

NLP- MTech AI & DS, MU

Assignment 1

Programming and research paper tasks

Instructor: Dr. Dheeraj

Start Date: 05-03-2025

End Date: 10-03-2025

Note that this is not a group project or assignment. You must work individually.

- a) Read the research papers which are uploaded in the Github (**Common for all**).
 - b) Text preprocessing on 10,000 texts, where each text contains a minimum of 200 characters (**Common for all**).
1. (**Roll Nos: 1-5**) Develop an NLP pipeline that performs:
 - c) Given a word and its possible meanings, select the correct sense based on a sentence.
 - d) Compare the overlapping words in the sentence and the sense definition.
 2. (**Roll Nos: 6-10**) Develop an NLP pipeline that performs:
 - c) Generate Text N-Grams Without Using NLTK:
Implement n-gram extraction (bi-grams, tri-grams, etc.) from a given text.
 - d) Ignore stopwords and punctuation in the process.
 3. (**Roll Nos: 11-15**) Develop an NLP pipeline that performs:
 - c) Compute TF-IDF scores for each word in a document.
 - d) Return the top N keywords with the highest scores.
 4. (**Roll Nos: 16-20**) Develop an NLP pipeline that performs:
 - c) Implement extractive text summarization by scoring sentences based on:
Word frequency, Sentence length, TF-IDF scores.
 - d) Find the most frequent POS (Part of Speech) tag in a given text.

5. (Roll Nos: 21-25) Develop an NLP pipeline that performs:

c) BPE to compress text data and reduce vocabulary size.

d) Given a BPE-encoded text, reconstruct the original words by merging subwords.

6. (Roll Nos: 26-33) Develop an NLP pipeline that performs:

c) Find the most frequently occurring word in a text, excluding common stopwords.

d) Find the most semantically similar sentence to a given input using TF-IDF.