Introduction

The project delves into clustering algorithms applied to articles from Kos Daily, a progressive American political blog. Our focus will be on clustering articles from Kos Daily, an American political blog known for its progressive perspective in news and opinion pieces.

The dataset (MB10.1 - CSV (dailykos)) contains information on 3,430 articles or blogs published in Kos Daily. These articles date back to 2004, coinciding with the United States presidential election, where the main contenders were incumbent President George W. Bush (Republican) and John Kerry (Democrat). Foreign policy, particularly the Iraq invasion of 2003, played a significant role in shaping the election discourse.

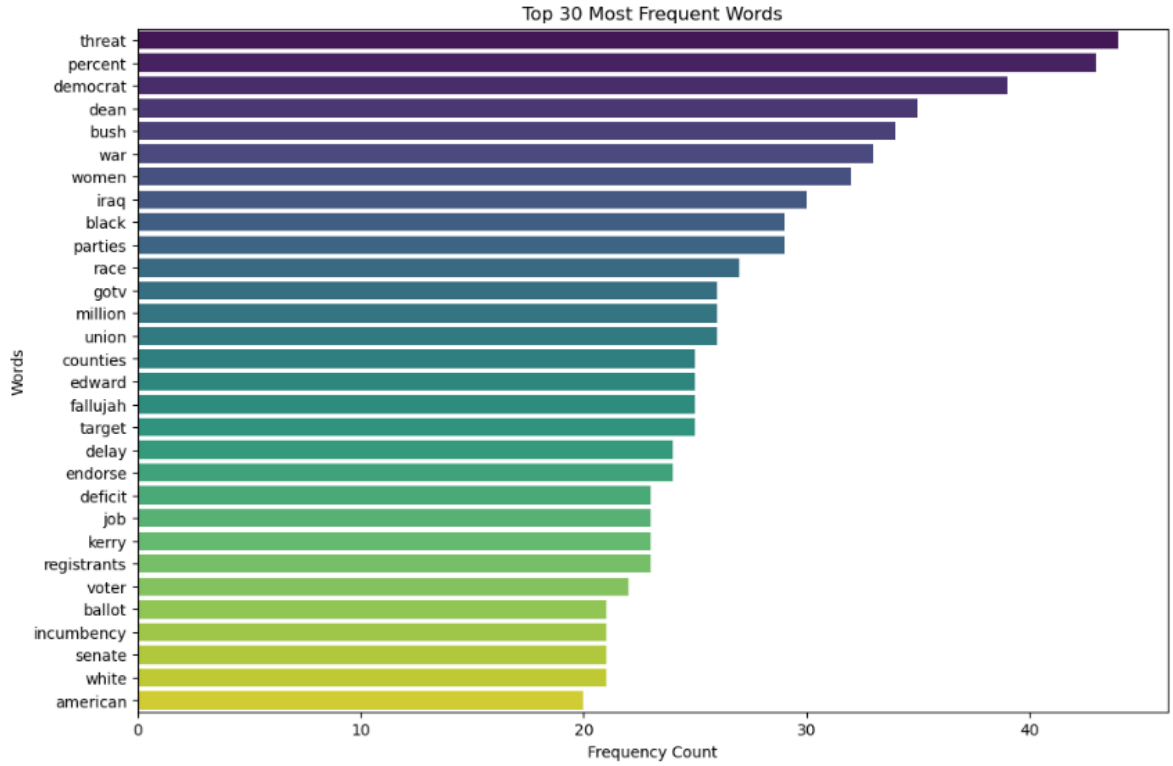
Each variable in the dataset represents a word that appeared in at least 50 different articles, totaling 1,545 unique words. The dataset has undergone preprocessing steps typical in text analysis, such as removing punctuation and stop words. For each document, the variable values indicate the frequency of each word's appearance in that document.

Methods and Techniques:

In the project I explores clustering techniques including DBSCAN and K-means, implemented from the sklearn library with default parameter settings adjusted to fit the dataset. For the algorithms metrics such as silhouette score, sum of squared errors (SSE) and Elbow Scaled Inertia were used to determine the optimal cluster count based on cluster quality. A PCA algorithm was performed after K-MEANS was performed in order to examine how it affects the clustering, whether the PCA is useful or not, This study aims to assess how PCA affects clustering quality compared to clustering without PCA, focusing on the achieved explained variance and clustering performance. In the project I employed both the t-SNE (t-distributed Stochastic Neighbor Embedding) and MDS (Multi-Dimensional Scaling) algorithms to visualize data points on a scatter plot graph.

Methodology

1. **Pre-Processing: Showing Dominant Words:**
   Analyzed and displayed dominant words within the dataset before proceeding with clustering algorithms.



On the basis of this barplot, it is difficult to determine what is the main topic in the documents and what is special between one group of documents to another, We do see that the dominant words are: the war in Iraq, Bush, democrat, women and the most prominent word - threat.

2. **Clustering Algorithms Implementation:**
Utilized sklearn library for implementing DBSCAN and K-means clustering algorithms.
Ensured appropriate default parameter settings for the algorithms based on the dataset and task requirements.

K-means algorithm Clustering Analysis:
The K-means algorithm divided the data into clusters, evaluating different cluster counts (2, 4, 6, 8) using metrics such as silhouette score and sum of squared errors (SSE) to determine the optimal cluster count based on clustering quality.
Metrics Overview:
1. SSE (Sum of Squared Errors):
- Measures squared distances between data points and their cluster centroids.
- It provides insights into how tightly grouped the data points are within their respective clusters; lower SSE implies denser clusters.

