Natural Language Processing
DATS 6312
Date: *17th March 2025*

Exam 1

- **Show ALL Work, Neatly and in Order.**
- **No** credit for Answers Without Work.
- Open Books, Open Notes.

1 Introduction

This exam consists in using NLP techniques to classify a given text to its class. The dataset includes some textual information and labels as Target. The goal of this exam is to use your NLP knowledge and data preprocessing that you learnt over time to perform text classification. The format will be as follows:

- Data folder: includes *train.xlsx*
- Sample_train folder: includes *nickname_train.py* & *nickname_inference.py*
- *Exam1_description.pdf*

The average of Quadratic Cohen Kappa score and F1 score is the metric for evaluation.

2 Rules of Exam

Please read these rules **carefully** and if you have any questions please let me know directly.

- You must change the **nickname**. Files with **wrong format will not be graded**. For example, if your assigned nickname is **Greta**, the submission files should be:
  - `Greta_train.py`.
  - `Greta_inference.py`.
  - `Greta_model.pickle`
  - `Greta_labelencoder.joblib`
  - `Greta_vectorizer.joblib`
- You can use Sklearn and any other packages for training **except gensim, transformers architecture and any neural network based models**, in this competition.
- You can only use the data you are given.

- You can do any kind of pre-processing with the training data.
- You are not allowed to share your results with others. If we find out you will get **zero** grade for the Exam.
- You are not allowed to copy code or ideas from any students in the class. If we find out you will get **zero** grade for the Exam.
- You are allowed to use internet. Any codes from external Github requires citation. No citation will cause reduce grade.
- You can use your own computer to do the coding, but you must run the code on **AWS instance**.

3 Clarification on Training

1. The split column has **train** and **test** values in train.xlsx. Use them accordingly for training purposes.
2. A sample training script is given. Please run it first to get a baseline result.
3. For any new packages you use, please add **os.system('pip3 install package_name')**
4. The sample run command using argparser is provided as the following:

   ```
   python3 nickname_inference.py --data_path ../Data/ --train_df train.xlsx --model_path . --split test
   ```

5. You must **run your inference file in the terminal using the above command** before you submit to avoid losing unnecessary points.

4 Deliverables

1. You must upload files with correct format to Blackboard in order to get your grade, don't zip any files. **Files with wrong format will not be graded.** For example, nickname is 'Greta', then you should submit:
   - `Greta_train.py`.
   - `Greta_inference.py`.
   - `Greta_model.pickle`
   - `Greta_labelencoder.joblib`
   - `Greta_vectorizer.joblib`
2. Make sure your NickName_inference.py runs on argparser, please test using the command before you submit.

   ```
   python3 nickname_inference.py --data_path ../Data/ --train_df train.xlsx --model_path . --split test
   ```

5 Grading

1. There is a reserved set that is called held out set. Your inference code will run on the separate split.

2. Your model is going to be evaluated on the held out set and the metric is going to be calculated.

3. There will be a leaderboard including your nickname and score, ranking based on the metric score.

4. The final score will be evaluated based on the leaderboard.