A PROJECT REPORT ON

TEXT CLUSTERING

CIS 6397 - 24353 – TEXT MINING

Group - 5

UNIVERSITY OF HOUSTON

Fall 2023

Professor: Dr. Peggy Lindner

Authors:

Akshitha Surabhi -2251696

(Report writing, editing & review)

Nikhita Peddi – 2215340

(Hierarchical Clustering, Refactoring and Report)

Srilakshmi Mannam – 2151535

(Dendrogram Visualization and clustering using K-means)

Link to Repository:
https://github.com/CIS-6397-Textmining-Spring-2023/miniproject2-miniproject2_group5

1. ABSTRACT

Text mining serves as a crucial method for gleaning valuable insights from vast amounts of unstructured textual data. In this project, we employ text mining and clustering techniques to examine a dataset encompassing 3,430 articles from the influential American political blog, Daily Kos. These articles were published in the run-up to the 2004 U.S. Presidential Election, a momentous political juncture in the United States. Our primary aim is to reveal inherent patterns and connections within the articles. This dataset captures the discussions revolving around significant subjects of that era, including the Iraq War and the candidacies of Howard Dean, John Kerry, and John Edwards. Through the process of article clustering, our objective is to expose concealed themes and narratives, providing crucial insights into the political discourse during this transformative period. The results of this project carry broader implications for content analysis and sentiment tracking, offering illumination on the evolution of political conversations.

2. INTRODUCTION

Clustering in text mining is a technique employed to group similar text documents or pieces of text based on their content or features. Its primary purpose is to unveil hidden structures or patterns within extensive unstructured text data. The process begins with feature extraction, where text documents are converted into a numerical format, usually using methods like TF-IDF or word embeddings. This numerical representation enables the calculation of document similarity using metrics like cosine similarity.

Once the similarity between documents is established, clustering algorithms such as K-means, hierarchical clustering, or DBSCAN are used to create clusters. These algorithms aim to minimize the dissimilarity within clusters and maximize dissimilarity between clusters.

Clustering text data has numerous applications, including improving information retrieval by efficiently locating relevant documents in large collections. It aids in document organization, categorizing news articles, research papers, or other documents into topics or themes for easier navigation. Furthermore, it plays a pivotal role in identifying topics or themes within a corpus, assisting in topic modeling and text classification tasks. Clustering can also be employed in recommendation systems and anomaly detection, where unusual or outlier documents can be detected by examining clusters.

2.1 Data Preprocessing

The dataset employed in this project is a meticulously assembled compilation of written content. It encompasses 3,430 articles and blog posts, which reflect the ever-changing content landscape of Daily Kos, a prominent American political blog with a progressive perspective. Each article is represented as a row, and the columns correspond to a thoughtfully chosen set of 1,545 words. These words have been selected based on their presence in a minimum of 50 different articles, ensuring that the dataset focuses on important and recurring themes. The values within the dataset indicate how frequently each word appears in a particular document, providing a quantitative measure of word usage. This preprocessed and refined dataset forms the foundation for our text clustering project, enabling us to identify patterns, themes, and connections within the political discourse presented on Daily Kos during the pivotal 2004 United States Presidential Election. The dataset was imported from a CSV file named 'dailykos.csv' into a Pandas DataFrame using the pd.read_csv() function. This step converts the raw data into a structured format suitable for analysis.

2.2 Hierarchical Clustering

Hierarchical clustering in text mining is a technique used to group similar text documents into a hierarchical structure. This method is valuable for understanding the relationships between documents, organizing large text corpora, and uncovering patterns within textual data. There are two primary approaches to hierarchical clustering: agglomerative and divisive.

Hierarchical clustering requires a similarity or dissimilarity measure, such as cosine similarity, Jaccard similarity, or Euclidean distance, to assess the similarity between documents. The choice of similarity measure and linkage criteria (e.g., single, complete, or average linkage) can significantly affect the clustering results.

The output of hierarchical clustering is often represented as a dendrogram, which visually displays the hierarchical relationships between documents. Users can select a level of clustering that best suits their analytical needs, making it a versatile tool for various text-mining tasks, including document organization, topic modeling, and summarization. However, it can be computationally intensive for large datasets.

2.3 K-means Clustering

K-means clustering is a widely used unsupervised machine learning technique employed in text mining to group similar text documents into clusters. The primary goal of K-means clustering is to partition a collection of documents into a predefined number (K) of clusters, where each cluster consists of documents that are more similar to each other in comparison to documents in other clusters.

