ML-ASSESSMENT-3

Shreyas Biradar

December 6, 2024

Abstract

This report presents the results of clustering analysis applied to a dataset of news headlines. Both K-Means and Hierarchical Clustering methods were employed, followed by a comparison of their results. Additionally, dimensionality reduction using Principal Component Analysis (PCA) was explored to optimize clustering performance. A comparative analysis between the original clustering and the PCA-reduced clustering is provided.

1 Introduction

Clustering is an unsupervised machine learning technique used to group similar data points. This report explores clustering techniques applied to news headlines using a TF-IDF feature matrix. The goals of this analysis include:

- Grouping similar headlines into meaningful clusters.
- Exploring the differences between K-Means and Hierarchical Clustering.
- Applying dimensionality reduction (PCA) and analyzing its impact on clustering.
- Interpreting and comparing cluster themes across different methods.

2 Methodology

2.1 Dataset Description

The dataset contains a collection of news headlines. A TF-IDF feature matrix was generated to represent the textual data in numerical form. Dimensionality reduction was performed using PCA for optimization.

2.2 Clustering Techniques

- **K-Means Clustering:** A centroid-based algorithm used to partition data into k clusters.
- **Hierarchical Clustering:** A tree-based algorithm that builds a hierarchy of clusters using a dendrogram.
- **Dimensionality Reduction:** PCA was applied to reduce the dimensionality of the TF-IDF matrix to 50 components before clustering.

3 Results and Analysis

3.1 Elbow Method for K-Means

The optimal number of clusters was determined using the Elbow Method. Figure 2 shows the Elbow graph, where the inertia begins to plateau at $k = 5$.

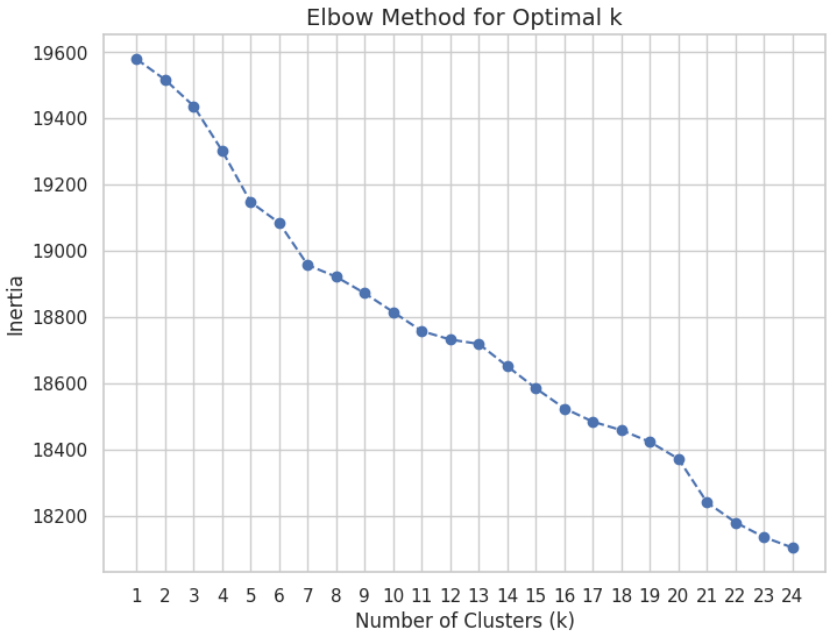


Figure 1: Elbow Method for Optimal k in K-Means

3.2 Distribution of Headlines across Clusters

The distribution visualizes the number of headlines assigned to each cluster, highlighting the relative size and prominence of topics identified by the clustering algorithm. This helps in understanding the prevalence of different themes in the dataset.

Figure 2: Distribution of Headlines across Clusters

3.3 Original K-Means Clustering

The clustering results for the original TF-IDF matrix are shown in Figure 3. Each color represents a cluster, and headlines within the same cluster exhibit similar themes.

