Institute of Business Administration
Introduction to Text Analytics
Assignment 03
Due Date: 9th March 2025 (11:55 PM)

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

Assignment: Clustering News Headlines Using Word2Vec Averaging and Doc2Vec Embeddings

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. Word2Vec – Averaged word embeddings to obtain headlines embeddings
2. Doc2Vec

Compare how these 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. **Text Vectorization:**
   - **Word2Vec Averaging:**
     - Train a Word2Vec model on the dataset (alternatively, you may use pre-trained embeddings like Google's Word2Vec or GloVe or FastText).
     - Compute headline embeddings via vector averaging.
   - **Doc2Vec:**
     - Train a Doc2Vec model to obtain embeddings for each headline.
2. **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.
3. **Analysis & Interpretation:**
   - Identify which embedding technique resulted in the best clustering.
   - Compare the performance of word2vec and doc2vec embeddings with those used in previous assignment (Assignment 02).

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 "A3_Assessment.docx".