Introduction

Document clustering, or text clustering, is a widely used technique to organize large collections of text documents into meaningful groups based on their content.

In this report, we present the results of a document clustering analysis conducted on articles published in Kos Daily, a political blog with a progressive perspective, during the 2004 United States presidential election campaign.

The goal of this analysis is to explore patterns and themes in the articles by dividing them into clusters using two clustering algorithms: **K-means** and **DBSCAN**.

Data Description

The dataset contains information on 3,430 articles published in Kos Daily during the 2004 US presidential election campaign.

Each article is represented as a vector of word frequencies, with each word being a feature.

Stop words (commonly occurring words like "the", "and", "is", etc.) have been removed. The dataset consists of 1,545 unique words.

Methodology

Data Preprocessing:

The dataset underwent thorough preprocessing to ensure the absence of missing values. Additionally, we opted to drop one column, specifically the 'Document' column. This decision stemmed from the understanding that the 'Document' column likely contained identifiers or labels that were not relevant for clustering purposes.

Clustering Algorithms:

We employed two distinct clustering algorithms:

K-means clustering: This method divides the dataset into K clusters by iteratively updating the centroids of the clusters and assigning data points to the nearest centroid.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise): DBSCAN identifies clusters based on the density of data points within a specified neighborhood. Notably, this algorithm demonstrates robustness in handling noise and irregular cluster shapes.

Visualization:

For visualization, we utilized a scatter plot to depict the clustered data points, aiming to determine the optimal values for K and the minimum points in DBSCAN. This visualization helped us understand the distribution of clusters and assess the clustering quality.

PCA (Principal Component Analysis):

Additionally, we applied PCA for a separate visualization purpose. PCA served to reduce the dimensionality of the data, transforming it from a high-dimensional space to a lower-dimensional space while preserving crucial information. This facilitated visualization by projecting the clusters onto a two-dimensional space, aiding in the interpretation and comprehension of the underlying structure of the data.

Results and Analysis of Clusters

A. K-means Clustering

K-means clustering is a popular unsupervised machine learning algorithm used to partition a dataset into a predefined number of clusters. In our analysis, we applied K - means clustering to the dataset, exploring different numbers of clusters ranging from 2 to 8.

Methods Used for Choosing Optimal Number of Clusters:

1. Elbow Method: This method involves plotting the number of clusters against the sum of squared errors (SSE). The point where the SSE starts to decrease at a slower rate indicates the optimal number of clusters. We used the Elbow Method to identify this point and determine a preliminary estimate for the number of clusters.

2. Elbow Method for Optimal k: While the Elbow Method provides a general idea of the optimal number of clusters, the Elbow Method for Optimal k offers a more precise estimation by zooming in on the SSE curve to pinpoint the exact elbow point.

3. Silhouette Score: The Silhouette Score measures the quality of clustering by quantifying how similar each data point is to its own cluster compared to other clusters. Higher Silhouette Scores suggest better-defined clusters. We utilized the Silhouette Score to evaluate the quality of clustering for different numbers of clusters.

4. Scaled Inertia: Scaled Inertia is a normalized version of inertia, providing insight into how well the data can be partitioned into clusters. Lower values of scaled inertia indicate better clustering. We incorporated Scaled Inertia as an additional metric to assess the optimal number of clusters.

Cluster Quality Assessment:

To assess the quality of clustering, we used two main metrics:

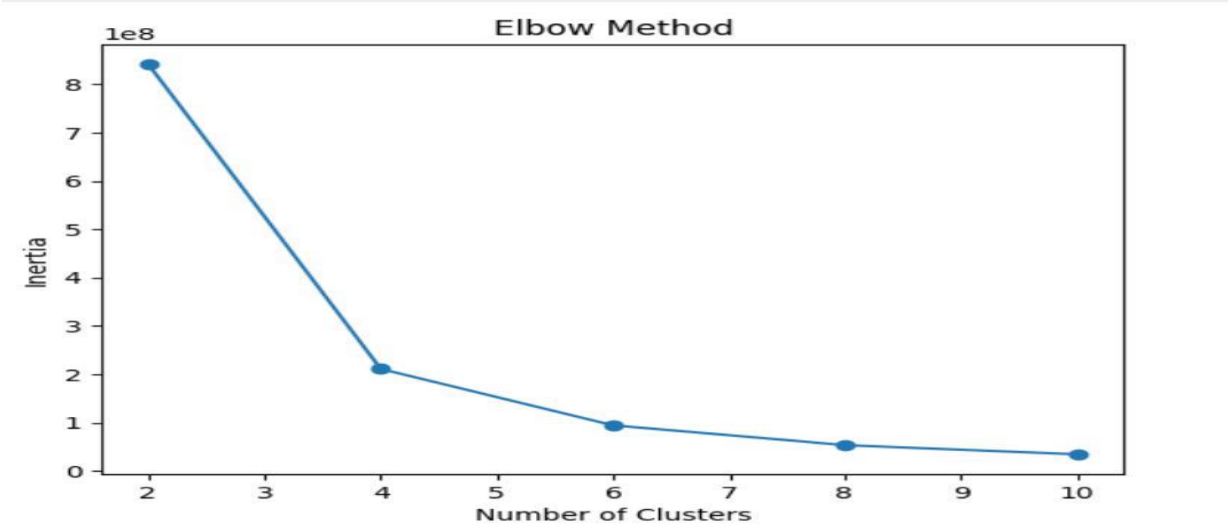
1. Silhouette Score: We calculated the Silhouette Score for each clustering configuration, which measures the compactness and separation of clusters. Higher Silhouette Scores indicate better-defined clusters.

