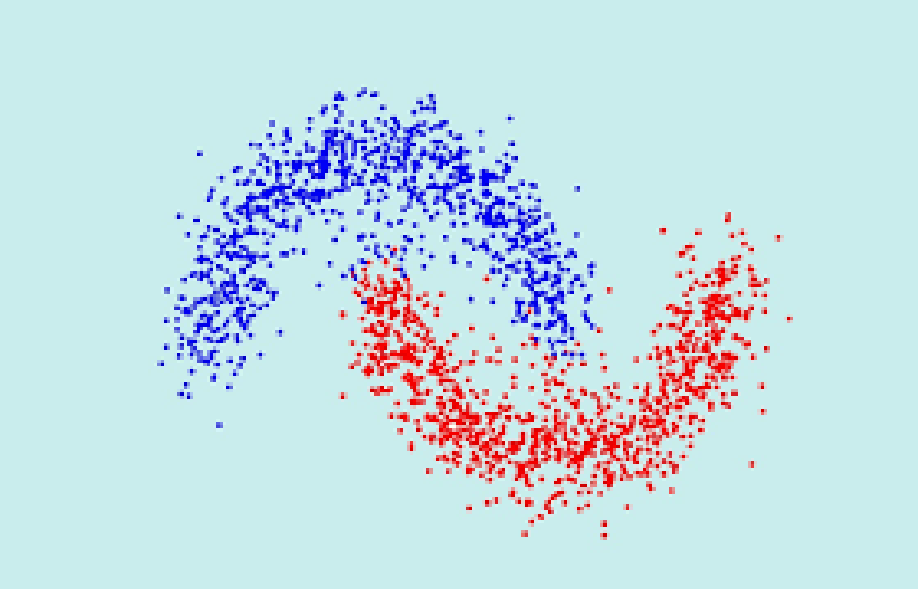
clustering



Amit Fallach

/https://www.linkedin.com/in/amitfallach

amitfallach (Amit Fallach) · GitHub

Abstract

This scientific report delves into the clustering analysis of 3,430 articles or blogs published in Kos Daily during the 2004 United States presidential election. Employing K-means and Hierarchical Clustering - Agglomerative Clustering, the study seeks to unravel hidden patterns in the political discourse. The K-means algorithm, tested with varying division methods, yielded a pivotal decision to form three clusters based on a comprehensive evaluation of silhouette scores and Sum of Squared Errors (SSE). Meanwhile, Hierarchical Clustering, utilizing a bottom-up approach with diverse linkage methods, showcased clear boundaries among clusters, with the exception of the single linkage method, which exhibited sensitivity to small differences between data points. This report presents an insightful comparative analysis of the effectiveness of these clustering algorithms in revealing underlying structures within the dataset, contributing to the broader understanding of political discourse during the 2004 US presidential election.

Introduction

The application of clustering algorithms, specifically k-means and Hierarchical Clustering - Agglomerative Clustering, plays a pivotal role in organizing and categorizing vast amounts of textual information. Document clustering, an established practice in information retrieval, has found widespread application in various search engines, such as PolyMeta, Helioid, and the gov.FirstGov portal, streamlining the accessibility of relevant information. In this study, we focus on applying these clustering techniques to 3,430 articles from Kos Daily, a prominent American political blog that offers a progressive perspective. Originating from the period leading to the 2004 United States presidential election, these articles encapsulate diverse viewpoints, emphasizing foreign policy, particularly the ramifications of the Iraq invasion in 2003.

The central inquiry of this investigation is twofold: firstly, to utilize the k-means and Hierarchical Clustering - Agglomerative Clustering algorithms to discern meaningful structures within the corpus of articles from Kos Daily; secondly, to evaluate the efficacy of these clustering techniques in uncovering nuanced narratives surrounding the 2004 US presidential election. Additionally, a significant part of the research objective is to delve into the inner workings of these algorithms, gaining insights into their mechanisms and behavior, as outlined in the task provided.

