TASK 3

Manish Kanuri
NUID: 002315456

Objective:
The notebook explores clustering techniques on textual data using the 20 Newsgroups dataset. The focus is on preprocessing text, extracting features, applying dimensionality reduction, and evaluating clustering performance.

Key Components:
Data Preprocessing:

Fetches 10 categories from the 20 Newsgroups dataset.
Performs tokenization, stopword removal, and vectorization (TF-IDF, CountVectorizer).
Feature Extraction:

Constructs TF, TF-IDF, and PPMI (Positive Pointwise Mutual Information) matrices.
Applies Singular Value Decomposition (SVD) for dimensionality reduction.
Clustering Techniques:

K-Means Clustering: Groups documents into clusters.
Word2Vec (CBOW Model): Trains word embeddings for document representation.
t-SNE Visualization: Projects high-dimensional embeddings into 2D space.
Evaluation Metrics:

Confusion Matrix: Assesses clustering quality.
Silhouette Score: Measures cluster compactness.
Adjusted Rand Index (ARI): Evaluates clustering agreement with ground truth.
Experiments Conducted:

Clustering performed for 3 groups and 10 groups.
Different embedding techniques (SVD-TF, SVD-TFIDF, CBOW) compared.
Summary of Findings:
The TF-IDF representation with SVD provided better clustering results.
Word embeddings (CBOW) captured contextual meaning but needed fine-tuning.
t-SNE visualizations showed some overlap among categories, indicating challenges in clustering purely based on text features.