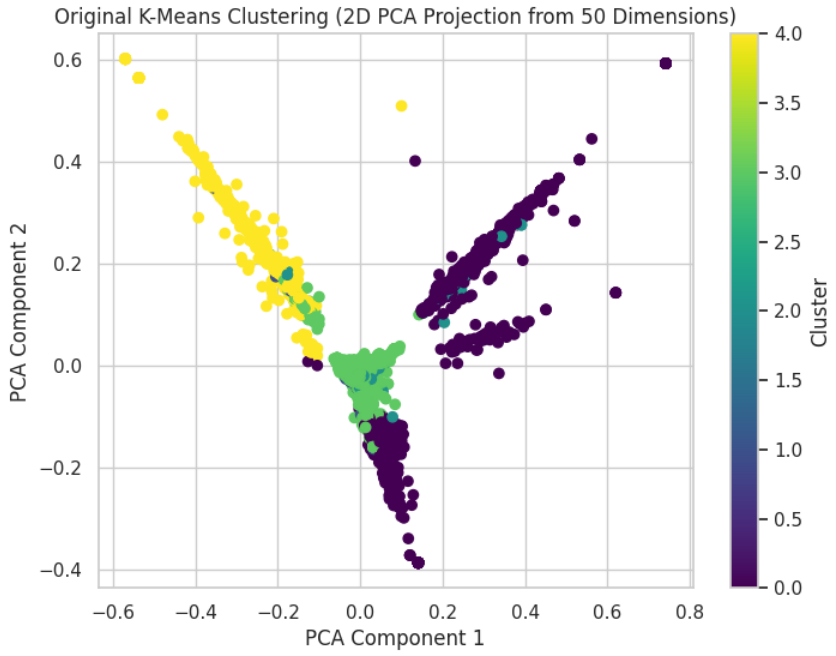


Figure 3: Original K-Means Clustering Results

3.4 PCA-Reduced K-Means Clustering

After applying PCA to reduce the TF-IDF matrix to 50 dimensions, K-Means was applied again. The clusters are visualized in Figure 4.

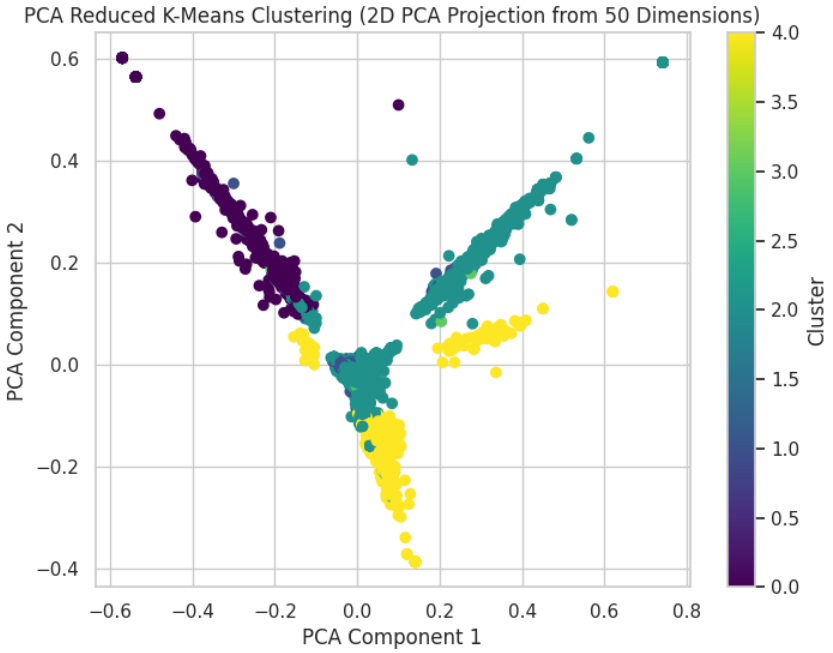


Figure 4: K-Means Clustering After PCA (2D Projection)

3.5 Hierarchical Clustering

A dendrogram for Hierarchical Clustering is displayed in Figure 5. The dendrogram is a tree-like diagram that visualizes the hierarchical structure of data, showing how data points are progressively grouped into clusters. The key to understanding the dendrogram lies in the relationship between the distance metric (typically Euclidean distance) and the formation of clusters.

- The x-axis represents the individual data points or samples, each positioned at the bottom.
- The y-axis shows the distance or dissimilarity between the clusters at each step of the merging process.
- The merging process begins at the bottom, with each sample initially in its own cluster. As the distance between clusters decreases, they are merged, and this is represented by horizontal lines connecting the clusters. The height at which two clusters are joined represents the dissimilarity or distance between them.
- Different colors in the dendrogram help distinguish between the individual clusters at different stages of merging.