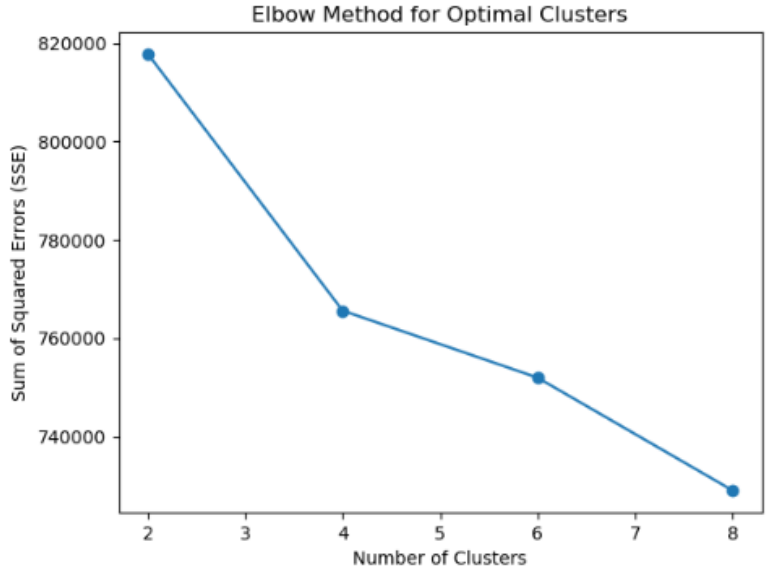
2. Silhouette Score:
- Evaluates cluster separation quality.
- Ranges from -1 to 1; higher values indicate better-defined clusters.
- Scores near +1 show strong separation, 0 indicates boundary proximity, and negatives suggest potential misclassifications.

3. Elbow Scaled Inertia:
- Helps identify optimal clusters by analyzing scaled inertia reduction rates.
- An "elbow" in the inertia graph indicates potential optimal cluster count.
- Balances model complexity and clustering effectiveness visually.

Inertia is the sum of squared distances of all points relative to their cluster centroids.

The SSE results and silhouette for the K-MEANS algorithm:

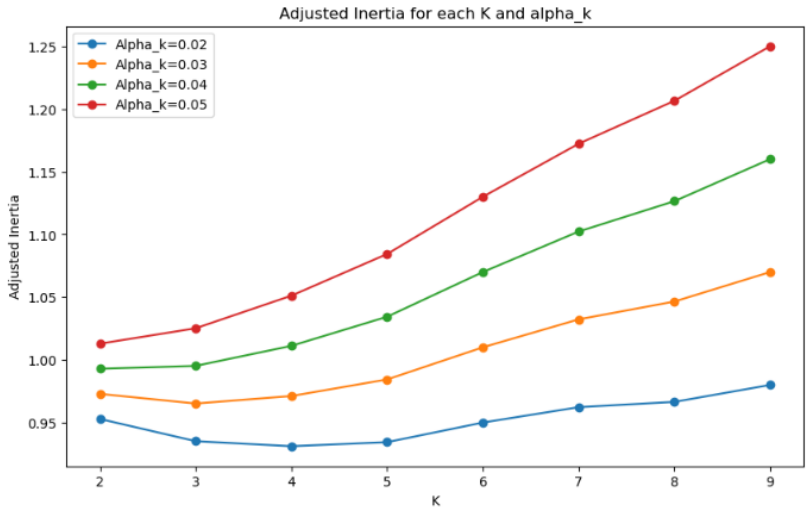
```
Clusters: 2, SSE: 817799.8968688044, Silhouette Score: 0.25231004882556324
Clusters: 4, SSE: 765583.4291995232, Silhouette Score: 0.21161748268224742
Clusters: 6, SSE: 751965.8666169838, Silhouette Score: 0.195011774044343 3
Clusters: 8, SSE: 729061.8162494708, Silhouette Score: 0.18235301530666687
```



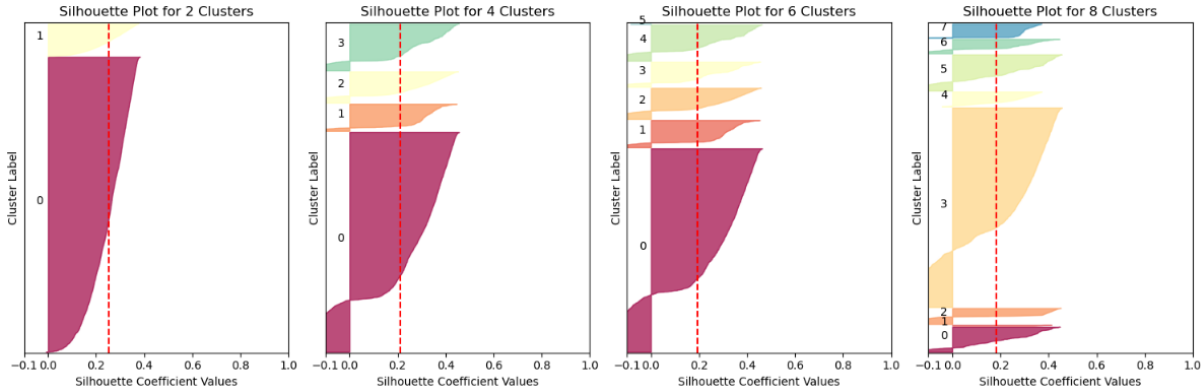
It can be seen that the lowest SSE is for division into 8, and the highest silhouette is for division into 2, from looking at the SSE and silhouette scores it seems that the optimal division is between 4 and 6.