Data Overview

Before delving into the detailed analysis and application of clustering algorithms, it's essential to provide an overview of the structure and content of the original data frame. The screenshot below presents a snapshot of selected columns, showcasing the data types and initial entries. Understanding the data structure is fundamental to interpreting the results and conclusions drawn from the subsequent clustering analyses.

[33]:

	Document	abandon	abc	ability	abortion	absolute	abstain	abu	abuse	accept	...	yeah	year	yesterday	york	youll	young	youre	youve	zogby	zone
0	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	1
1	2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	3	0	0	0	0	0	1	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	4	0	0	0	0	0	0	0	0	0	...	0	0	2	0	0	1	0	0	0	0
4	5	0	0	0	0	0	0	0	0	0	...	0	1	1	0	0	1	0	0	0	0

5 rows × 1546 columns

Figure 1: Snapshot of Selected Columns in the Original Data Frame.

The columns displayed include essential information about the frequency of words in each document, offering a glimpse into the variables that form the basis of the clustering algorithms. This visual representation sets the stage for a comprehensive exploration of the clustering results and the subsequent interpretation of the clustered articles from Kos Daily.

K-means Algorithm

K-means is a widely used clustering algorithm designed to partition a dataset into distinct groups based on similarities among data points. The "K" in K-means represents the predetermined number of clusters the algorithm aims to identify within the data. The algorithm iteratively assigns data points to the nearest cluster centroid and updates the centroid based on the mean of the assigned points. This process continues until convergence, resulting in well-defined clusters that encapsulate similar patterns or characteristics in the dataset. K-means is known for its simplicity, efficiency, and effectiveness in uncovering inherent structures within data.

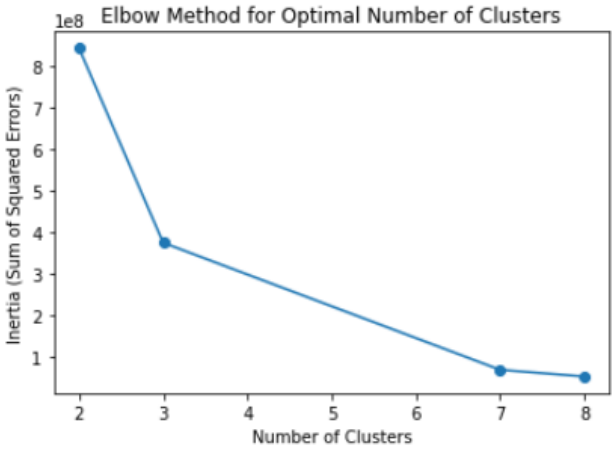
Methodology

For the K-means algorithm, four division methods (2, 3, 7, 8) were tested. The algorithm calculates the Sum of Squared Errors (SSE) and silhouette for each method.

When evaluating clustering models, the Silhouette Score serves as a valuable metric if the primary objective is to assess the separation and distinctiveness of clusters. Higher Silhouette Scores indicate well-defined and separated clusters. On the other hand, if the focus is on determining the optimal number of clusters and assessing their compactness, the Sum of Squared Errors (SSE) becomes a useful metric. Particularly when employed in conjunction with the elbow method, SSE helps in identifying a suitable number of clusters by measuring the compactness of the clusters.

Work process and results

A function was developed to implement the algorithm for each division method, providing the required parameters. Results were presented in an ordered data frame, and an elbow method graph was used to find the optimal K according to SSE.



[4]:

	Clusters	SSE	Silhouette	Sample Count	Min Count	Max Count
0	2	8.415947e+08	0.625761	0 1716 1 1714 dtype: int64	1714	1716
1	3	3.745378e+08	0.588230	1 1145 2 1143 0 1142 dtype: int64	1142	1145
2	7	6.951579e+07	0.540412	3 492 4 492 0 490 1 490 6 489 2...	488	492
3	8	5.343367e+07	0.534682	0 433 7 432 4 430 2 429 5 428 1...	425	433

Figure 2: Elbow Method.

