Institute of Business Administration
Introduction to Text Analytics
Assignment 02
Due Date: 23rd February 2025 (11:55 PM)

Note: This is an individual assignment; hence, everyone must submit it separately.

Assignment: Clustering News Headlines Using Different Embedding Techniques

Objective:
The goal of this assignment is to explore how different text vectorization techniques impact clustering results. You will apply **K-Means clustering** on a dataset of news headlines using the following different embedding methods:

1. **Bag of Words (BoW)** – Count Vectorizer and TF-IDF
2. **Latent Semantic Analysis (LSA)** – Dense representations obtained via Singular Value Decomposition (SVD)

Compare how preprocessing choices (e.g., stopword removal, stemming, lemmatization) and vectorization techniques affect the clustering quality, as measured by **Within-Cluster Sum of Squares (WSS)** and **Silhouette Score**.

Dataset:
You are provided with a dataset (news_Feb_14.csv) containing around 450 news headlines.

Tasks:

1. **Preprocessing & Tokenization**
   - Convert text to lowercase.
   - Explore the impact of:
     - **Stopword removal** (with and without stopword removal)
     - **Stemming vs Lemmatization**
     - **N-grams** (e.g., unigrams vs bigrams)
2. **Text Vectorization:**
   - **BoW:**
     - Count Vectorizer
     - TF-IDF
   - **LSA:**
     - Apply TruncatedSVD on TF-IDF vectors.
     - Experiment with different numbers of dimensions (e.g., 50, 100, 200).
3. **Clustering using K-Means:**
   - Perform K-Means clustering using fixed values of k = 5, 9, and 13. Please set random_state parameter to your ERP ID in K-Means initialization. For instance, km = KMeans(n_clusters =4, random_state=12345) if your ERP_ID = 12345.
   - Report the Within-Cluster Sum of Squares (WSS) (kmeans.inertia_ in sklearn) and Silhouette Score.
   - Compare the results across different embeddings and preprocessing techniques.
4. **Analysis & Interpretation:**
   - Identify which embedding technique resulted in the best clustering.
   - Discuss how preprocessing choices impacted the results.
   - Provide sample headlines from different clusters to analyze coherence.

Evaluation Criteria:
Your submission will be evaluated based on:

- Correct implementation of vectorization and clustering techniques
- Comparison and justification of different approaches
- Quality of analysis and interpretation of clustering results
- Proper use of evaluation metrics (WSS, Silhouette Score)
- Code clarity and documentation

Deliverables:

1. Python code notebooks that you used for experimentation.
2. Filled version of the attached document “A2_Assessment.docx”.