The hierarchical clustering process continues until all data points are merged into a single cluster at the top of the dendrogram. This method is particularly valuable for understanding the data's inherent structure, allowing users to choose an appropriate

threshold for cluster formation. The dendrogram visually indicates which clusters are most similar to each other and when they should be merged based on their distance.

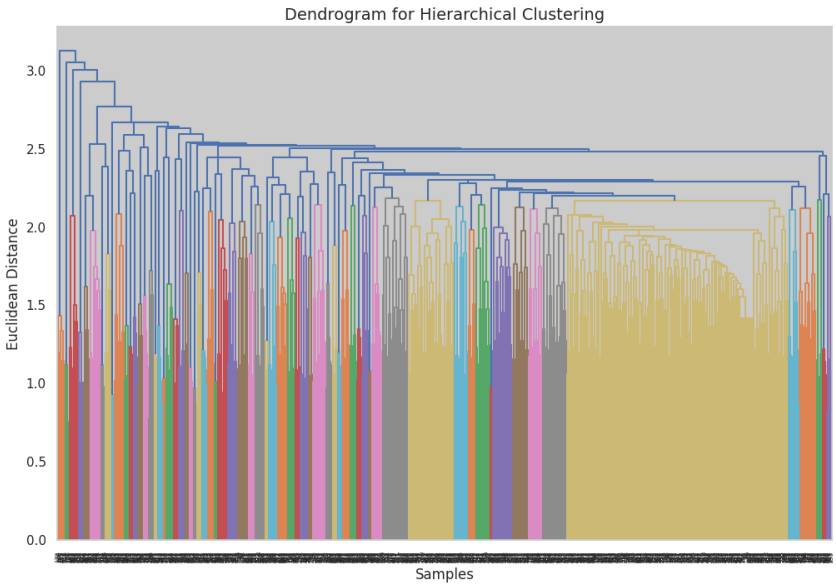


Figure 5: Dendrogram for Hierarchical Clustering

3.5.1 Insights from the Dendrogram and Hierarchical Clustering

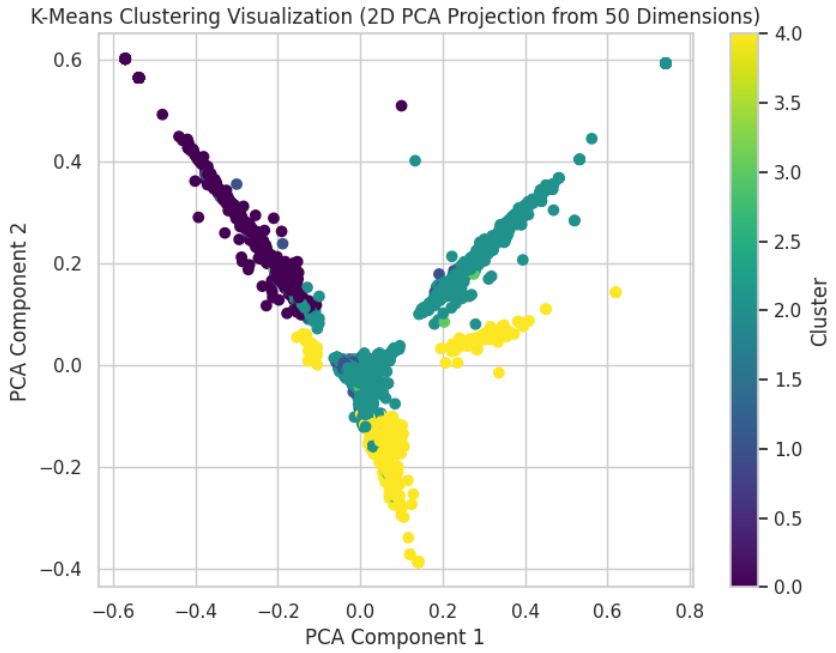
The dendrogram provides valuable insights into the structure and relationships within the data, which can guide further analysis and decision-making:

- Cluster Similarity: The height at which two clusters merge in the dendrogram indicates their similarity. A low merging height suggests that the clusters are highly similar, while a higher merging height indicates that the clusters are more distinct.
- Determining Optimal Clusters: By examining the dendrogram, a threshold can be selected to cut the tree and identify a desired number of clusters. A common approach is to select the threshold where the horizontal lines are the longest, suggesting the most significant differences between clusters.
- Outlier Detection: Outliers or isolated points can be identified as those that do not merge with others until the final stages of the dendrogram. These points are often distant from other clusters and may represent anomalies in the dataset.
- Cluster Hierarchy: The dendrogram illustrates the hierarchical nature of the data, showing how smaller subgroups are nested within larger groups. This hierarchical structure is useful for understanding the levels of granularity at which data can be analyzed.
- Grouping Strategy: The dendrogram helps to explore different grouping strategies by adjusting the threshold at which clusters are merged. By cutting the tree at different heights, one can explore how clusters form at varying levels of similarity.

