):**

Use the following code – Siamese network using contrastive loss.

https://keras.io/examples/vision/siamese_contrastive/

Train a Siamese network using Quora Question pairs dataset to predict whether a question pair duplicate or not duplicate (https://huggingface.co/datasets/AlekseyKorshuk/quora-question-pairs). Use any recently published pre-trained model to generate fixed size embeddings for each question pairs, and feed both question embeddings into a cosine similarity layer (distance layer), use the distance layer as input to a classification layer (sigmoid). Train the network for minimum 50 epochs. Using train and validation split, present model accuracy learning curve, and contrastive loss learning curve. Finally evaluate the model performance on the test dataset.

**You are required to follow:**

1. Submit **one** MS/PDF/Scanned document:
    - Include all the steps of your calculations.
    - Include the summary of the model.
    - Attach screenshots of your code.
    - Attach screenshots – showing first few epochs of model training.
    - Attach screenshots of the important code outputs such as confusion matrices, learning curves, and classification reports.
2. Source code:
    a. Python (Jupyter Notebook)
    b. Ensure it is well-organized with comments and proper indentation.

- **Failure to submit the source code will result in a deduction of full/partial points.**
- Format your filenames as follows: "your_last_name_HW1.pdf" for the document and "your_last_name_HW1_source_code.ipynb" for the source code.
- Before submitting the source code, please double-check that it runs without any errors.
- Must submit the files separately.
- Do not compress into a zip file.
- HW submitted more than 24 hours late will not be accepted for credit.