The process of K-means clustering in text mining can be summarized as follows:

Initialization: Initially, K cluster centroids are randomly or intelligently chosen from the document dataset. These centroids represent the center of each cluster.

Assignment: Each document is assigned to the nearest cluster centroid based on a similarity metric, such as cosine similarity or Euclidean distance. The document belongs to the cluster whose centroid is closest in terms of the chosen metric.

Update: The centroids of the clusters are recalculated based on the documents assigned to them. The centroid of a cluster is typically set as the mean (average) of all the document vectors in that cluster.

Reassignment: Documents are reassigned to the nearest cluster based on the updated centroids. This process of updating centroids and reassigning documents is repeated iteratively.

Convergence: The iterations continue until a stopping criterion is met. This criterion could be a maximum number of iterations or until the centroids no longer change significantly.

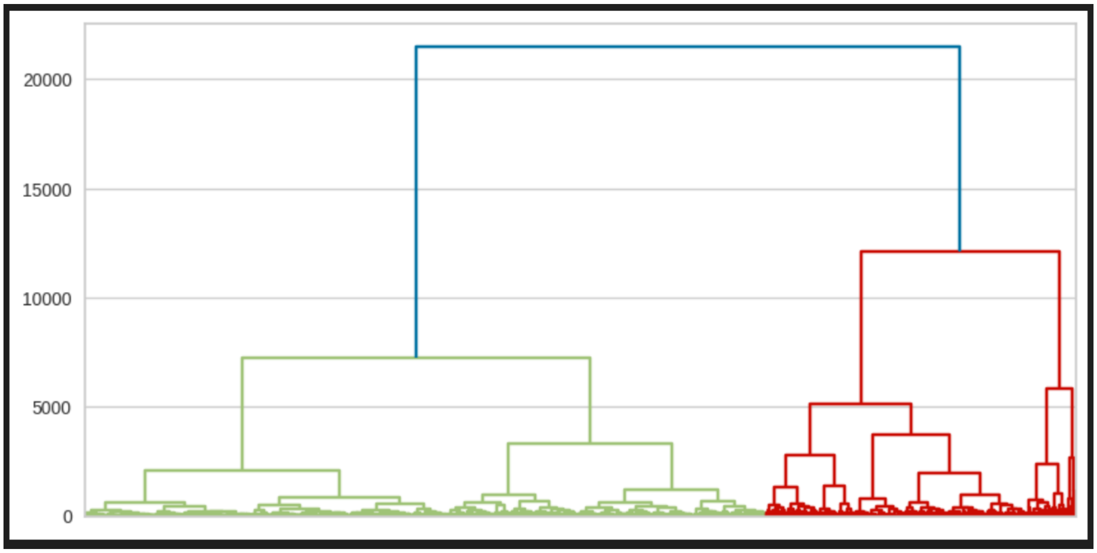
K-means clustering is an efficient and scalable method for text document grouping. However, its success heavily depends on the choice of K (the number of clusters) and the selection of an appropriate similarity metric. In text mining, it's common to represent documents as numerical feature vectors, often using techniques like TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings (e.g., Word2Vec or Doc2Vec) to facilitate the calculation of document similarities.

K-means clustering can be used for various text mining applications, including document categorization, topic modeling, and text summarization, providing insights and organization in large collections of textual data.

In this project, the K-means clustering algorithm is employed with a value of K set to 7, leading to the creation of 7 distinct clusters. Subsequently, an analysis of each cluster's attributes, including the most frequently occurring words, is done based on the cluster assignments. Additionally, the outcomes of the hierarchical clustering were compared with the assignments of the clusters to draw comparisons.

2.4 Euclidean distance

The Euclidean distance is a metric used to quantify the similarity or dissimilarity between two data points in a multi-dimensional space. This project was applied to calculate the dissimilarity between pairs of articles or documents by considering the word frequency in each document. The resulting distance matrix was then utilized as input for both hierarchical clustering and k-means clustering algorithms. While k-means clustering assigns data points to a fixed number of clusters by minimizing the sum of squared distances from each point to its cluster center, hierarchical clustering organizes data points into hierarchical clusters based on their distances from each other.