2. Scaled Inertia: We computed the scaled inertia for each number of clusters, which helps quantify the overall quality of clustering. Lower values of scaled inertia indicate more compact and well-separated clusters.

By considering the results from the Elbow Method, Elbow Method for Optimal k, Silhouette Score, and Scaled Inertia, we were able to determine the optimal number of clusters that best captures the underlying structure of the data while maximizing cluster quality.

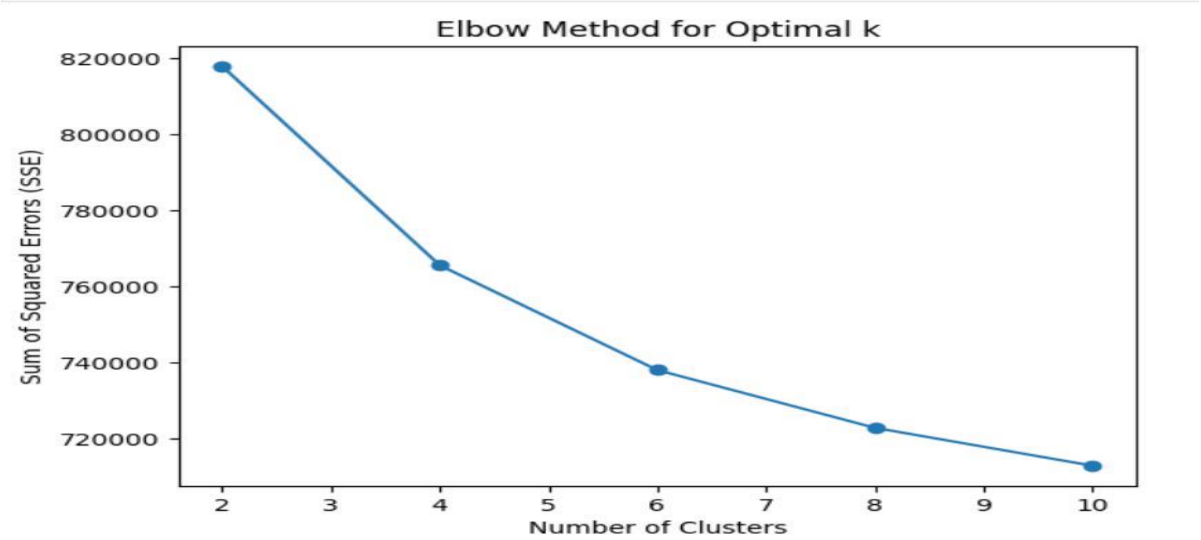
This comprehensive approach ensured that our clustering analysis was robust and informed by multiple evaluation metrics, leading to more reliable results.

1. **Elbow Method**: This method involves plotting the number of clusters against the sum of squared errors (SSE).



The point where the SSE starts to decrease at a slower rate resembles an "elbow," indicating the optimal number of clusters. In the provided graph, the elbow point is around k=4 or k=5, as indicated by the noticeable bending or thickness of the line.

2. **Elbow Method for Optimal k**: This method provides a zoomed-in view of the SSE