Based on the graph and the Elbow method, it seems that division into 4 is optimal (Elbow Point) - between 2 and 4 clusters there is a sharp decrease and from 4 clusters to 8 clusters there is a slower decrease, but it is not possible to decide unequivocally, so we will check the Adjusted Inertia.

```
Best k for alpha_k=0.02: 4
Best k for alpha_k=0.03: 3
Best k for alpha_k=0.04: 2
Best k for alpha_k=0.05: 2
```



We can see that when K=4 - this is the minimum point among options 2, 4, 6 or 8
And the silhouette:



The broader the color, the larger the cluster size it represents, indicating a bigger group. Silhouette scores less than 0 indicate overlap with another cluster. When a silhouette plot crosses the red dashed line, it indicates the extent of separation from other clusters.
Also according to the silhouette we can say that 4 clusters will be ideal.

<u>Implemented the K-means algorithm with k=4 to create the desired cluster division.</u>
Introducing The cluster that has the maximum number of samples and the cluster that has the minimum number

```
Cluster with max examples: 2 ,Count: 2188
Cluster with min examples: 0 ,Count: 304
```

The analysis identified clusters with varying sample sizes, with one cluster having the maximum number of samples and another the minimum, indicating distinct patterns within the data. Separate CSV files were created for each cluster to facilitate in-depth analysis and topic interpretation based on cluster content.

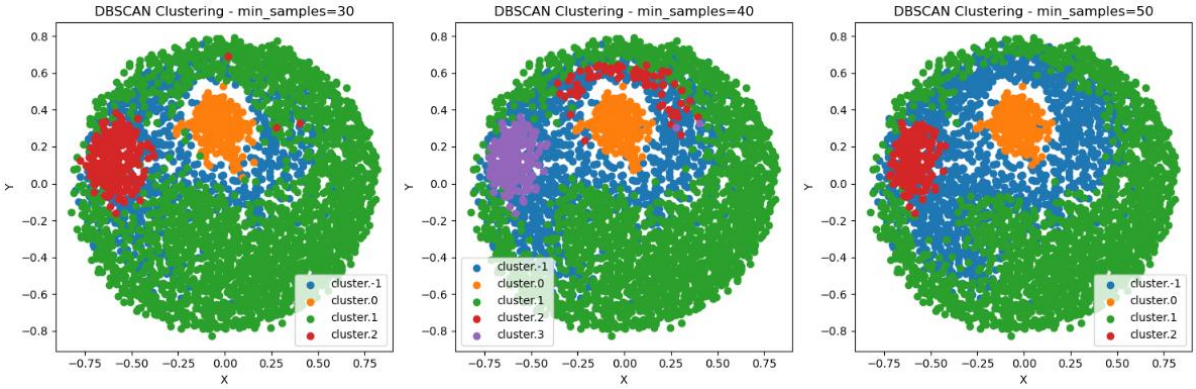
```
Cluster 0 data exported to cluster_ 0 .csv
Cluster 1 data exported to cluster_ 1 .csv
Cluster 2 data exported to cluster_ 2 .csv
Cluster 3 data exported to cluster_ 3 .csv
```

DBSCAN algorithm Clustering Analysis:

The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm is a density-based clustering method used to group data points based on their density distribution in a given space. Unlike traditional centroid-based clustering algorithms like K-means, DBSCAN does not require specifying the number of clusters beforehand and can identify clusters of arbitrary shapes.
For the DBSCAN algorithm, varying `min_samples` values were tested to assess their impact on clustering quality, ultimately selecting the optimal `min_samples` value based on clustering evaluation criteria.
Through analysis of silhouette scores and SSE, we identified the optimal `min_samples` value that resulted in well-defined clusters with meaningful separation and compactness.

```
min_samples=30: Silhouette Score = 0.1301675763310476, SSE = 65304.98160641705
min_samples=40: Silhouette Score = 0.08886439247019853, SSE = 63684.8628578615
min_samples=50: Silhouette Score = 0.15218247323377584, SSE = 63553.93879991758
```

Plotted the clusters derived from DBSCAN on the MDS network to visualize their distribution and relationships:



The visualization of DBSCAN clustering on the MDS network provides insights into the spatial arrangement of clusters in a lower-dimensional space, allowing us to observe:
Cluster Separation: Clusters are visually separated, indicating distinct groupings based on content similarities within the Kos Daily articles.
Cluster Proximity: Proximity of clusters on the MDS network reflects similarities or thematic relationships between articles within clusters.
You can see from the graphs and the silhouette value that min_samples=40 returns the highest silhouette value and the smallest SSE and divides the data into 4 clusters, but there are many outliers

<u>Therefore I chose: min_samples=40 (4 clusters)</u>
Introducing The cluster that has the maximum number of samples and the cluster that has the minimum number

```
Cluster with max examples: 1 ,Count: 1933
Cluster with min examples: 2 ,Count: 68
```

The analysis identified clusters with varying sample sizes, with one cluster having the maximum number of samples and another the minimum, indicating distinct patterns within the data. Separate CSV files were created for each cluster to facilitate in-depth analysis and topic interpretation based on cluster content.

```
Cluster 0 data exported to cluster_DBSCAN_ 0 .csv
Cluster 1 data exported to cluster_DBSCAN_ 1 .csv
Cluster 2 data exported to cluster_DBSCAN_ 2 .csv
Cluster 3 data exported to cluster_DBSCAN_ 3 .csv
```

3. **Dimensionality Reduction:**

In the analysis of the Kos Daily dataset using the K-means clustering algorithm, I integrated the PCA (Principal Component Analysis) technique to reduce dimensionality while retaining essential information. This study aims to assess how PCA affects clustering quality compared to clustering without PCA, focusing on the achieved explained variance and clustering performance.