3. METHODOLOGY

3.1 Data Visualization

After generating the dendrogram, we observed a dispersion of data points across a range. The Euclidean distance metric was used to compute distances between pairs of data points. Due to the high dimensionality of the data, which is determined by the number of unique words in each article, the computation of distances is also time-consuming and computationally expensive. In this scenario, with n data points, there are n(n-1)/2 pairs to calculate distances for, resulting in a time complexity of O(n^3). Creating the distances may take time because there are many articles (n) and a high number of dimensions (1545 variables).

3.2 Data Exploration

We conducted both Hierarchical and K-means clustering on the data provided in the problem statement. In addition, we identified the top 6 words for each cluster. Word clouds would have been a better option if we did not get the context from top 6 words. To enhance the visibility and formatting of the code, we opted to declare individual functions. The outcomes of these analyses are presented in the following section.

4. EXPERIMENTAL RESULTS

4.1 Choosing the number of clusters

To cluster news articles, a good choice for the number of clusters could be 5. To determine the value of k, we used the elbow method. This method helps us in determining the most optimal value for k. This allows readers to choose between articles biased towards

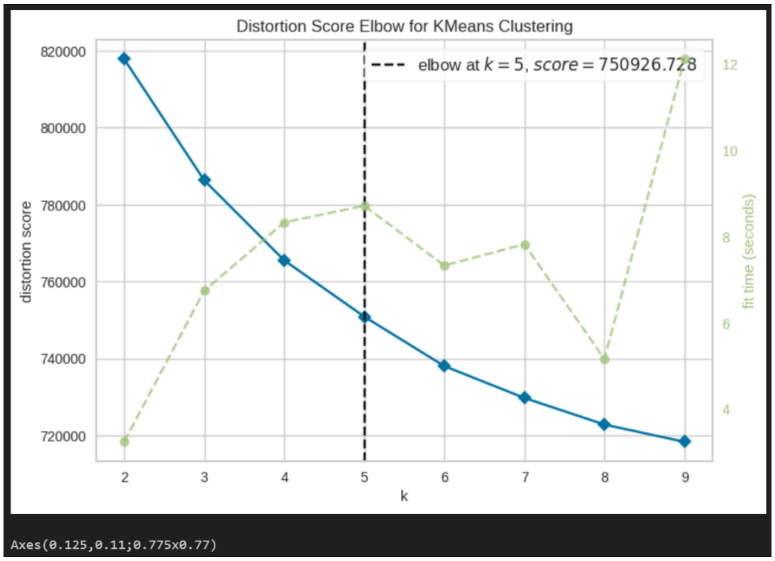
1. Democrats

2. Republicans

3. War

4. Foreign Policy

5. Elections

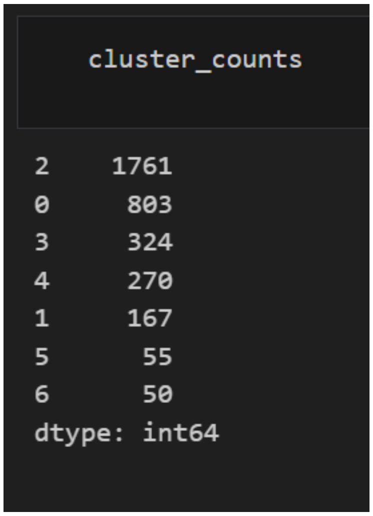


```
Axes(0.125,0.11;0.775x0.77)
```

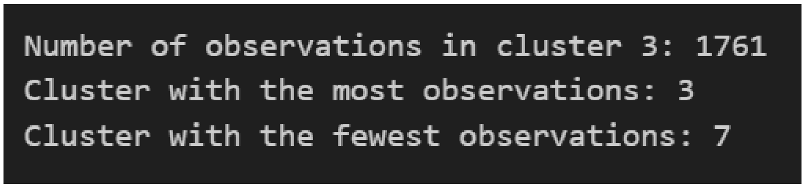
4.2 Analyzing data using Hierarchical Clustering

The hierarchical clustering resulted in value counts, revealing that Cluster 2 is the largest, containing 1761 articles. It is followed by Cluster 0 with 803 articles and Cluster 3 with 324 articles. Clusters 4, 1, 5, and 6 are smaller, comprising 270, 167, 55, and 50 articles, respectively.

These findings indicate that the majority of the dataset's articles belong to Cluster 2. The smaller clusters, although less extensive, might contain articles that are distinct from those in the largest cluster. Further analysis, such as the examination of keywords and topics associated with each cluster, could offer additional insights into the nature of the articles within each cluster.