Based on silhouette and SSE, the choice was between 2 and 3 clusters. K=2 had the highest silhouette, while K=3 had the lowest SSE.

In-Depth Analysis of Clusters in K-means

After identifying that the optimal number of clusters falls between 2 and 3 based on silhouette scores and SSE, a more detailed exploration was conducted. Silhouette analysis for clusters ranging from 2 to 10 was visualized in a matrix. The goal was to assess the quality of each cluster division and make a nuanced decision.

Key Considerations

1. Average Silhouette Value

- The silhouette plots displayed the silhouette scores for individual clusters within each division. It was crucial to ensure that the average silhouette score exceeded a certain threshold. A higher average silhouette suggests well-defined and distinct clusters.

2. Consistent Cluster Sizes

- A critical aspect of the analysis was the examination of cluster sizes. For robust clustering, it was essential that the clusters were approximately the same size. This ensures that no single cluster dominates the others, allowing for a balanced representation of distinct patterns within the data.

By scrutinizing the silhouette plots and considering both the average silhouette score and the uniformity of cluster sizes, a comprehensive understanding of the optimal cluster division was achieved. These additional insights helped in making an informed decision regarding the number of clusters that best represented the underlying structure of the dataset.

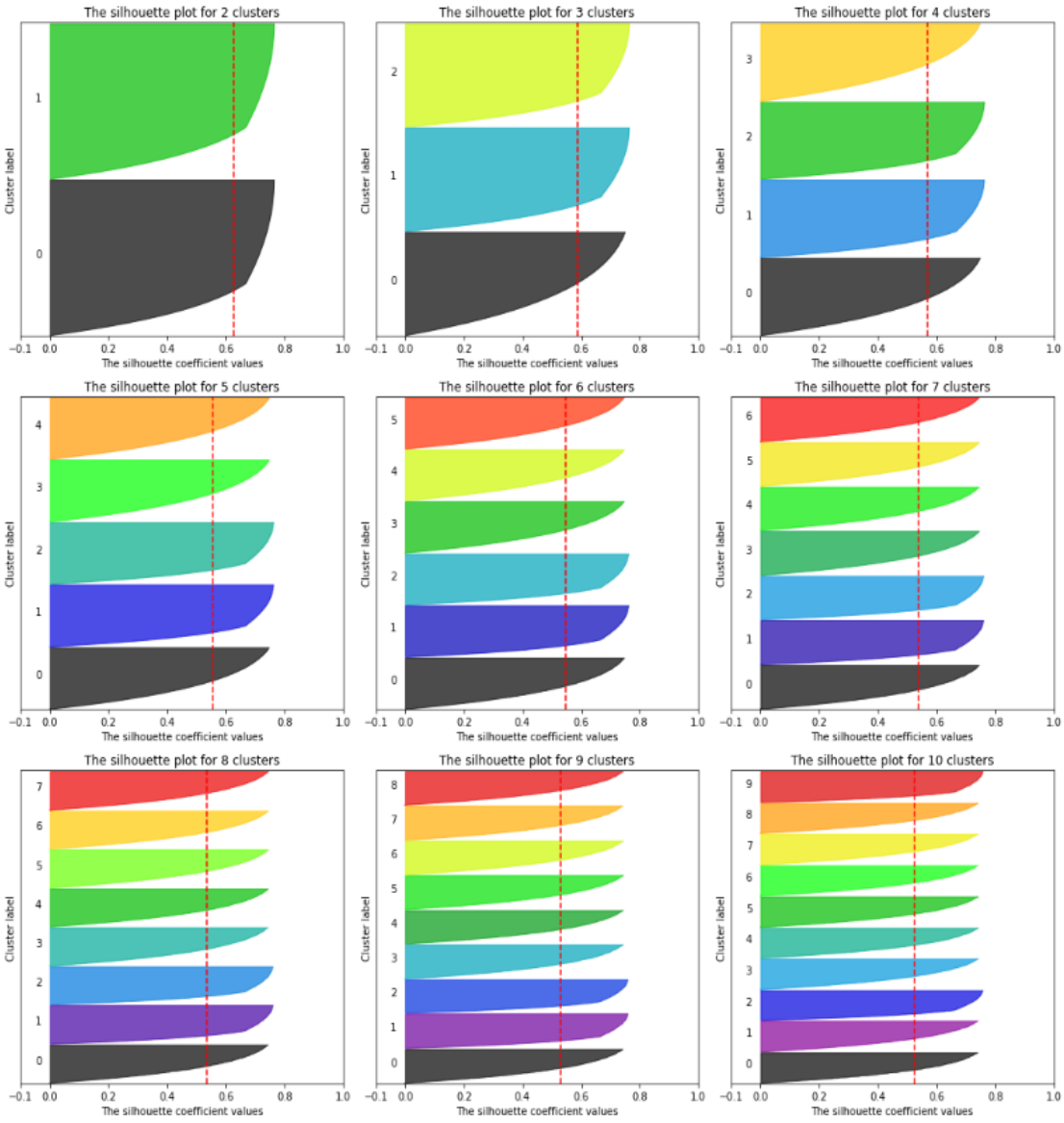


Figure 3: In-Depth Silhouette Analysis.

My decision is to split the data into 3 clusters.

It can be seen that his SSE score is the lowest (through the graph of the elbow method) and in addition the Silhouette is relatively high. After researching the topic, I did not choose the division into 2 clusters (even though the silhouette of the division into 2 is the highest) because in my opinion, the division into 2 clusters is too weak to classify articles with thousands of words. In addition, because the silhouette is relatively balanced for all divisions, I chose to give more weight to SSE in the final decision.

Cluster Analysis