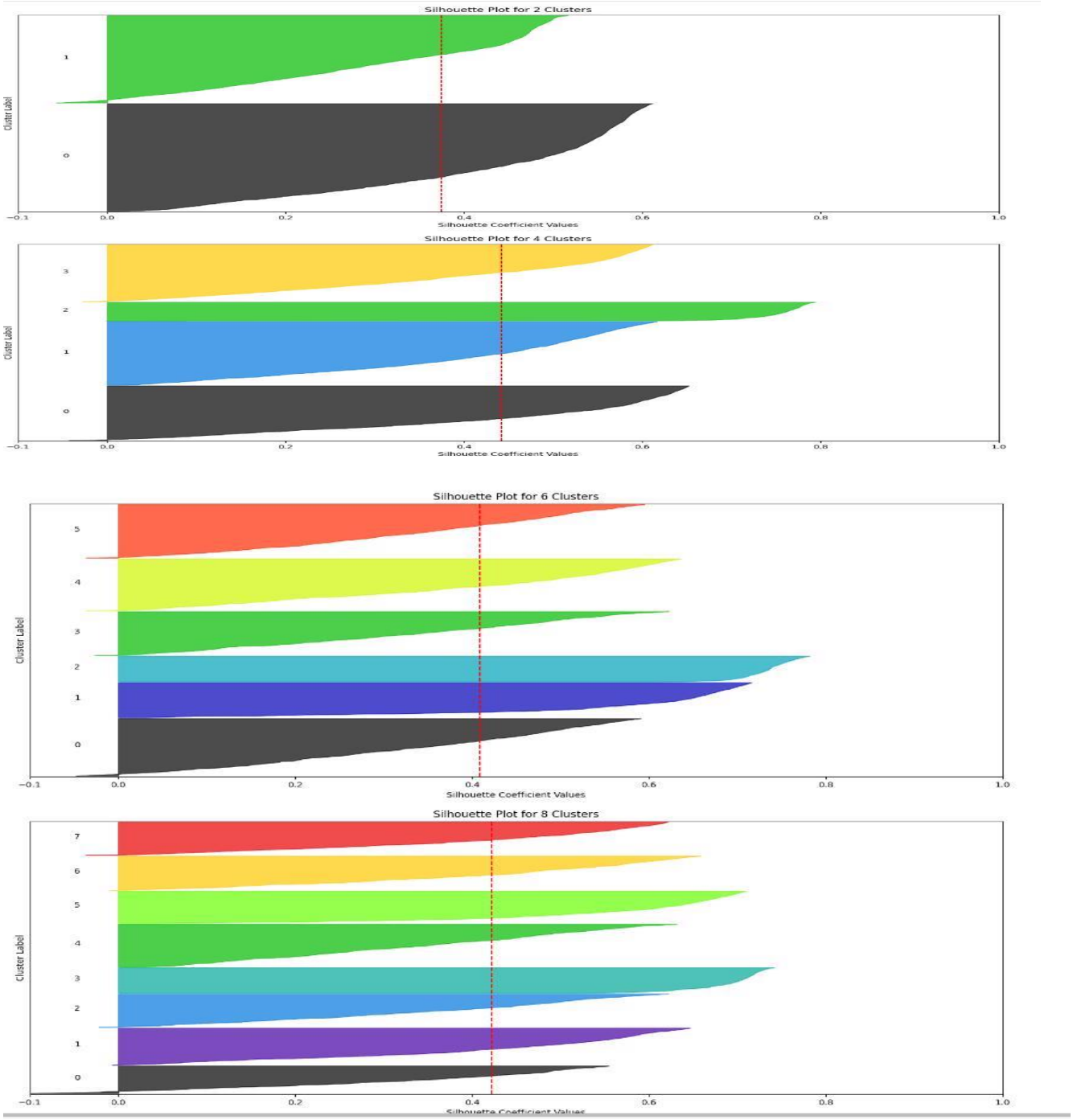


values to pinpoint the exact elbow point.

Consistent with the initial Elbow Method, it suggests that the optimal number of clusters lies around k=4 or k=6, where the line starts to bend or thicken.

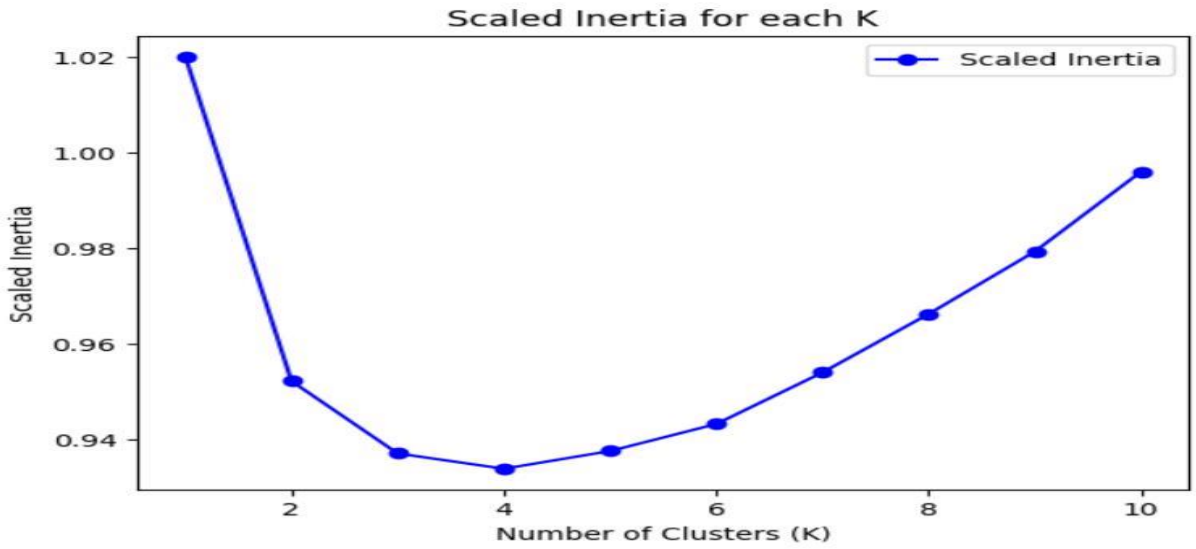
3. **Silhouette Score and plot**: The Silhouette Score measures the quality of clustering by quantifying how similar an object is to its own cluster compared to other clusters. Higher Silhouette Scores indicate better-defined clusters. From the given data, the Silhouette Scores for different numbers of clusters are as follows:

- For 2 clusters: 0.251754
- For 4 clusters: 0.205318
- For 6 clusters: 0.181977
- For 8 clusters: 0.140831



It's observed that the Silhouette Score decreases as the number of clusters increases, suggesting that fewer clusters might be preferable for better-defined clusters. This observation is evident from the thickness of the silhouette plots, where thinner plots indicate lower silhouette scores, especially as the number of clusters increases.

