phase 2&3

Data Cleaning and K-Means Clustering, and optimizing the algorithms

In this phase of the project, we meticulously conducted various tasks using Python:

Data Cleaning:

We employed advanced data cleaning techniques to preprocess the dataset, enhancing its quality and reliability.

Dimensionality Reduction:

Utilizing sophisticated dimensionality reduction algorithms, we aimed to streamline the dataset effectively.

- **PCA (Principal Component Analysis):** Applied PCA to the cleaned data, reducing its dimensionality.
- **SVD (Singular Value Decomposition):** Explored SVD as an alternative dimensionality reduction technique.

K-Means Clustering:

We delved into K-Means clustering to unveil hidden patterns within the cleaned data, experimenting with different values of k for optimal cluster identification.

- **Version 1:** Utilized K-Means without removing outliers.
- **Version 2:** Executed K-Means after the removal of outliers for improved cluster analysis.

Results Visualization:

To present our findings comprehensively, we visualized the results through scatter plots in both two and three dimensions. Each cluster was uniquely color-coded for clarity.

Explore the visualizations and analyses:

- **Code with Outliers:** Link to Visualizations with Outliers
- **Code without Outliers:** Link to Visualizations without Outliers

Feel free to navigate through the code repositories for an in-depth understanding of our data cleaning, dimensionality reduction, and K-Means clustering processes.

Text Preprocessing Integration:

In this enhancement, I incorporated advanced text preprocessing techniques using NLTK:

1. **Tokenization:** Utilized NLTK's tokenizer to convert each document into a list of tokens, enhancing text analysis granularity.
2. **Lemmatization:** Employed lemmatization with WordNet to derive the base form of each word, standardizing vocabulary for improved accuracy.
3. **Stopword Removal:** Implemented the removal of common stopwords, focusing on content-bearing words and reducing dataset noise.
4. **Integration with Previous Phases:** Seamlessly integrated text preprocessing into both Phase 1 (TF-IDF) and Phase 2 (Data Cleaning and K-Means Clustering) processes.

Impact:

- Enhanced dataset quality, providing a more accurate representation of content.
- Improved focus on content-bearing words, contributing to better dimensionality reduction and clustering outcomes.

Explore the updated code to observe the transformative effects of text preprocessing on the overall project.