Each cluster was exported to a separate data frame, and a function identified the 30 most frequent words in each cluster. The analysis revealed.

Cluster 0

This cluster seems to focus on the broader political landscape, the election process, and issues related to the 2004 US elections. The most frequent words include terms like "bush," "poll," "Democrat," "Kerry," "Republican," "vote," "elect," "war," "Iraq," and "campaign." The words suggest a concentration on the candidates, political parties, voting, and the overall electoral environment.

Top words for df0 (Rows: 1142):

Word	Count
bush	2902
poll	1707
democrat	1537
kerry	1526
republican	1327
november	1231
vote	1211
state	1041
elect	999
general	890
war	871
iraq	818
presided	815
campaign	793
time	781
senate	778
voter	771
house	739
nation	640
race	637
parties	596
year	569
candidate	555
challenge	553
report	544
political	492
media	484
people	479
oct	454
american	452

Figure 4: cluster 0 top words.

Cluster 1

This cluster appears to be more specific and may center around the candidates themselves and the dynamics of the election campaign. The most frequent words include "bush," "Kerry," "poll," "November," "Democrat," "Republican," "vote," "elect," "war," "State," "General," "Senate," and "house." The inclusion of terms like "race," "people," and "candidate" suggests a focus on the individuals involved and the campaign activities.

Top words for df1 (Rows: 1145):

Word	Count
bush	2501
kerry	1732
poll	1441
november	1320
democrat	1275
republican	1208
Cluster	1145
vote	1119
elect	1024
war	889
state	880
general	875
senate	853
house	817
campaign	800
race	785
iraq	737
time	736
presided	674
people	616
voter	600
challenge	597
candidate	551
parties	547
report	546
political	518
percent	513
media	501
year	499
nation	489

Figure 5: cluster 1 top words.

Cluster 2

When I researched a bit on the Internet, I saw that words like "primaries", "Clark" and "Edwards" can indicate that the cluster is centered around the primaries and the course of the preliminary elections in the Democratic Party. John Edwards ran against Kerry and eventually supported him in the Democratic Party. Wesley Clark was a military general from Arkansas. Retired on February 11 and supported Kerry.