Overall, the dendrogram serves as an essential tool in hierarchical clustering, enabling a clear and interpretable view of the clustering process. It aids in understanding the intrinsic structure of the data and helps in making informed decisions about the appropriate level of granularity for clustering.

3.6 Comparison of K-Means and Hierarchical Clustering

The comparative results of K-Means and Hierarchical Clustering are shown in Figure 6. While both methods produced similar themes, there were minor differences in cluster assignments.



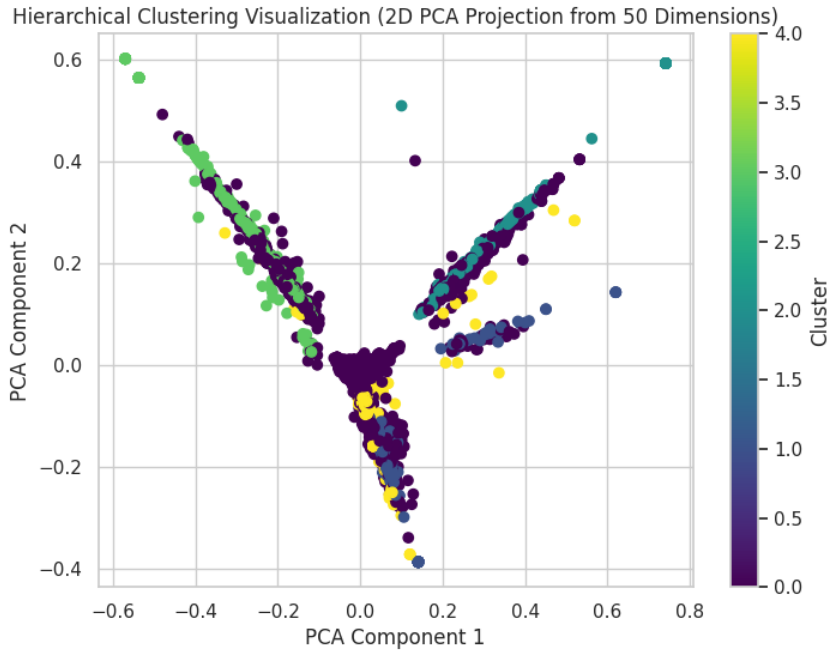


Figure 6: Comparison of K-Means and Hierarchical Clustering

4 Discussion

- The Elbow Method suggested $k = 11$ as the optimal number of clusters.
- PCA significantly reduced the dimensionality of the dataset while preserving clustering performance.
- Hierarchical Clustering produced similar themes to K-Means but with different grouping dynamics due to its tree-based approach.
- The comparative analysis indicates that both clustering methods are effective, but K-Means is more computationally efficient.

The themes for each cluster, as identified by both K-Means and Hierarchical Clustering, are explicitly detailed in the Colab notebook. In the notebook, the word clouds for each cluster are presented with their corresponding dominant themes, offering a clear and visual representation of the clustering results. This enables a deeper understanding of the key themes and patterns emerging from the dataset.

Jacobian Matrix: The heatmap below shows the overlap between clusters generated by K-Means and Hierarchical Clustering using the Jaccard similarity index. Higher similarity scores (closer to 1) indicate overlapping themes, while lower scores highlight divergence in clustering dynamics.

Word Clouds: The word clouds below illustrate the dominant themes within each cluster for both methods. These visualizations help in qualitatively assessing thematic coherence.