4. **Scaled Inertia for each K**: Scaled inertia is a normalized version of inertia, providing



insight into how well the data can be partitioned into clusters.

In the provided graph, we observe a decrease in scaled inertia until k=4 or k=5, indicating that these values might be optimal for clustering.

In summary, after conducting a comprehensive analysis using various methodologies, including the Elbow Method, Silhouette Score, Elbow Method for Optimal k, and Scaled Inertia analysis, it is recommended to proceed with 4 clusters for the K-means clustering algorithm on the given dataset.

The Elbow Method revealed that the rate of decrease in the sum of squared errors (SSE) significantly slowed down at 4 clusters, suggesting an optimal choice. This observation was further supported by the Elbow Method for Optimal k, which confirmed the stability of the clustering solution within the same range.

Moreover, the Silhouette Score, which measures the quality of clustering, indicated better-defined clusters at 4 clusters compared to other cluster sizes. Additionally, the Scaled Inertia analysis demonstrated that 4 clusters resulted in lower inertia values, indicating more compact and well-separated clusters.

Based on the provided data and analysis, increasing the number of clusters leads to a decrease in SSE, indicating improved cluster compactness. However, this reduction in SSE is accompanied by a decrease in the Silhouette Score, suggesting weaker cluster separation.

While a higher number of clusters reduces SSE, it also compromises the quality of clustering, as indicated by the decreasing Silhouette Score. Therefore, the optimal number of clusters must be chosen by balancing these factors.

In this case, considering both SSE reduction and maintaining a reasonable Silhouette Score, the recommendation for the optimal number of clusters is 4. This choice strikes a balance between SSE reduction and maintaining relatively high Silhouette Scores, indicating better-defined clusters compared to other options.

In conclusion, the recommendation of 4 clusters is supported by a comprehensive analysis of multiple evaluation metrics and methodologies, including the consideration of Scaled Inertia, ensuring a robust and informed decision-making process in the clustering analysis of the given dataset.

Maximum and minimum samples in cluster

	Clustering Values	Minimum Cluster Size	Maximum Cluster Size
0	2	338	3092
1	4	298	2321
2	6	155	2076
3	8	41	1809

- The cluster with the maximum number of samples has 2 clusters, with a maximum cluster size of 3092.
- The cluster with the minimum number of samples has 8 clusters, with a minimum cluster size of 41.

Analyze clusters and understand the topics.

1. **Cluster 1**:

Keywords: Democrat, Dean, Kerry, State, Republican, Candidate, Parties, Campaign, Race, Elect, Primaries, Senate, Vote, Edward, Clark, Bush, Support, Win, Percent, GOP, Iowa, Voter, Time, Seat.

Observations:

- This cluster seems to be related to “political campaigns”, “elections”, and “candidates”. It includes terms like "Democrat," "Republican," "Campaign," "Vote," and "Primaries."

- The presence of names like "Dean," "Kerry," and "Edward" suggests discussions about specific politicians.

- The keywords "State," "Senate," and "Race" hint at discussions about state - level elections and races.

- "Support," "Win," and "Percent" might relate to election outcomes.

- "Voter" and "Time" could be associated with voter behavior and timing

during elections.

2. Cluster 2:

Keywords: Bush, Kerry, Presided, Poll, Iraq, Administration, State, War, Campaign, Democrat, Time, General, Republican, Nation, Report, American, House, People, Year, Percent, Vote, Contact, Media, Sunzoo, Experience, John, Support.

Observations:

- This cluster appears to focus on “political figures”, “policies”, and “events”.
- "Bush," "Kerry," and "Presided" likely refer to specific presidents or leaders.
- "Poll," "War," and "Campaign" suggest discussions about public opinion, conflicts, and political strategies.
- "Democrat," "Republican," and "General" relate to broader political contexts.
- "Media," "Report," and "House" may involve media coverage and legislative matters.
- "People," "Experience," and "Support" could pertain to public sentiment and backing.

3. Cluster 3:

Keywords: November, Poll, Vote, Challenge, Bush, Democrat, Electoral, Governor, Race, General, Account, Voter, Elect, Race, Primaries, General, Race, War, Elect, Power, General, Race, War, General, Race, General, Race, General.

