TASK 1

Name: Manish Kanuri

NUID: 002315456

Objective:
The primary objective of this task is to apply Singular Value Decomposition (SVD) on different term-document matrices (Term Frequency, TF-IDF, and PPMI) to reduce dimensionality and analyze the resulting representations.

Steps Performed:
Importing Libraries:

The necessary libraries such as numpy, pandas, scikit-learn, and scipy were imported to handle data manipulation, text processing, and matrix operations.

Loading the Dataset:

The 20 Newsgroups dataset was loaded using fetch_20newsgroups with specific categories selected. The dataset was preprocessed by removing headers, footers, and quotes to focus on the main content of the documents.

Checking Dataset Integrity:

A check was performed to ensure that the dataset was not empty. If the dataset was empty, an error was raised to prompt a review of the selected categories.

Sample Document Inspection:

A sample document from the dataset was printed to inspect the content and ensure that the data was loaded correctly.

Term-Document Matrix Preparation:

Term Frequency (TF) Matrix: A term-frequency matrix was created using CountVectorizer. The matrix was checked to ensure it was not empty.

TF-IDF Matrix: A TF-IDF matrix was created using TfidfVectorizer.

PPMI Matrix: A Positive Pointwise Mutual Information (PPMI) matrix was computed from the TF matrix. The PPMI matrix was calculated to capture the association between terms and documents.

Applying SVD:

SVD was applied to reduce the dimensionality of the term-document matrices (TF, TF-IDF, and PPMI) to 100 dimensions. The TruncatedSVD function from scikit-learn was used for this purpose.

Output of Transformed Matrices:

The shapes of the transformed matrices after applying SVD were printed to confirm the dimensionality reduction:

TF SVD Shape: (18846, 100)

TF-IDF SVD Shape: (18846, 100)

PPMI SVD Shape: (18846, 100)

Key Observations:
Dataset Loading and Preprocessing:

The dataset was successfully loaded, and preprocessing steps (removing headers, footers, and quotes) were applied to focus on the main content of the documents.

A sample document was inspected to ensure the data was loaded correctly.

Term-Document Matrices:

The Term Frequency (TF) matrix was created successfully, and a sample of the matrix was displayed, showing non-zero terms for the first five documents.

The TF-IDF matrix was computed to weigh terms based on their importance in the document and across the corpus.

The PPMI matrix was calculated to capture the association between terms and documents, which is useful for understanding term co-occurrence.

Dimensionality Reduction with SVD:

SVD was applied to reduce the dimensionality of the term-document matrices to 100 dimensions. This reduction is crucial for computational efficiency and for capturing the most important latent features in the data.

The shapes of the transformed matrices confirmed that the dimensionality reduction was successful, with all matrices reduced to (18846, 100).

Conclusion:
The task successfully demonstrated the application of SVD on different term-document matrices (TF, TF-IDF, and PPMI) to reduce dimensionality. The resulting matrices can be used for further analysis, such as clustering, classification, or topic modeling.

The use of PPMI provided an additional layer of insight into term-document associations, which could be useful for tasks requiring a deeper understanding of term co-occurrence.

The dimensionality reduction step ensures that the data is more manageable for downstream tasks while retaining the most important features.

Future Work:
Clustering and Classification: The reduced-dimensionality matrices can be used for clustering or classification tasks to group similar documents or predict document categories.

Topic Modeling: The SVD-reduced matrices can be used for topic modeling to identify latent topics within the document corpus.

Hyperparameter Tuning: Experimenting with different values for the number of dimensions in SVD to find the optimal balance between dimensionality reduction and information retention.

Code Quality and Structure:
The code is well-structured and follows a logical flow from data loading to dimensionality reduction.

Error handling is implemented to ensure the dataset is not empty and the term-frequency matrix is valid.

The use of sparse matrices (csr_matrix) ensures efficient memory usage, especially when dealing with large datasets.

Final Remarks:
This task provides a solid foundation for understanding the application of SVD in natural language processing tasks. The successful reduction of dimensionality using SVD opens up possibilities for more advanced analyses and applications in text mining and information retrieval.