- 1. Term Weighting Schemes
  - Description
- 2. Topic Modeling
  - Description
- 3. Dimensionality Reduction
  - Description
- Estimating Time and Difficulty:
- Quiz:


In Week 3 of the course EE6405, you've covered two primary topics: **"Term Weighting Scheme"** and **"Topic Modelling"**, as well as the associated concept of **"Dimensionality Reduction"**. The learning materials used were primarily from the slide deck titled "Term Weighting Scheme and Topic Modelling" for students. Here's a breakdown of what you've learned:

# 1. Term Weighting Schemes

- **Pages**: 2-7, 10-14
- **Concepts Covered**:
  - **TF-IDF** (Term Frequency-Inverse Document Frequency): Pages 3-7
  - **BM25**: Pages 11-14

**Description**

TF-IDF and BM25 are foundational concepts in understanding how terms are weighted in the context of information retrieval and natural language processing. TF-IDF quantifies the importance of a word relative to other words in the document set, while BM25 is a probabilistic approach that refines the basic TF-IDF approach to improve retrieval effectiveness.

# 2. Topic Modeling

- **Pages**: 15-42
- **Sub-concepts**:
  - **LDA (Latent Dirichlet Allocation)**: Pages 15-24
  - **LSA (Latent Semantic Analysis)**: Pages 25-30

- **pLSA (Probabilistic Latent Semantic Analysis)**: Pages 31-32
- **NMF (Non-negative Matrix Factorization)**: Pages 34-41

**Description**

You explored different algorithms that help in identifying topics in large sets of documents. These algorithms assist in uncovering hidden thematic structures in document collections, making them crucial for summarization, classification, and retrieval tasks.

# 3. Dimensionality Reduction

- **Pages**: 42-57
- **Sub-concepts**:
  - **PCA (Principal Component Analysis)**: Pages 43-51
  - **SVD (Singular Value Decomposition)**: Pages 52-57

**Description**

Dimensionality reduction techniques, such as PCA and SVD, are critical for reducing the number of variables under consideration, by obtaining a set of principal variables. This is particularly useful in processing and visualizing high-dimensional data sets, and in the context of NLP, for reducing the complexity of models.

# Estimating Time and Difficulty:

- **Term Weighting Schemes**: Approximately 2 hours. Medium difficulty due to technical details.
- **Topic Modeling**: Approximately 3 hours. High difficulty because of complex mathematical concepts.
- **Dimensionality Reduction**: Approximately 2 hours. High difficulty due to advanced algebra and statistics involved.

# Quiz:

**Question:** Which topic modeling method is based on a probabilistic framework and generalizes the LSA?

- A) LDA
- B) pLSA
- C) NMF
- D) SVD

**Answer:** B) pLSA

This quiz question tests your understanding of the theoretical underpinnings of different topic modeling techniques, specifically focusing on the probabilistic aspect and relationship with LSA.