Top words for df2 (Rows: 1143):

Word	Count
bush	2536
Cluster	2286
democrat	2041
kerry	2035
poll	1670
dean	1524
november	1230
state	1114
republican	1015
primaries	943
house	937
senate	849
vote	848
elect	842
campaign	809
iraq	746
clark	743
time	708
war	703
administration	699
race	694
presided	692
edward	678
political	657
nation	625
candidate	620
general	615
people	590
parties	589
challenge	572

Figure 6: cluster 2 top words.

It seems like Cluster 0 is broader, covering the general political and electoral landscape, while Clusters 1 and 2 are more specific, honing in on the candidates and the dynamics of the election campaign.

Hierarchical Clustering - Agglomerative Clustering

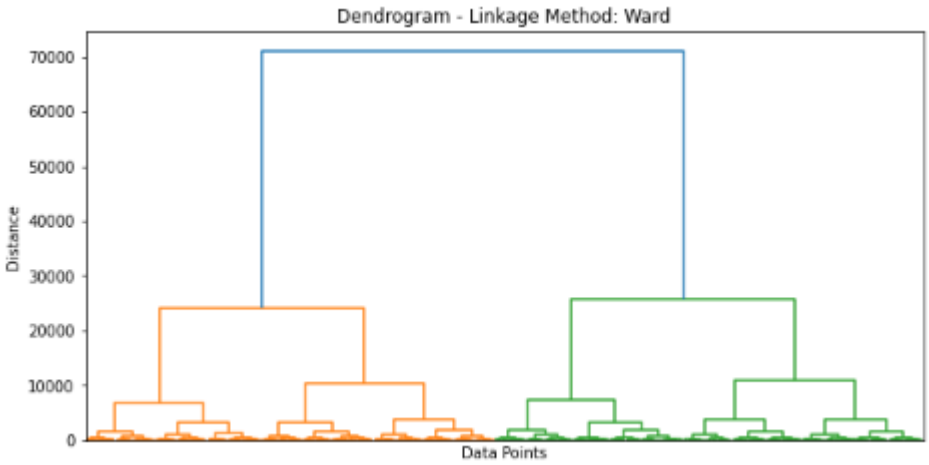
Agglomerative Clustering is a hierarchical clustering algorithm that begins with each data point as an individual cluster and iteratively merges the closest clusters until a single cluster encapsulates all data points. This process creates a tree-like structure, known as a dendrogram, representing the hierarchy of relationships between clusters. Agglomerative Clustering is characterized by its flexibility, allowing for the identification of clusters at various levels of granularity. This method is particularly useful for revealing hierarchical structures within datasets, providing a visual representation of how data points are progressively grouped into clusters.

Methodology

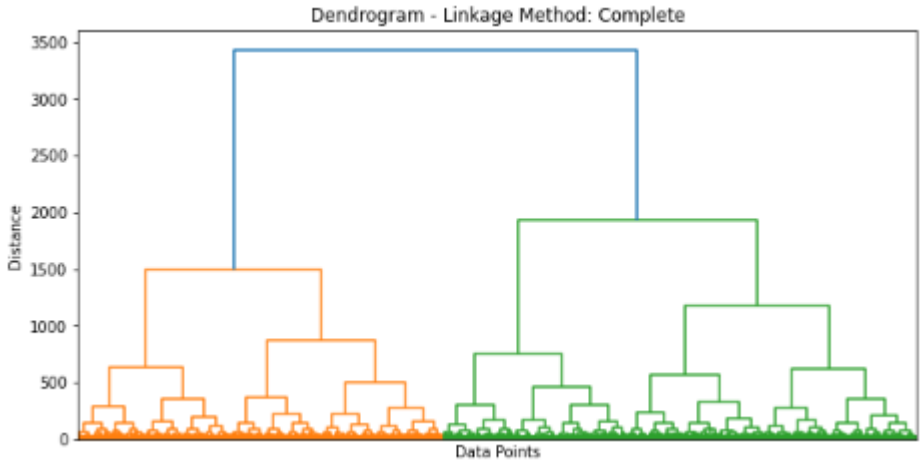
A bottom-up approach was employed for Hierarchical Clustering, specifically using Ward, Maximum (complete linkage), Average linkage, and Single linkage methods.