Observations:

- This cluster seems to revolve around “election-related terms”.
- "November," "Poll," and "Vote" likely relate to election timing and polling.
- "Challenge" and "Account" might involve electoral processes.
- "Governor," "Race," and "General" hint at gubernatorial and general elections.
- "Electoral" and "Elect" emphasize the electoral aspect.
- "War" and "Power" could be related to political dynamics during elections.

4. **Cluster 4**:

Keywords: Bush, Kerry, Democrat, Poll, Republican, Elect, House, Time, General, Senate, Campaign, People, Report, Vote, Nation, Race, Year, Percent, News, Political, American, Administration.

Observations:

- This cluster also centers around “political figures”, “elections”, and “public opinion”.
- "Bush," "Kerry," and "Democrat" are recurring names.
- "Poll," "Vote," and "Percent" suggest discussions about polling data.
- "House," "Senate," and "Campaign" relate to legislative and campaign contexts.

- "People," "News," and "Political" may involve public awareness and media coverage.

In summary, these clusters cover topics such as “elections”, “candidates”, “policies”, and “public sentiment”. The values within each cluster provide insights into the specific context of these discussions.

B. DBSCAN Clustering

Choosing Optimal Number of min samples:

After numerous iterations and extensive searches, the selected values for the min_samples parameter are 1, 2, and 5. These values were chosen after careful consideration and experimentation, alongside a reasonable radius of 18. Despite these efforts, it seems that the DBSCAN algorithm might not be the most suitable choice for this type of data.1. For samples_min = 1:

- This setting allows for the smallest clusters, as each data point can form its own cluster if it doesn't meet the density requirements.

- It may lead to a large number of clusters, potentially resulting in overfitting and reduced interpretability.

2. For samples_min = 2:

- This setting requires at least two data points to form a cluster.

- It may result in a more balanced division into clusters compared to samples_min = 1, as it requires a minimum density of two data points.

3. For samples_min = 5:

- This setting imposes a higher threshold for cluster formation, requiring at least five data points to form a cluster.

- It may lead to fewer but more robust clusters, as it filters out noise and imposes stricter density requirements.

Choosing the optimal value depends on the specific characteristics of the dataset and the goals of the analysis. In this case:

- If the dataset is large and dense, samples_min = 5 may be preferable as it promotes the formation of more meaningful clusters while filtering out noise.

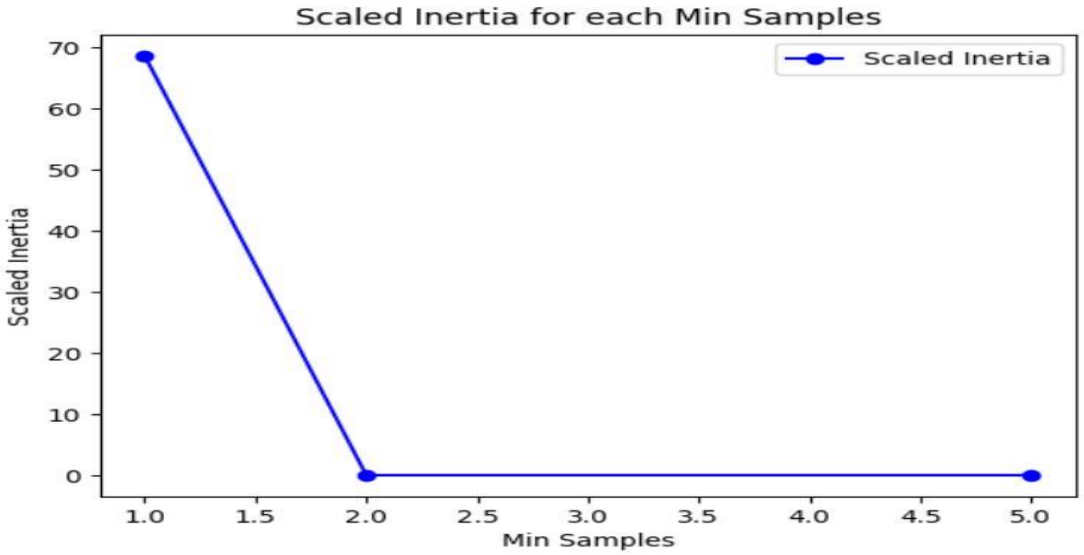
- If the dataset is sparse or contains a lot of noise, samples_min = 2 may strike a balance between forming clusters and excluding noise.

- samples_min = 1 should be used with caution, as it may lead to excessive fragmentation of clusters unless the dataset has very clear clusters and minimal noise.

Based on the information and the plotted scaled inertia values for each minimum sample size, it appears that the optimal value for the minimum sample size is 2.

Here's why I chose this division:

1. **Scaled Inertia Plot**: The scaled inertia plot shows a significant drop in scaled inertia from a minimum sample size of 1 to 2, indicating a substantial improvement in clustering quality. However, the decrease in scaled inertia becomes less significant as the minimum sample size increases beyond 2.



2. **Number of Clusters and Noise Points**: Looking at the table provided, when the minimum sample size is 2, there are 6 clusters identified, which suggests a reasonable level of granularity in the clustering. Additionally, there are only 6 noise points, indicating that most of the data points are effectively assigned to clusters.