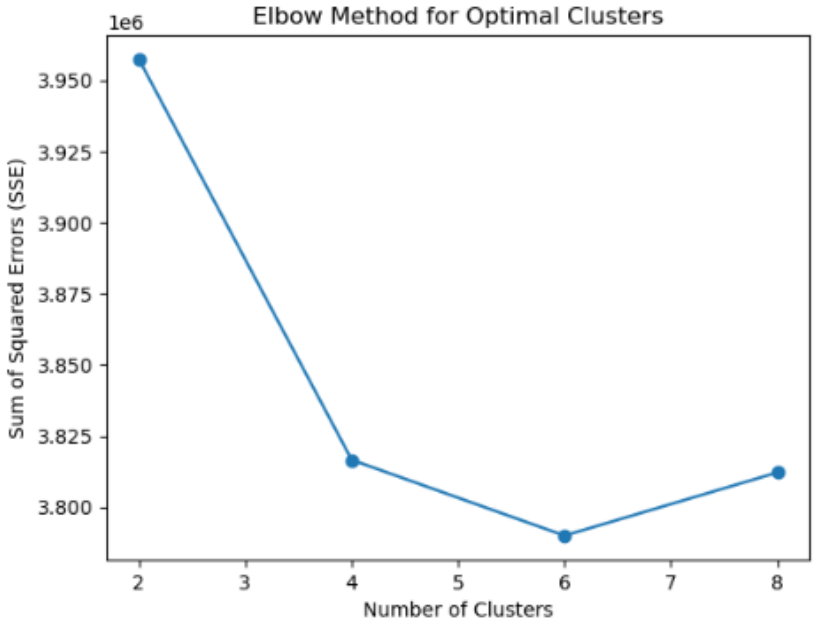
PrincipalDF

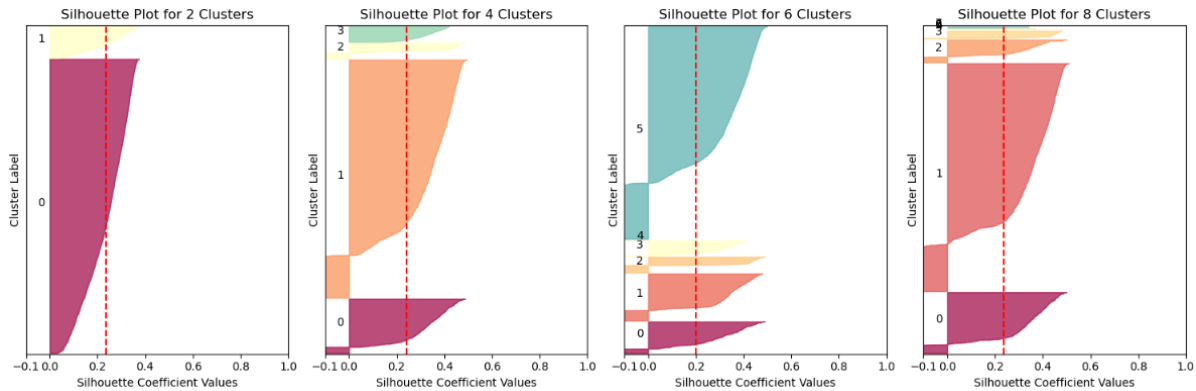
	Principal Component 0	Principal Component 1	Principal Component 2	Principal Component 3	Principal Component 4	Principal Component 5	Principal Component 6	Principal Component 7	Principal Component 8	Principal Component 9	...
0	-2.952117	0.533620	-0.080091	0.123273	-1.450242	0.888589	-3.749650	-1.206127	2.601633	-0.373846	...
1	-3.392105	-2.831885	-0.236883	-0.293806	-2.257524	-0.992251	-1.007744	0.323338	-1.540891	-1.506924	...
2	23.519528	-4.994131	-17.507357	-11.191110	0.214823	-12.132932	-1.644152	-0.537791	-0.024939	1.174827	...
3	-2.011269	1.448427	-0.807848	-0.064569	0.118439	0.541330	-0.786769	-1.528725	-1.883282	-2.052580	...
4	-2.616584	2.422425	-1.164630	-1.287248	2.141052	0.318815	1.537912	1.228598	3.334900	-0.540625	...
...	...	...	...	...	...	...	...	...	...	...	...
3425	-2.782257	-3.905839	-0.520214	0.225470	0.211909	-0.373975	0.455298	0.462642	1.319445	-0.887737	...
3426	-3.284308	-5.505195	-0.241980	-0.041105	-0.772550	-0.290369	0.333589	-0.282766	0.541593	0.185417	...
3427	-3.065183	-1.938143	0.014906	0.024544	-0.891824	-0.178110	1.245538	-1.270409	-2.247356	-0.317249	...
3428	-3.435958	-3.577170	-0.132683	-0.345534	0.695065	0.500796	-0.443959	-1.150099	-1.245223	2.379888	...
3429	-2.854465	-1.779311	-0.096770	-0.331454	-0.335383	0.184510	-0.999053	-0.213443	0.163473	2.438347	...

3430 rows × 579 columns

Applied PCA algorithm to reduce the dataset's dimensionality while retaining 80% of the explained variance. And extracted principal components to represent the data in a lower-dimensional space.
Conducted K-means clustering both with and without PCA-transformed data. And evaluated clustering quality metrics such as silhouette scores and cluster separations for both scenarios.

```
Clusters: 2, SSE: 3957345.3106166106, Silhouette Score: 0.2365940966156151
Clusters: 4, SSE: 3816539.94781942, Silhouette Score: 0.24338896214469263
Clusters: 6, SSE: 3789924.7389656184, Silhouette Score: 0.2017796916906539
Clusters: 8, SSE: 3812113.469780933, Silhouette Score: 0.23811663037439393
```





Out of the following options 2, 4, 6, 8 is the most suitable based on SEE and this silhouette is divided into 6 clusters.

