Natural Language Processing - AIGC 5501
Mid-Term (20% of Final Grade)
Exam Window: 11.59AM – 11.59PM

Assignment Title: Sentiment Analysis using Word Embeddings

Introduction:
In this assignment, you will perform sentiment analysis on movie reviews using word embeddings. The dataset provided "Bag of Words Meets Bags of Popcorn," where the task is to predict the sentiment of movie reviews as positive or negative. In the mid-term exam, we will work till creating those word embeddings from the data.

Objective:
The objective of this assignment is to utilize word embeddings, particularly Word2Vec, to convert movie reviews into numerical representations.

Tasks:
Data Exploration: Explore the provided dataset to understand its structure, including the format of the reviews and the sentiment labels (positive or negative).

Preprocessing: Preprocess the text data by removing HTML tags, punctuation, and stopwords. Tokenize the reviews and convert them to lowercase.

Word Embeddings: Train a Word2Vec model on the preprocessed movie reviews to generate word embeddings. Experiment with different vector dimensions and window sizes to optimize the embeddings.

Evaluation Criteria:
Read the notebook - **Mid-Term Word Embeddings with Gensim.ipynb,** run all the codes and prepare a step-by-step report on what we are trying to achieve.

Points:
a) Run the code from your system (Data attached)
b) Add code comments in each step.
c) Prepare MS Word report/ppt to present your findings and recommendations.

PS:

- *Start early the code runs may take time*

Submit:

a) The executed notebook in html/.ipynb/pdf format
b) The Report/PPT