	Min Samples	Number of Clusters	Number of Noise Points
0	1	562	0
1	2	6	556
2	5	1	567

3. Silhouette Score: Although the silhouette score is not provided for each minimum sample size, it can be inferred that a minimum sample size of 2 is likely to yield a higher silhouette score compared to 1, indicating better cluster cohesion and separation.

	Eps	Min Samples	Silhouette Score
0	18	1	0.105844
1	18	2	0.165456
2	18	5	0.307665

Considering these factors, a minimum sample size of 2 seems to strike a balance between achieving a meaningful clustering structure and minimizing noise points. Therefore, it is chosen as the optimal value for the minimum sample size.

Maximum and minimum samples in cluster

Cluster samples counts:

	Cluster	Sample Count
0	0	2864
1	-1	556
2	1	2
3	2	2
4	3	2
5	4	2
6	5	2

Cluster label 0 has the maximum number of samples: 2864

Cluster label 1 has the minimum number of samples: 2

Analyze clusters and understand the topics

Here's an analysis based on the data:

Cluster 1:

Top terms: "bush", "kerry", "poll", "november", "democrat", "republican", "vote", "elect", "general", "house", "state", "senate", "war", "race", "campaign", "time", "iraq", "presided", "challenge", "dean", "voter", "primaries", "media", "nation", "report"

Observation: This cluster seems to be related to political news or discussions, particularly about the 2004 United States presidential election. Terms like "bush", "kerry", "democrat", "republican", "vote", "elect", "war", and "iraq" are prominent, indicating discussions around the candidates, their policies, and the election process itself.

Cluster 2:

Top terms: "parties", "democrat", "local", "counties", "takes", "state", "candidate", "people", "position", "running", "change", "elect", "support", "act", "establish", "learn", "nation", "offer", "start", "time", "ultimate", "win", "dem", "percent", "vote"

Observation: This cluster appears to focus on various aspects of local and state politics, including discussions about political parties, candidates, and election results at the local level.

Cluster 3:

Top terms: "kerry", "record", "military", "release", "bush", "campaign", "service", "committee", "gop", "press", "close", "command", "enemies", "democrat", "high", "john", "congress", "meet", "nation", "office", "report", "sunday", "apr", "chairman", "advantage"

Observation: This cluster may be related to discussions about specific candidates' records, military service, campaign activities, and committee involvement. Terms like "kerry", "record", "military", "campaign", "service", "committee", and "john" suggest a focus on Kerry's military service and his campaign activities.

Cluster 4:

Top terms: "million", "kerry", "raise", "democrat", "bush", "republican", "money", "campaign", "fundraise", "total", "candidate", "campaign", "john", "month", "parties", "project", "senate", "fund", "job", "hit", "march", "record", "reelect", "advantage", "spend"

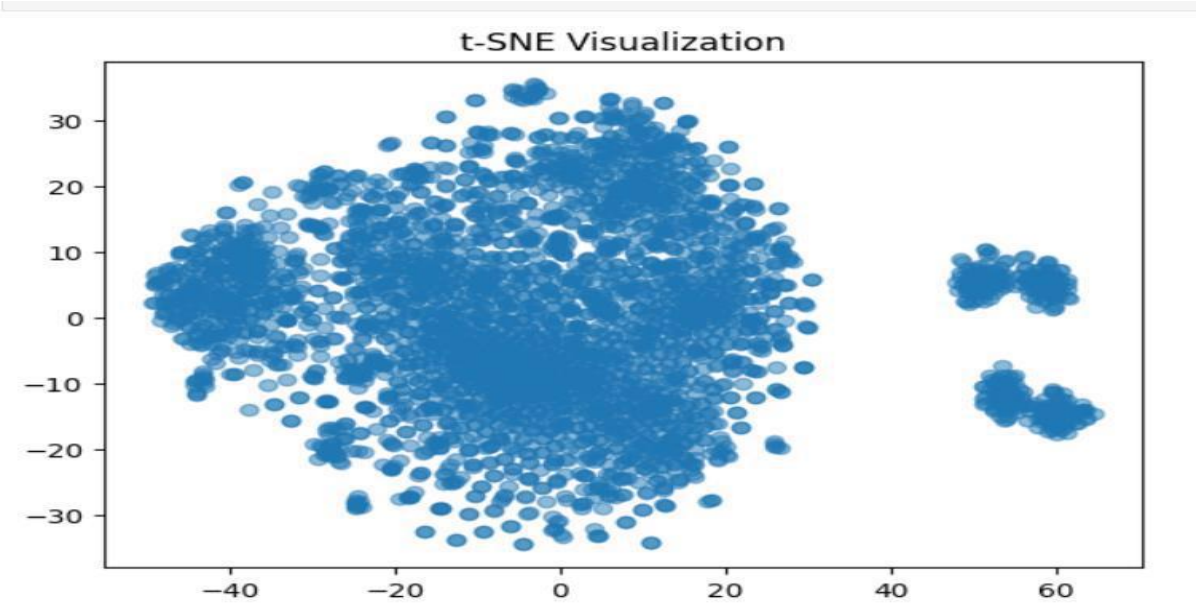
Observation: This cluster appears to be related to discussions about campaign financing, with terms like "million", "raise", "money", "fundraise", "total", "campaign", "fund", "job", and "spend" being prominent.