Introducing The cluster that has the maximum number of samples and the cluster that has the minimum number

```
Cluster with max examples: 1 ,Count: 2429
Cluster with min examples: 5 ,Count: 79
```

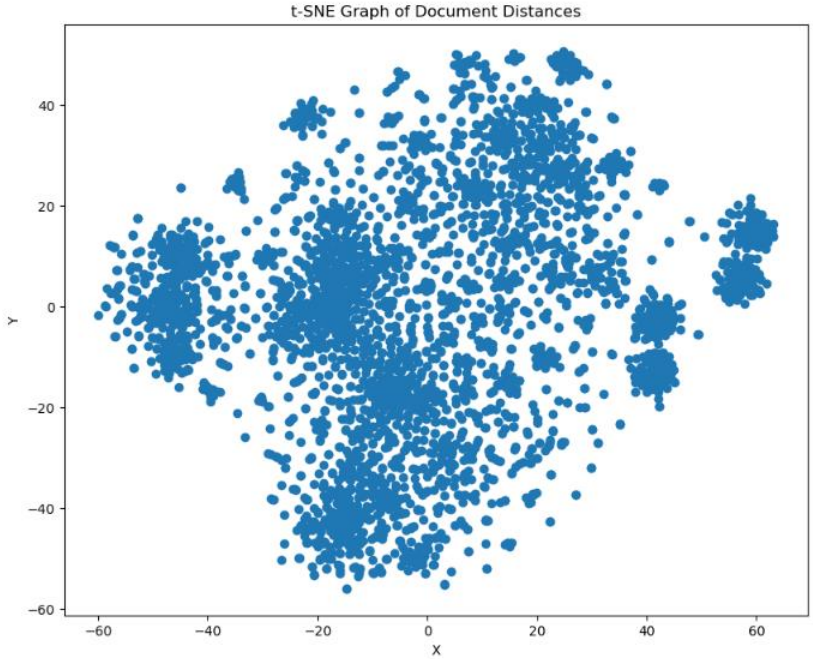
The analysis identified clusters with varying sample sizes, with one cluster having the maximum number of samples and another the minimum, indicating distinct patterns within the data. Separate CSV files were created for each cluster to facilitate in-depth analysis and topic interpretation based on cluster content.

```
Cluster 0 data exported to cluster_ 0 .csv
Cluster 1 data exported to cluster_ 1 .csv
Cluster 2 data exported to cluster_ 2 .csv
Cluster 3 data exported to cluster_ 3 .csv
Cluster 4 data exported to cluster_ 4 .csv
Cluster 5 data exported to cluster_ 5 .csv
```

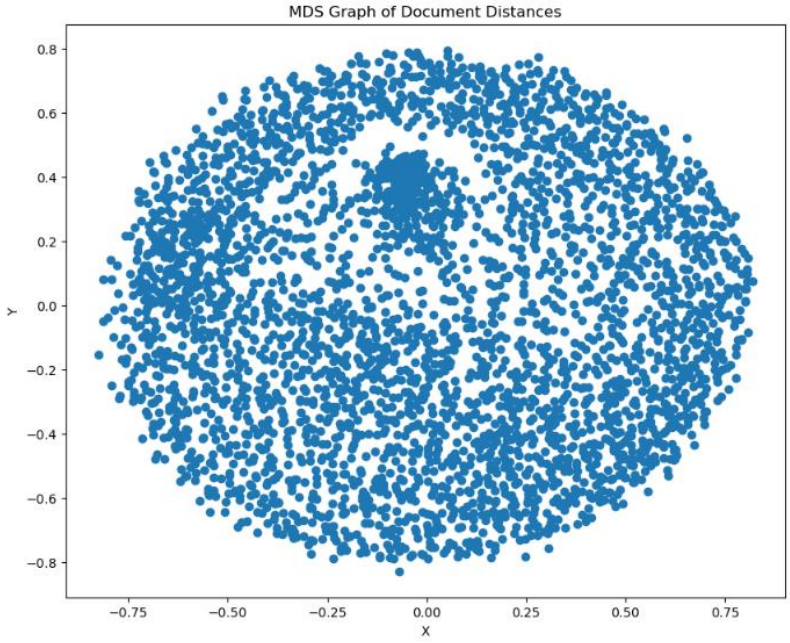
By evaluating the impact of PCA on K-means clustering, I gain a nuanced understanding of how dimensionality reduction techniques influence clustering outcomes. This analysis aids in making informed decisions regarding the appropriate balance between dimensionality reduction and clustering accuracy based on the dataset's characteristics and analysis goals.

4. **Visualization:**

A graphical depiction of the scatter of each data point representing an article or blog from Kos Daily was performed. The locations of points on the scatterplot were determined based on dimensionally reduced representations derived from t-SNE or MDS, allowing visualization of clustering patterns, data structures, and similarity/dissimilarity between data points. These simulations helped in gaining insights into the relationships and the basic structures that exist in the data set, and helped in further analysis and interpretation of the data.

