# National University
**of Computer & Emerging Sciences**

Program: BS (SE & AI)
Semester: Spring 2025
Target CLO: CLO 5
Course(s): AI2002 - Artificial Intelligence
AI4007 - Applied Artificial Intelligence

Assignment#02
Due Date: 23-03-2025
Marks: 10
Instructor: Waqas Ali

1. You are given two files (tab separated values) named **film-genres-train.tsv** and **film-genres-test.tsv**. As the names suggest one file contains the training dataset and the other contains the test dataset. Both the test and train datasets contains some text describing each movie and the genre of the movie. You have to solve a classification problem where you are required to predict the genre of a movie (where the description of the movie is available). You are required to use the Naive Bayes classifier to solve this problem.

   You are required to provide the number of **correct predictions** and and the number of **incorrect predictions** for **each genre** from the **test dataset**.

## Strategy

You may use the following strategy to solve this problem. Other solutions also exists, you are free to use any of them:

To classify movie genres using the Naive Bayes algorithm, begin by splitting each movie description into individual words using spaces. Next, remove any special characters or non-alphabetic symbols from the text to ensure only meaningful words remain. After that, eliminate common English stop words—words like "the," "is," "and," and "of"—as they occur frequently but do not contribute significantly to distinguishing genres. This step helps in focusing only on content-rich terms. A python list containing the stop words are in the file named **stop-words.py**.

For each cleaned description, calculate the frequency (counts) of every remaining word. This results in a dictionary where each key is a vocabulary word (excluding stop words), and the corresponding value is the number of times that word appears in the description. These frequency dictionaries serve as feature representations for the movie texts.

Using these dictionaries as input features and the known genres as target labels, train a Naive Bayes classifier. Once trained, apply the classifier to the test dataset to predict the genres based on word frequencies. Use the standard steps of the Naive Bayes algorithm to perform classification and obtain the results.

## Note

While popular Python libraries provide built-in functions to automate most steps of text preprocessing and classification, this assignment must be completed without relying on them to help you understand each part of the process in depth. The use of any python library will lead to deduction of marks.