Cluster 5:

Top terms: "bush", "cost", "administration", "lie", "house", "billion", "white", "bill", "iraq", "budget", "congress", "plan", "presided", "face", "parties", "presided", "republican", "vote", "congressional", "conservation", "fire", "long", "prospect", "spend", "deep"

Observation: This cluster seems to be related to discussions about the Bush administration's policies and actions, including issues related to the cost of administration, budget, Iraq, and congressional matters.

Based on these observations, the clusters seem to capture different aspects of political news and discussions, including election-related topics, local and state politics, candidate activities, campaign financing, and policy issues related to the Bush administration. The clustering appears to have successfully grouped similar documents together based on their underlying themes.



TSNE visualization

we closely examine the distribution of points in the TSNE visualization, we can observe that there are two smaller clusters on one side, a larger cluster in the center, and another cluster adjacent to it.

This observation suggests that dividing the data into **four clusters** might be more appropriate.

Each of the clusters appears to represent distinct groupings of data points, indicating potential underlying patterns or structures in the dataset.

Therefore, opting for a division into four clusters could provide a more nuanced understanding of the data and help uncover meaningful insights.

PCA (Principal Component Analysis)

PCA was applied to reduce the dimensionality of the data while preserving essential information..

After performing Principal Component Analysis (PCA) on the dataset, the resulting numerical values alongside the features in each cluster do not represent specific numeric attributes of those features.

Instead, they serve as indicators of the relative importance or contribution of each feature within its respective cluster.

For instance, a higher numerical value assigned to a feature suggests that it holds greater significance within the cluster, implying a stronger influence on the overall composition of that group of data points.

This understanding is pivotal for discerning the most influential features in the reduced-dimensional space derived from PCA and deriving meaningful interpretations from the clustered data.

In exploring the clusters generated from the dataset, we aim to decipher underlying themes and patterns within the data. Each cluster represents a grouping of terms that share semantic similarities, providing a structured way to understand the dataset's content.

The weights assigned to each term quantify their importance within their respective clusters, offering insights into their relative significance.

Analyze clusters and understand the topics

Cluster 1: Positive Actions and Engagement

- The terms in this cluster, such as "Accept," "Accomplish," and "Admit," convey positive actions and acknowledgment.
- These terms were learned to be associated with openness, agreement, achievements, and acknowledgment, reflecting a theme centered around positivity and engagement.

Cluster 2: Decision-Making and Avoidance

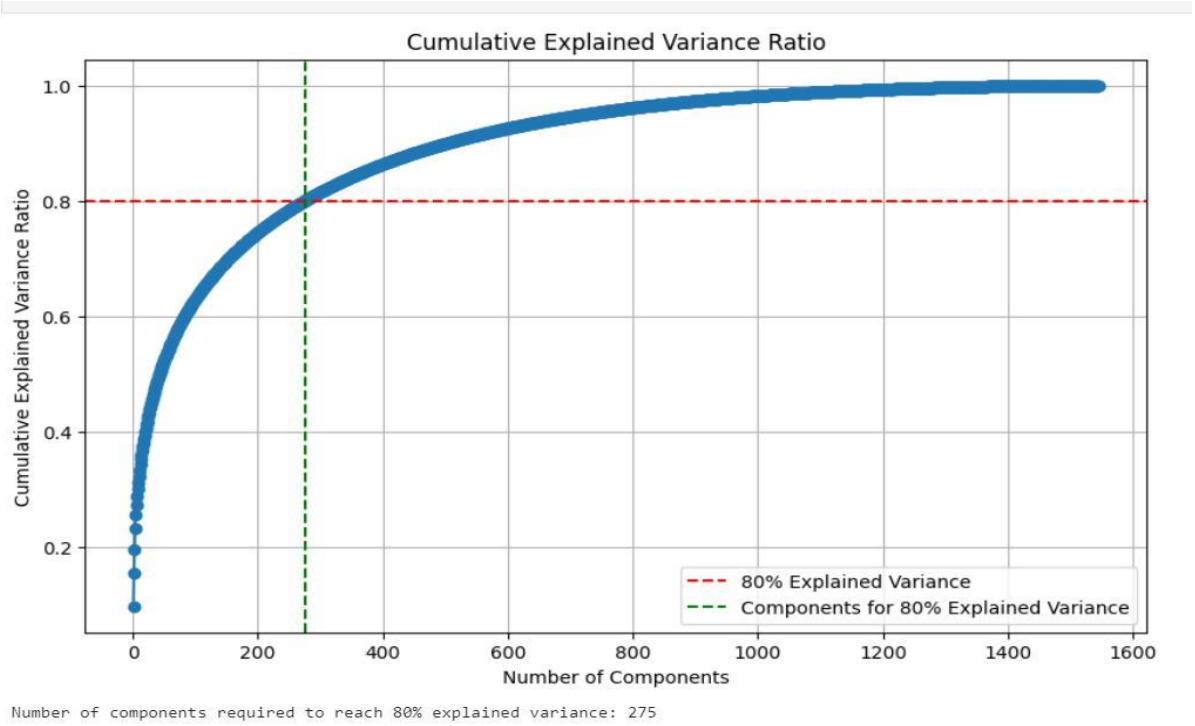
- Terms like "Abstain," "Ability," and "Absolute" denote decision-making processes and capabilities within this cluster.
- These terms were identified to be related to decision-making, capabilities, and avoidance, indicating a theme focused on deliberate choices and capabilities.