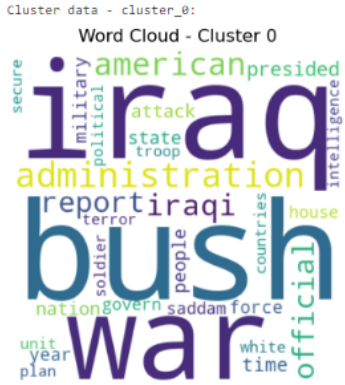


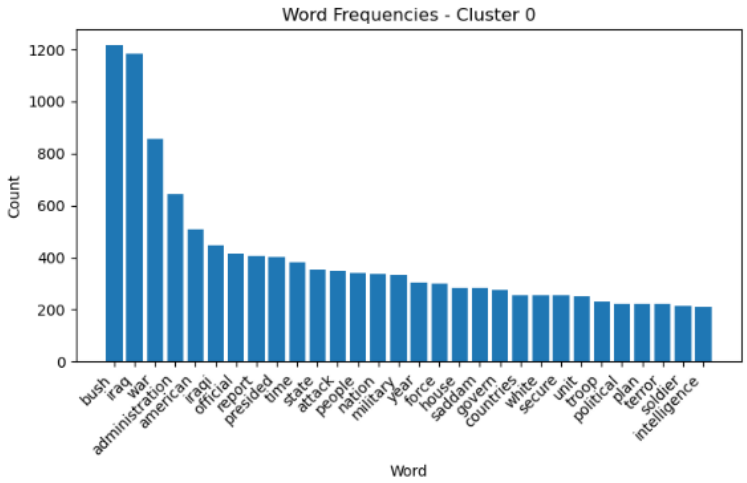
From looking at the T-SNE graph, it can be seen that there are certain clear groups that apparently deal with the same topics (more on the edges), and there is the center of the graph, an area that is difficult to identify what its clear shape is. Perhaps these documents in this area will deal with relatively identical topics, but each document is from a slightly different point of view.



Results/Discussion

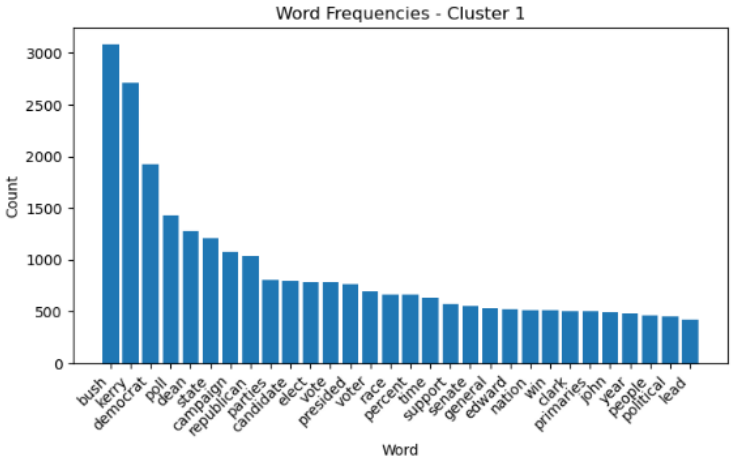
The results of the division into clusters according to K-MEANS:





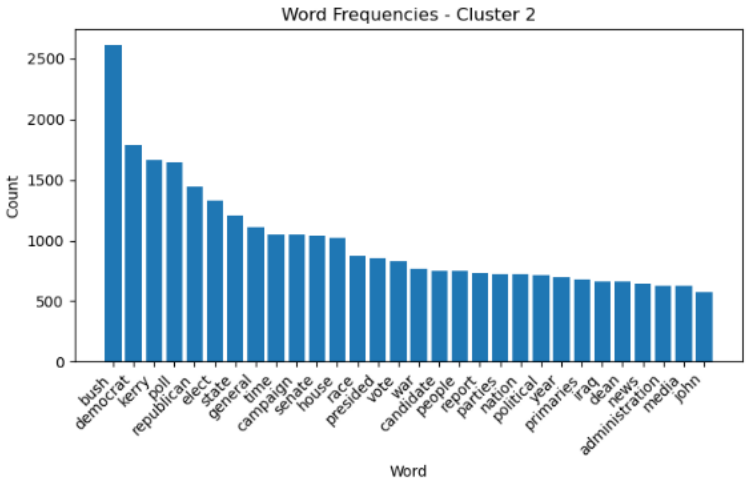
In cluster 0, the central words are "Bush," "Iraq," and "war." These documents delve into the invasion and attack by the U.S. in Iraq during Bush's presidency. Bush saw Iraq as a threat due to its alleged weapons development, leading to Congress approving military action, starting the Iraq War in 2003.

Cluster data - cluster_1:



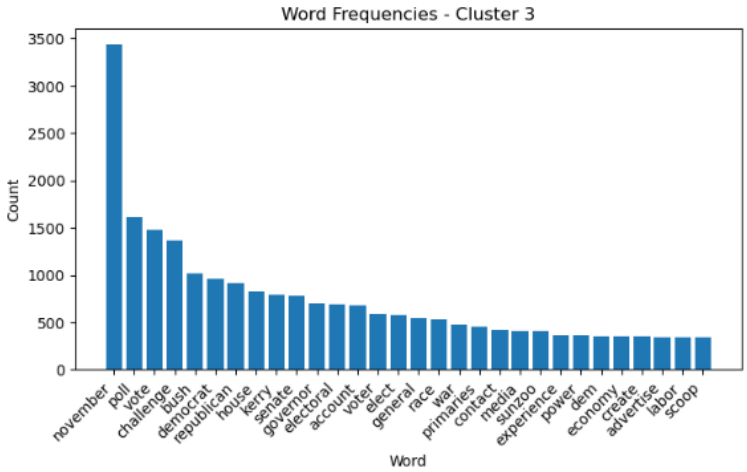
In cluster 1, these documents deal with pre-election polls leading up to the US presidential election. Key figures include George W. Bush, the Republican incumbent seeking re-election, John Kerry as the Democratic candidate, and Howard Dean, a prominent Democratic contender during the primaries.

