**Introduction**

After tokenizing the features and obtaining the high-dimensional feature representations, the next steps involved dimensionality reduction and clustering. This report details the process of reducing the dimensionality of the features using both PCA and t-SNE, performing K-means clustering on the reduced data, and evaluating the clustering results using various metrics.

**Dimensionality Reduction**

**PCA (Principal Component Analysis)**

PCA was applied to reduce the dimensionality of the feature set. Two approaches were used:

1. **Elbow Method**: The elbow method was used to determine the optimal number of components. This involved plotting the cumulative explained variance against the number of components and identifying the 'elbow point' where additional components contributed marginally to the explained variance.

2. **Explained Variance Threshold**: PCA was also performed to retain 95% of the variance in the data. The cumulative explained variance was computed to determine the number of components required to achieve this threshold.

**t-SNE (t-distributed Stochastic Neighbor Embedding)**

t-SNE was used to transform the high-dimensional data into a 2-dimensional space for visualization and clustering. Various perplexity values were tested, and the best results were obtained with a perplexity of 30 and 1000 iterations.

**Clustering**

**Determining the Optimal Number of Clusters**

The optimal number of clusters was determined using the Silhouette Score. A range of cluster numbers was tested, and the silhouette scores for each number of clusters were plotted to identify the best number of clusters.

**K-means Clustering**

Based on the silhouette scores, the optimal number of clusters was chosen and K-means clustering was performed on the data reduced by both PCA and t-SNE. The clustering results were evaluated using three metrics: Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index.

**Results**

**t-SNE with K-means Clustering:**

- **Silhouette Score:** 0.4773882031440735

- **Davies-Bouldin Index**: 0.6705032092993364

- **Calinski-Harabasz Index**: 8481.6525313706

**PCA with K-means Clustering:**

- **Silhouette Score**: 0.03951304405927658

- **Davies-Bouldin Index**: 3.89209418426256

- **Calinski-Harabasz Index**: 175.04811157298917

**Observations:**

- The t-SNE transformation followed by K-means clustering produced significantly better results compared to PCA. This is indicated by the higher Silhouette Score and Calinski-Harabasz Index, and the much lower Davies-Bouldin Index.

- The clustering using t-SNE gave better results compared to PCA, as indicated by the higher Silhouette Score and Calinski-Harabasz Index, and the lower Davies-Bouldin Index. This suggests that t-SNE was more effective at preserving the local structure of the data and creating well-defined clusters.

**Saving Clustered Images**

After clustering, the images were saved into directories corresponding to their cluster labels. The clustered images can be found on the following link: [Clustering Results](https://drive.google.com/drive/folders/1iiIQels-ue352-7KodvsvuJWnEzuJ87w?usp=drive\_link).

**Conclusion**

Both PCA and t-SNE were used for dimensionality reduction, followed by K-means clustering. The clustering results were evaluated using Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. t-SNE with K-means clustering showed significantly better results with higher cluster quality as indicated by the evaluation metrics. The images were then saved into cluster-specific directories for further analysis and validation.