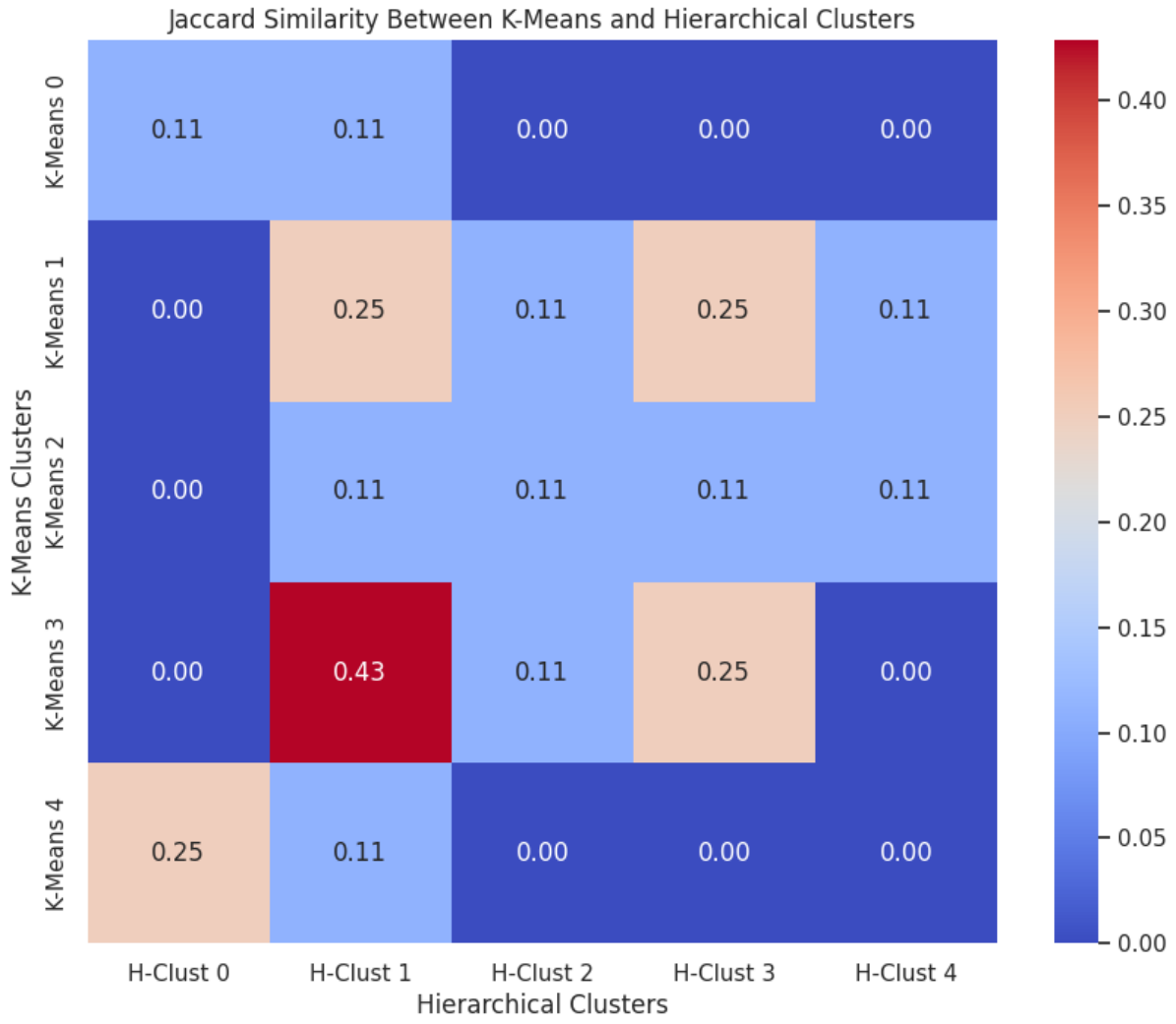


Figure 7: Jacobian Matrix Comparing Themes Across Clustering Methods

5 Conclusion

This report demonstrates the application of clustering techniques to textual data using a TF-IDF matrix. Both K-Means and Hierarchical Clustering effectively grouped similar headlines. PCA proved useful for dimensionality reduction, enabling faster computations while retaining clustering quality.

The Jacobian matrix highlights the thematic overlap between clusters from both methods, while the word clouds provide qualitative insights into the representative themes of each cluster. The findings emphasize the importance of selecting appropriate clustering techniques and preprocessing steps for textual data analysis.

Figure 8: Word Cloud for K-Means Clusters

Figure 9: Word Cloud for Hierarchical Clustering Clusters