Cluster data - cluster_2:



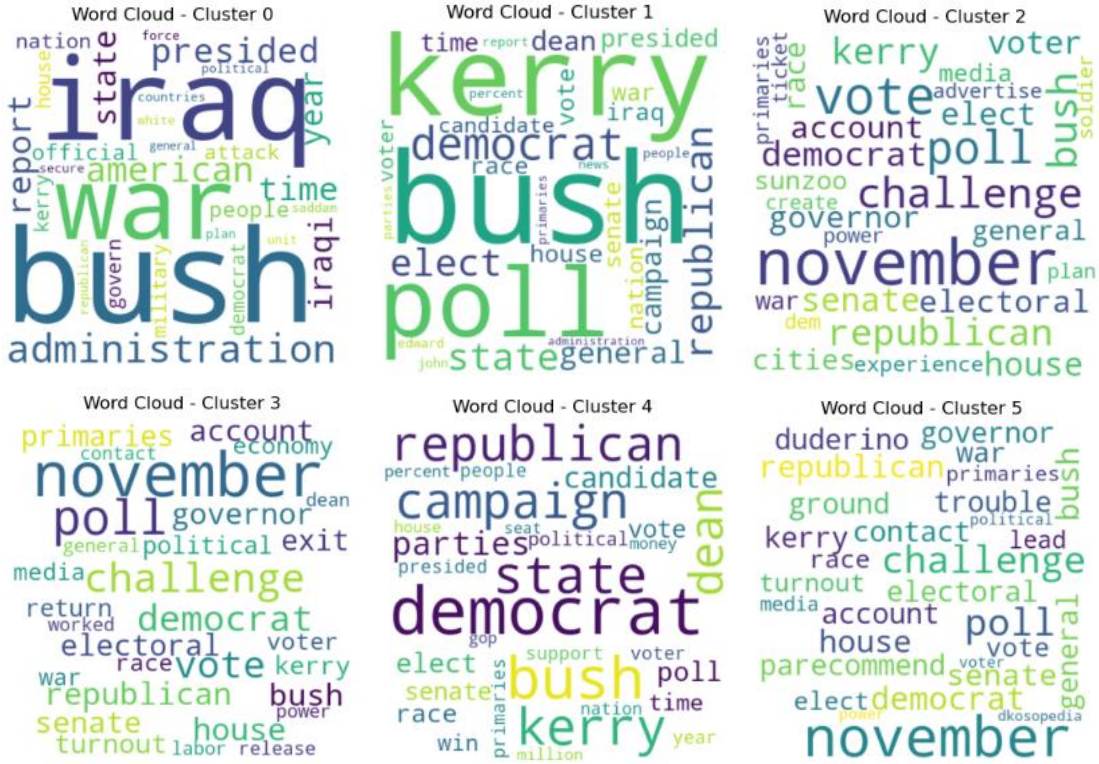
Cluster 2 deals with the political competition between the two largest parties in the United States, the Democrats and the Republicans. Specifically, it focuses on the rivalry between George W. Bush, representing the Republican Party, and John Kerry, representing the Democratic Party. This cluster examines the likely contrasting policies, campaign strategies, and public perceptions that define the broader partisan dynamics in American politics.

Cluster data - cluster_3:



In cluster 3 the emphasis is on the timing of the elections, with a specific emphasis on the month of November when the elections are held. The word "November" stands out significantly, followed by the words: "survey", "vote" and "challenge". This cluster will likely delve into discussions about election processes, voter turnout, survey data, and the challenges and overall electoral dynamics that emerge during this critical period in November.

The results of the division into clusters according to PCA & K-MEANS:



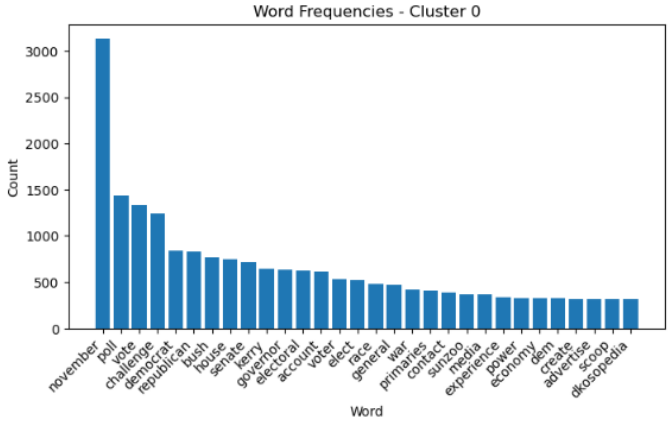
First of all we see that here the division is into 6 clusters (before PCA for k-means there were 4 clusters).
Here too cluster 0,*deals with the invasion of the USA and the attack on Iraq during Bush's presidency, the key words are "Bush", "Iraq" and "war".
Cluster 1, deals with the results of the polls between Bush and Kerry.
Cluster 4, deals with the comparison and competition between the Democratic Party and the Republican Party.

Clusters 2,3,5 are quite similar:
Cluster 2, deals with polls and elections in preparation for election day in November, in the contest between Bush and Kerry.
Cluster 3, deals with polls and elections in preparation for election day in November. Emphasizes how the electoral system works and deals less with the competitors. Gives more emphasis on the Republican side.
Cluster 5, in terms of words, is the same as cluster 3, but gives more emphasis on the Republican side.

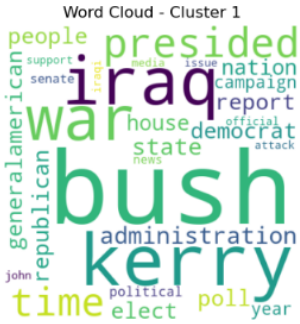
From looking at the SSE and silhouette results with and without PCA, it seems that after PCA the values are better for each type of partition.

