EAS 595
Fundamentals of Artificial Intelligence
Project 2
Total Marks 100

1 Task

The task of this project is to implement a movie genre prediction model. It's a multi-label classification problem. Your task is to classify the movies based on their synopses. You are provided the training and test datasets. Deadline to submit the code and the rerport is **April 30th 11:59 pm EDT**.

2 Dataset

There are 20 classes of genres in the dataset:

Drama, Comedy, Romance Film, Thriller, Action, World cinema, Crime Fiction, Horror, Black-and-white, Indie, Action/Adventure, Adventure, Family Film, Short Film, Romantic drama, Animation, Musical, Science Fiction, Mystery and Romantic comedy

A single movie can have multiple genres associated with it. E.g., Titanic has the following genres associated with it: *'Drama'*, *'Action/Adventure'*, *'Romance Film'*, *'Romantic drama'*.

Following are the first three entries in the training dataset:

movie_id	movie_name	plot	genre
23890098	Taxi Blues	Shlykov, a hard-working taxi driver...	['World cinema', 'Drama']
31186339	The Hunger Games	The nation of Panem consists...	['Action/Adventure', 'Action', 'Science Fiction', 'Drama']
20663735	Narasimham	Poovalli Induchoodan is sentenced...	['Musical', 'Action', 'Drama']

movie_id: Unique integer to identify the movie.
movie_name: String describing the name of the movie.
plot: String describing plot of the movie.
genre: List of strings describing the genres a movie belongs to.

3 Plan of Work

1. **Data pre-processing:** Process the original CSV data files into a Numpy matrix of Pandas Dataframe. Remove the stopwords from the data and pre-process it.
2. **Create a Machine Learning model**: Create a machine learning model using any algorithm and use the information provided in the training dataset to predict the genres associated with the movie. You should implement Term Frequency-Inverse Document Frequency (TF-IDF) and use these features as vectors for the machine learning model.
3. **Evaluation**: Generate the predictions for the test dataset using your trained model and report the F1 score.

4 Evaluation

1. **Task 1**: 20 points for removing the stopwords and pre-processing the data.
2. **Task 2**: 60 points for creating the machine learning model.
3. **Task 3**: 20 points for evaluating the model's performance on the test dataset. (Points also depend upon the F1 score obtained.)

5 Deliverables

There are two deliverables: Report and code. After finishing the project, you may be asked to demonstrate it to the TAs, particularly if your results and reasoning in your report are not clear enough.

1. **Report**: The report should explain the logic that you have used to implement each part of the assignment and describe the results from each of the three tasks. Your report should be named lastname_firstname_proj1.pdf.
2. **Code**: You can submit multiple files, but the name of the entrance file should be main.py or main.ipynb. Please provide necessary comments in the code. Python code and data files should be packed in a ZIP file named lastname_firstname_ proj1code.zip.
3. **Video**: 5-10 mins Panopto Video with the demo, which can be embedded separately.

Submit the Python code and pdf on UBLearns.