كلية الحاسبات وتقنية المعلومات
FACULTY OF COMPUTING
& INFORMATION TECHNOLOGY

CS-681: Assignment #1

Due date: February 27, 2025 (11:59)

Objective:
The aim of this assignment is to give you hands-on experience for Natural Language Processing (NLP) tasks using traditional methods and different pre-trained embedding.

Instructions:
In this assignment, you will work with an NLP dataset of your choice. You will create different word embeddings and build logistic regression and naïve bayes models to solve the NLP problem.

Part 1: Data Selection and Preprocessing

1. **Dataset:**
   - Sentiment analysis (IMDB reviews),
   - Text classification (News Groups)
     - from sklearn.datasets import fetch_20newsgroups
     - cats = [' rec.autos', comp.graphics ']
     - newsgroups_train = fetch_20newsgroups(subset='train', categories=cats)

   - Load the dataset and explore its structure, labels, and features.

Part 2: Implementing Word Embeddings

1. **Embedding Techniques:**
   - Explore different embedding techniques for representing text:
     - Traditional methods: TF-IDF or Count Vectorizer.
     - Pre-trained embeddings: Word2Vec, GloVe, or FastText.
   - Implement at least two different embedding techniques for the same dataset to see how they affect the model's performance.

Part 3: Building Logistic Regression and Naïve Bayes models

1. **Models Development:**
   - Apply Logistic Regression and Naïve Bayes classifiers using traditional methods and pre-trained embeddings.
   - Report accuracy and other metrics on training data and testing datasets

Submission Requirement

- A Jupyter Notebook or code script with clear documentation and comments.
- A report summarizing your approach evaluation results