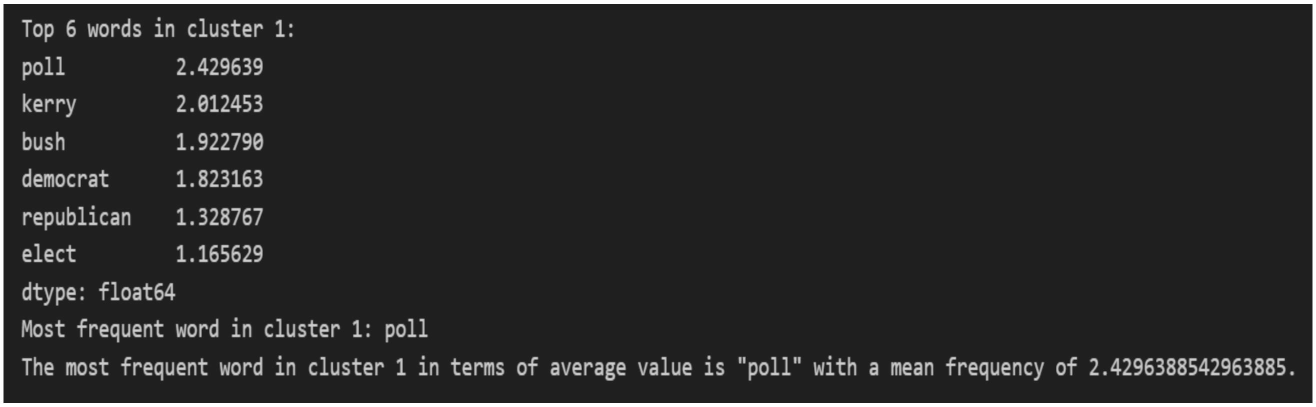


```
    cluster_counts

2    1761
0     803
3     324
4     270
1     167
5      55
6      50
dtype: int64
```



```
Number of observations in cluster 3: 1761
Cluster with the most observations: 3
Cluster with the fewest observations: 7
```

Most frequent words in cluster 1:

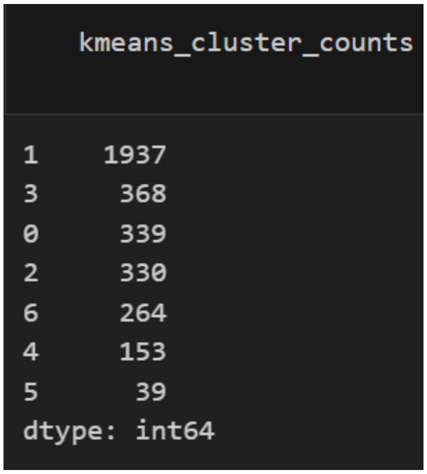


```
Top 6 words in cluster 1:
poll          2.429639
kerry         2.012453
bush          1.922790
democrat      1.823163
republican    1.328767
elect         1.165629
dtype: float64
Most frequent word in cluster 1: poll
The most frequent word in cluster 1 in terms of average value is "poll" with a mean frequency of 2.4296388542963885.
```

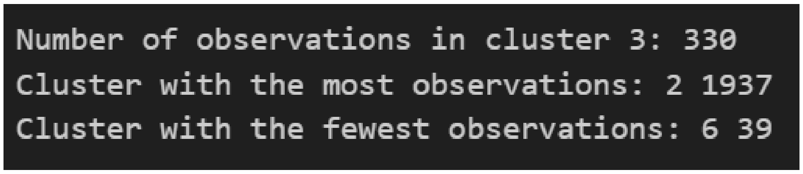
4.3 Analyzing data using K-means Clustering

The k-means clustering revealed that the majority of articles were assigned to cluster 1, with 1937 articles. This is followed by clusters 3 (368 articles), 0 (339 articles), and 2 (330 articles). Clusters 6, 4, and 5 had fewer articles, with 264, 153, and 39 articles, respectively.

These findings indicate that there might be distinct groups of articles in the dataset characterized by their word content. Further examination of these clusters could unveil patterns related to the topics or themes of the articles.



