Work Integrated Learning Programmes Division
M.Tech (Data Science and Engineering)
Natural Language Processing
DSECLZG530
2023 -24

# Assignment 1

**Data set Link:**
https://drive.google.com/file/d/1fQYedyo5KGYT6g6v5dJKnO9qOuJ8q4M8/view?usp=sharing

## Problem Statement:

The goal of Part I of the task is to use raw textual data in language models for recommendation based application.

The goal of Part II of task is to implement comprehensive preprocessing steps for a given dataset, enhancing the quality and relevance of the textual information. The preprocessed text is then transformed into a feature-rich representation using a chosen vectorization method for further use in the application to perform similarity analysis.

## Part I

### Sentence comparison using N-gram: (3 Marks)

Let a search engine powered by language model recommend which of the below sentences are most relevant w.r.t to given training corpus. Design a probabilistic language model to compare below test sentences for recommendation using bigram. Use all the instances in the dataset as a training corpus.

**Test Sentence 1:** **“**Petter Mattei's 'Love in the Time of Money' is a visually stunning film to watch.”
**Test Sentence 2:** “I sure would like to see a resurrection of an updated Seahunt series with the tech they have today”

## Part II

Perform the below sequential tasks on the given dataset.

### i) Text Preprocessing: (2 Marks)

a. Tokenization
b. Lowercasing
c. Stop Words Removal
d. Stemming
e. Lemmatization

### ii) Feature Extraction: (2 Marks)

Use the pre-processed data from previous step and implement the below vectorization methods to extract features.

Word Embedding using TD-IDF

### iii) Similarity Analysis: (3 Marks)

Use the vectorized representation from previous step and implement a method to identify and print the names of top two similar words that exhibit significant similarity. Justify your choice of similarity metric and feature design. Visualize a subset of vector embedding in 2D semantic space suitable for this use case. **HINT: (Use PCA for Dimensionality reduction)**