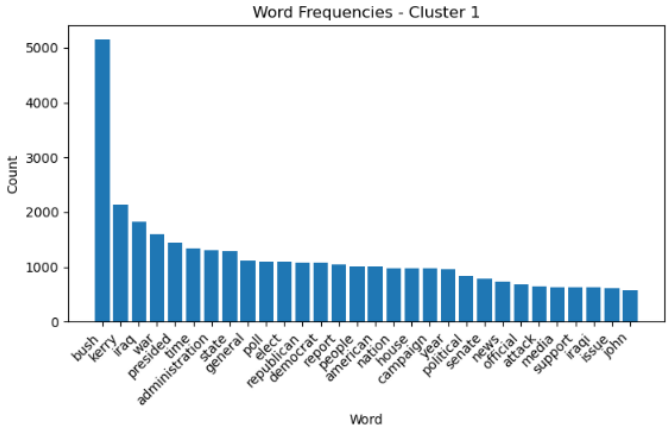
The results of the division into clusters according to DBSCAN:

Cluster data - cluster_0:

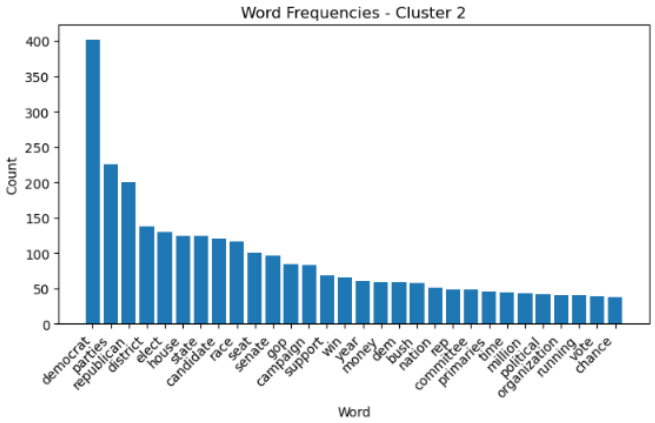


Cluster data - cluster_1:

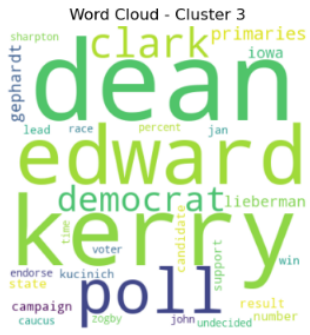


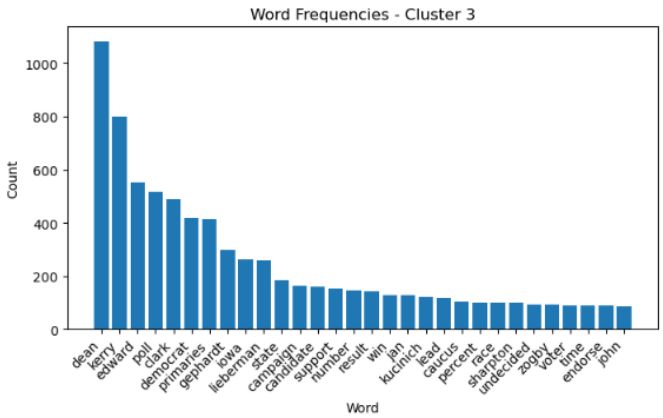


Cluster data - cluster_2:



Cluster data - cluster_3:





In cluster 0, the preoccupation is with the polls for election day which will be in November, the preoccupation with the number of voters from each party.

In cluster 1, the key words are "Bush", "Kerry", "Iraq" and "war". These documents are in the war in Iraq and the follow-up measures that the president chooses.

In cluster 2, Comparison between the parties, Democrats and Republicans (the Senate, the House, and the competition).

In Cluster 3, the focus is on primary election polls related to the competition for leadership within the Democratic Party, between Dean and Kerry.

Conclusion

In concluding my analysis, I gained valuable insights into the structure and themes of the Kos Daily dataset. Within each cluster, I identified unique topics such as discussions on the Iraq War, election polls, party rivalries, and primary elections. This enabled me to distinguish clusters not just by content but also by their temporal context, providing a comprehensive understanding also of the dataset's evolution over time.

The implementation of PCA significantly improved clustering quality by preserving essential variance while reducing noise. This led to clearer and more meaningful cluster formations, enhancing interpretability and uncovering underlying patterns within the data. The combined use of clustering algorithms, visualization methods, and PCA analysis provided deep insights into both the content and structural relationships within the dataset. For the PCA analysis, the optimal number of clusters was determined to be 6.

I utilized both the K-means and DBSCAN clustering algorithms and evaluated their performance using metrics such as SSE (Sum of Squared Errors) and silhouette scores. The SSE metric helped assess the compactness of clusters, while silhouette scores aided in evaluating cluster separation.

The K-means algorithm, optimized based on SSE and silhouette scores for optimal cluster counts, delivered valuable clustering results, with the optimal number of clusters determined to be 4. For the DBSCAN algorithm, with its density-based approach and parameter tuning for optimal `min_samples`, provided insights into nuanced cluster structures, particularly in managing noise and outliers effectively. Also for DBSCAN, the optimal number of clusters was determined to be 4. This dual approach ensured a robust analysis of the dataset's clustering patterns and characteristics, contributing to a deeper understanding of its underlying dynamics.

Appendices

https://towardsdatascience.com/an-approach-for-choosing-number-of-clusters-for-k-means-c28e614ecb2c

https://medium.com/accel-ai/pca-algorithm-tutorial-in-python-93ff19212026

https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/

https://github.com/ShirSaadi/clustering-project.git