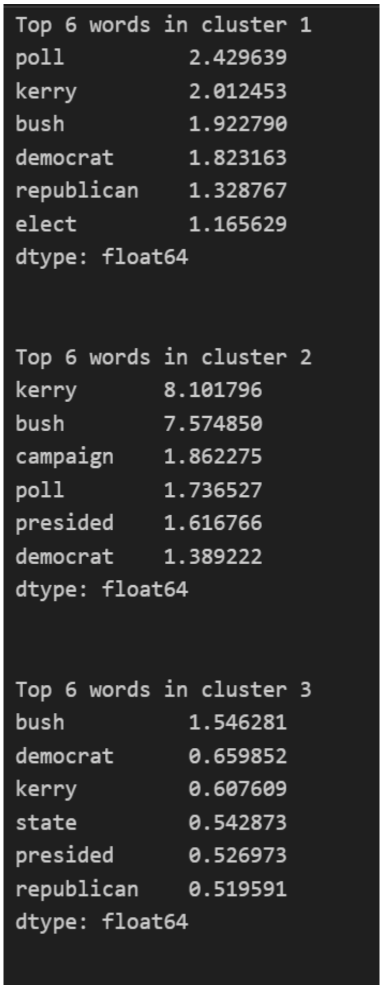
```
    kmeans_cluster_counts

1    1937
3     368
0     339
2     330
6     264
4     153
5      39
dtype: int64
```



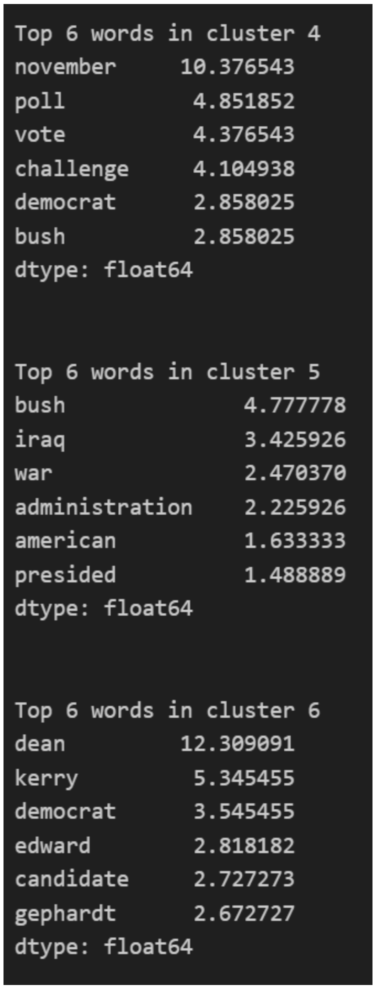
```
Number of observations in cluster 3: 330
Cluster with the most observations: 2 1937
Cluster with the fewest observations: 6 39
```

4.4 Top 6 words using Hierarchical Clustering and K-means Clustering

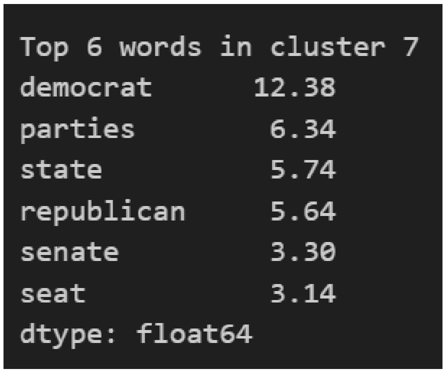
Hierarchical Clustering: The top 6 words using hierarchical clustering in each of the 7 clusters with their frequency are obtained. Using this we can find the clusters related to the Iraq war and the Democratic party.



```
Top 6 words in cluster 1
poll          2.429639
kerry         2.012453
bush          1.922790
democrat      1.823163
republican    1.328767
elect         1.165629
dtype: float64


Top 6 words in cluster 2
kerry       8.101796
bush        7.574850
campaign    1.862275
poll        1.736527
presided    1.616766
democrat    1.389222
dtype: float64


Top 6 words in cluster 3
bush          1.546281
democrat      0.659852
kerry         0.607609
state         0.542873
presided      0.526973
republican    0.519591
dtype: float64
```