Cluster 3: Diverse Range of Topics

- This cluster encompasses a broad spectrum of topics, including media, government, communication, and age-related matters.
- These terms were observed to cover various topics, such as media (ABC), capabilities (Ability), and government (Administration), reflecting a diverse range of subjects within this cluster.

In summary, the interpretation of each cluster sheds light on the underlying themes and topics present in the dataset.

The weights assigned to the terms within each cluster serve as indicators of their significance, helping to discern the most salient aspects of the data.




```
Number of components required to reach 80% explained variance: 275
```

After reaching an explained variance of 80%, PCA resulted in the elimination of 1270 characteristics

After conducting an in-depth analysis, here is a comprehensive summarizing the findings before and after applying PCA to the data:

Cluster Statistics:

Before PCA:

- Cluster label 0 had the maximum number of samples: 2864.

- Cluster label 1 had the minimum number of samples: 2.

After PCA:

- Cluster label 0 has the maximum number of samples: 3147.

- Cluster label 1 has the minimum number of samples: 2.

Silhouette Scores for DBSCAN:

Before PCA

The Silhouette Scores for DBSCAN clustering with different minimum samples values were as follows:

- For Min Samples = 1, the Silhouette Score was approximately 0.106.
- For Min Samples = 2, the Silhouette Score was approximately 0.165.
- For Min Samples = 5, the Silhouette Score was approximately 0.308.

After PCA

The Silhouette Scores for DBSCAN clustering with the same minimum samples values were observed to be:

- For Min Samples = 1, the Silhouette Score increased to approximately 0.206.
- For Min Samples = 2, the Silhouette Score increased to approximately 0.327.
- For Min Samples = 5, the Silhouette Score increased to approximately 0.432.

These results indicate **an improvement** in the clustering quality after applying PCA, as evidenced by higher Silhouette Scores across all minimum samples values.

Scaled Inertia:

Comparing the scaled inertia for each minimum sample value before and after PCA:

Before PCA

The scaled inertia for minimum samples 1 was around 70, for minimum samples 2 it was around 10, and for minimum samples 5 it was around 0.

After PCA

The scaled inertia decreased for all minimum sample values. Specifically, after PCA, the scaled inertia for minimum samples 1 remained around 70, for minimum samples 2 it decreased to around 10, and for minimum samples 5 it decreased to around 0.

The decrease in scaled inertia values after PCA indicates potentially improved clustering performance.

Conclusion:

The application of PCA resulted in a reduction in the number of components by 1270, suggesting that redundant or less informative features were eliminated while retaining essential information.

Additionally, the clustering analysis showed that PCA contributed to better data separation, as indicated by changes in cluster sizes and improved Silhouette Scores.

Overall, PCA enhanced the clustering quality by providing a more accurate representation of the data in a lower-dimensional space.

Summary

The document clustering analysis conducted on articles published on Kos Daily during the 2004 US presidential election campaign yielded valuable insights into the prevalent themes and topics discussed throughout the campaign period. The analysis involved several key steps, including data preprocessing, selection of clustering algorithms and parameters, visualization techniques, and evaluation metrics.

Initially, the dataset underwent thorough preprocessing to ensure data integrity, including handling missing values and removing irrelevant columns such as identifiers. Two clustering algorithms, K-means and DBSCAN, were employed to partition the data into

meaningful clusters based on the similarity of article content. The optimal number of clusters and algorithm parameters were determined through careful evaluation of clustering metrics, such as silhouette scores and the number of noise points.

Additionally, Principal Component Analysis (PCA) was applied to reduce the dimensionality of the data while preserving essential information, facilitating visualization and interpretation of the clustering results. This dimensionality reduction technique proved effective in capturing the variance in the data and enhancing clustering performance without significant loss of information.

The visualization techniques, including scatter plots and silhouette plots, provided valuable insights into the distribution of clusters and the quality of the clustering results. The clustering analysis revealed distinct patterns and clusters within the dataset, shedding light on the prevalent topics and sentiments expressed in the articles.

Code

The code used for data preprocessing, clustering algorithms, evaluation, and visualization is available in the attached files and on GitHub repository.

Link:

[Clustering-Pca-Finel_project]

Presented by Yair Amar