भारतीय प्रौद्योगिकी
संस्थान जम्मू
INDIAN INSTITUTE OF
TECHNOLOGY JAMMU

Discipline of
Computer Science & Engineering

Assignment 1 - Executive M.Tech(AI)
Course: NLP **Due Date: 31/03/2025**

Instructions

1. Assignment submissions will be accepted only via Google Classroom. Submissions through email or any other methods will NOT be accepted. Please join Google Classroom using the following link: https://classroom.google.com/c/NzUyOTIwOTIxOTU2?cjc=y66aldcx
2. This is a graded assignment (10 points). Penalties may be applied to those who do not submit the assignment before the due date.
3. The submission deadline is 31/03/2025. Please submit a single.pdf file using the nomenclature 'NLP<AssignmentNumber>_<EnrollmentNumber>.pdf,'

   for example, 'NLP1_2022RCS2021.'

1. Write a Python program to download and preprocess a text corpus of your choice, such as from NLTK or Hugging Face. Implement functions to compute Term-Based Metrics (TBM), including Term Frequency (TF), Document Frequency (DF), and Inverse Document Frequency (IDF). Calculate the raw and normalized term frequency of words in the corpus and compare different weighting schemes. Next, compute the TF-IDF score for each term and analyze its significance in identifying important words. Finally, visualize the top-ranked terms using a bar chart or word cloud and discuss how preprocessing impacts the results.
2. Write a Python program to download and preprocess a text corpus, then train word embeddings using Continuous Bag of Words (CBOW) and Skip-gram models with the Word2Vec algorithm. Additionally, explore other embedding techniques such as GloVe or FastText. After training the embeddings, use PCA and t-SNE to reduce their dimensionality and visualize the word relationships in a 2D or 3D space. Compare the differences in word clusters and interpret how CBOW, Skip-gram, and other embeddings capture semantic relationships.