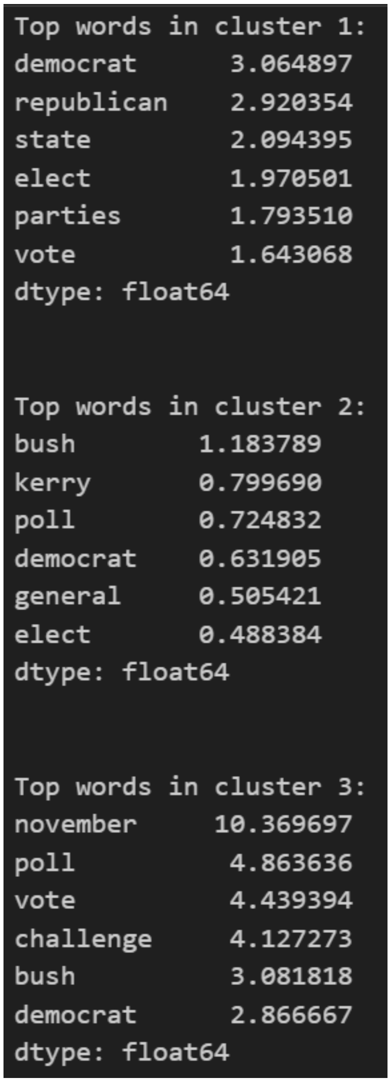
```
Top 6 words in cluster 4
november     10.376543
poll          4.851852
vote          4.376543
challenge     4.104938
democrat      2.858025
bush          2.858025
dtype: float64


Top 6 words in cluster 5
bush              4.777778
iraq              3.425926
war               2.470370
administration    2.225926
american          1.633333
presided          1.488889
dtype: float64


Top 6 words in cluster 6
dean         12.309091
kerry         5.345455
democrat      3.545455
edward        2.818182
candidate     2.727273
gephardt      2.672727
dtype: float64
```



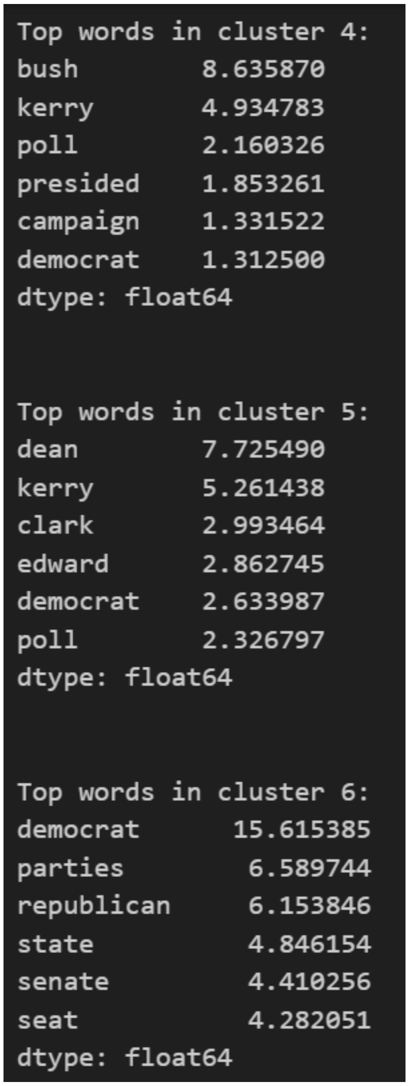
```
Top 6 words in cluster 7
democrat      12.38
parties        6.34
state          5.74
republican     5.64
senate         3.30
seat           3.14
dtype: float64
```

According to the above data, top 6 frequent words in the 7 clusters, Cluster 5 relates to the Iraq War since it has the most frequent words related to Iraq War such as 'Iraq' and 'war', and Cluster 2 is related to the Democratic Party since it has most frequent words related to Democratic party such as 'democrat', 'kerry', 'candidate'. Though the words for the democratic party are in other clusters as well (1, 3, and 6), the mean of the most frequent word 'kerry' is the highest in Cluster 2.

