K-Means Clustering: Optimal Number of Clusters

Preprocessing Steps: The dataset was converted to lowercase and tokenized into words. Non-ASCII characters and punctuation were removed, and word stems were extracted while filtering out stop words. It was ranked by top 10,000 words frequency excluding those with a frequency lower than 100. Finally, a co-occurrence matrix was constructed with a window size of 2, capturing word relationships for clustering. Additionally, the co-occurrence matrix was normalized using the normalization to balance high- and low-frequency word relationships, ensuring consistent scaling and improving clustering accuracy.

Methodology: The K-means algorithm with KMeans++ initialization was applied to cluster high-frequency words extracted from a text-based dataset. A co-occurrence matrix, constructed with a window size of 15, was used to represent the relationships between words. Each word was treated as a data point, and clustering was performed to identify groups of semantically or contextually similar words. The range of K values tested was from 2 to 9. KMeans++ initialization was chosen to improve the convergence speed and clustering quality by selecting initial centroids that are well-separated. To ensure robust evaluation, K-Fold cross-validation was employed to compute the average Silhouette Score, its standard deviation, and the average Within-Cluster Sum of Squares (WCSS) for each K value. A boxplot analysis was conducted to facilitate the selection of the optimal K value, offering insights into clustering stability and performance.

Application of Occam's Razor: Occam's Razor was applied by selecting the simplest model (smallest K) that achieved satisfactory clustering results. This approach avoided overfitting while maintaining meaningful patterns in the data.

Results:

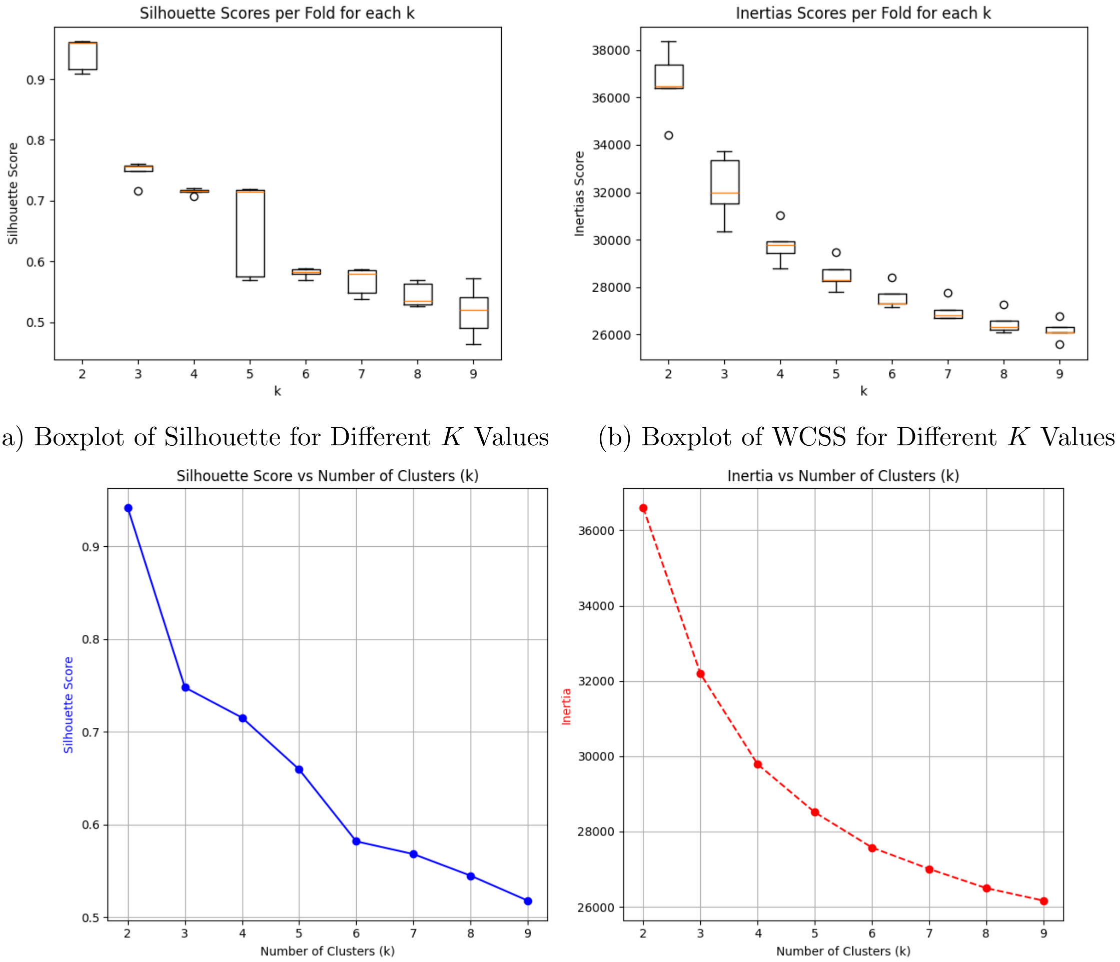
K Value	AVG WCSS	AVG Silhouette Score	STD
2	36591.92	0.9414	0.0235
3	32189.00	0.7478	0.0162
4	29787.08	0.7152	0.0048
5	28511.36	0.6593	0.0708
6	27577.06	0.5817	0.0068
7	27009.75	0.5680	0.0204
8	26498.59	0.5448	0.0180
9	26171.05	0.5172	0.033

Table 1: Clustering Performance for Different K Values (Including STD)

The optimal K value was determined to be 4, as it balanced performance and simplicity.

Analysis:

Selecting $K = 4$, Subplots (a) and (b) illustrate the distributions of Silhouette Scores and WCSS values for different k values under KFold cross-validation. The boxplots show that most folds achieve good intra-cluster cohesion and inter-cluster separation, with small variability in Silhouette Scores and WCSS values across folds for each k. Therefore, the use of average scores is a reliable representation of the overall Silhouette and WCSS coefficients.



(a) Boxplot of Silhouette for Different K Values

(b) Boxplot of WCSS for Different K Values

(c) WCSS and Silhouette Scores Across K Values

Figure 1: Clustering Performance Analysis

Observing subplot (c) for WCSS (i.e., Inertia), it is evident that as k increases, the intra-cluster error decreases overall. However, beyond $k = 4$, the reduction in WCSS becomes significantly smaller, indicating diminishing marginal returns. Furthermore, the Silhouette Score decreases gradually as k increases, but at $k = 3$ or $k = 4$, the scores remain relatively high and within an acceptable range.

Finally, subplot (c) combines the trends of both metrics as k changes, providing a clearer visualization that $k = 4$ offers a good balance. It avoids the coarse grouping observed with very few clusters (e.g., $k = 2$) and prevents the issues of low Silhouette Scores and increased complexity that arise with excessive cluster numbers. Following the principle of Occam's Razor, $k = 4$ represents a solution that is "simple enough while effectively capturing data differences." It neither over-complicates the model nor oversimplifies the clustering, making $k = 4$ an optimal choice for this clustering analysis.

References:

1. Zhi-Hua Zhou. *Machine Learning.* Chapter 2 and Chapter 11.
2. J. Xu et al. *A K-means Algorithm for Financial Market Risk Forecasting.* arXiv preprint arXiv:2405.13076, 2024.