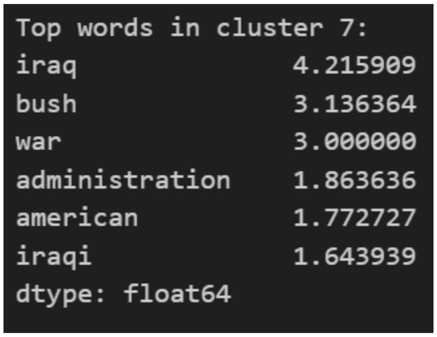
K – means Clustering: The top 6 words using K-means clustering in each of the 7 clusters with their frequency is obtained. Using this we can find the clusters related to the Iraq war and the Democratic party.



```
Top words in cluster 1:
democrat      3.064897
republican    2.920354
state         2.094395
elect         1.970501
parties       1.793510
vote          1.643068
dtype: float64


Top words in cluster 2:
bush        1.183789
kerry       0.799690
poll        0.724832
democrat    0.631905
general     0.505421
elect       0.488384
dtype: float64


Top words in cluster 3:
november     10.369697
poll          4.863636
vote          4.439394
challenge     4.127273
bush          3.081818
democrat      2.866667
dtype: float64
```



```
Top words in cluster 4:
bush        8.635870
kerry       4.934783
poll        2.160326
presided    1.853261
campaign    1.331522
democrat    1.312500
dtype: float64


Top words in cluster 5:
dean        7.725490
kerry       5.261438
clark       2.993464
edward      2.862745
democrat    2.633987
poll        2.326797
dtype: float64


Top words in cluster 6:
democrat      15.615385
parties        6.589744
republican     6.153846
state          4.846154
senate         4.410256
seat           4.282051
dtype: float64
```

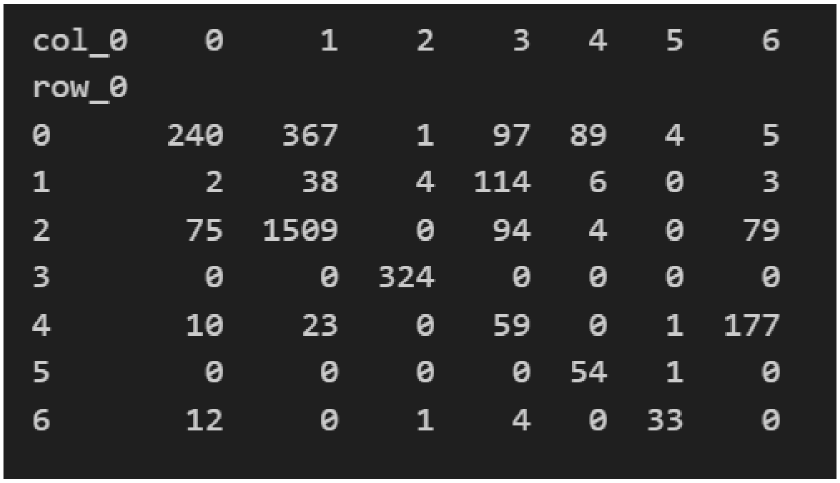


```
Top words in cluster 7:
iraq              4.215909
bush              3.136364
war               3.000000
administration    1.863636
american          1.772727
iraqi             1.643939
dtype: float64
```

According to the above data, top 6 frequent words in the 7 clusters, Cluster 5 corresponds to the Iraq War since it has terms like 'Iraq', and 'war' in its most frequent words. Clusters 1, 2, 3, and 6 have terms related to the democrat Party such as 'democrat' and 'Kerry' but among all cluster 2 is more suitable for the democratic party because the term 'kerry', the presidential candidate for the democratic party, has the highest mean frequency in cluster 2.

4.5 Cross-tab Comparison

Cross-tab comparison between hierarchical and k-means clustering in text mining is a method to evaluate the performance of these clustering algorithms. It involves creating a table that shows how data points or documents are assigned to clusters by each method. By comparing the results, you can assess the consistency of cluster assignments between the two algorithms. This analysis helps in understanding which algorithm is more suitable for grouping text data, considering factors like cluster size and data distribution. It aids in choosing the most effective clustering approach based on the specific characteristics and goals of the text-mining task.



```
col_0    0     1    2    3   4   5    6
row_0
0      240   367    1   97  89   4    5
1        2    38    4  114   6   0    3
2       75  1509    0   94   4   0   79
3        0     0  324    0   0   0    0
4       10    23    0   59   0   1  177
5        0     0    0    0  54   1    0
6       12     0    1    4   0  33    0
```

The hierarchical cluster that best corresponds to K-means cluster 2 is 3.

The hierarchical cluster that best corresponds to K-means cluster 3 is 4.

5. CONCLUSION

In this project, we employed hierarchical and k-means clustering to categorize news articles and blog posts from the Daily Kos political blog. We created seven clusters, and for both clustering methods, we identified the six most significant words in each cluster.

Through this clustering analysis, we successfully identified key themes and subjects within the Daily Kos blog, allowing us to organize articles based on their content. This could prove valuable for presenting articles to readers in a structured manner and for conducting further research on specific topics or issues.