Work process and results

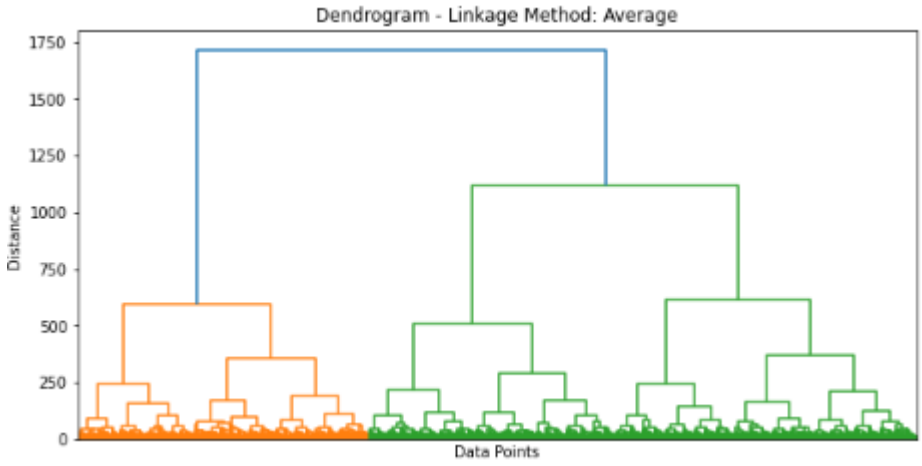
Ward Method: Minimizes the variance within each cluster, suitable for compact and - equally sized clusters.



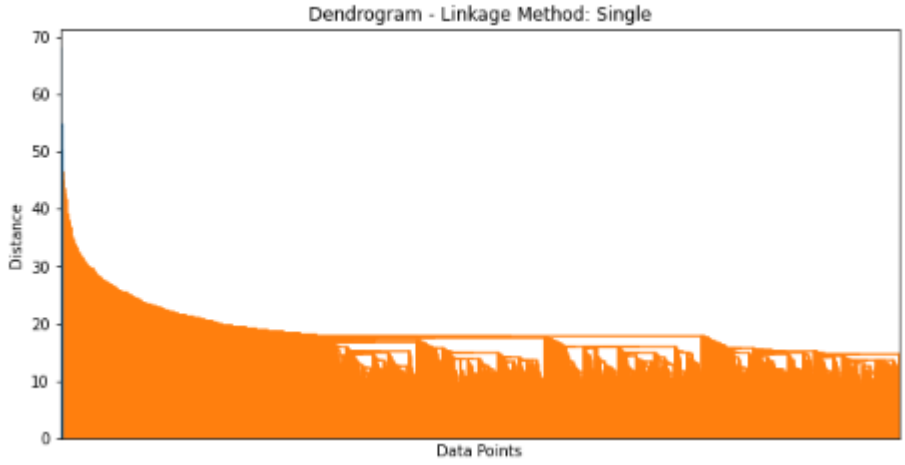
Maximum (Complete Linkage) Method: Measures dissimilarity between clusters by considering the maximum distance between their members.



Average Linkage Method: Calculates the average distance between all pairs of members in different clusters.



Single Linkage Method: Measures dissimilarity by considering the minimum distance - between clusters.



The relative interval graphs of the dendrograms for the three hierarchical clustering methods - Ward, Complete, and Average linkage, reveal distinctive characteristics. In the Ward method, the lower part of the tree exhibits relatively high cluster connections, with a slightly higher prevalence of mixed connections compared to the other two methods. The connections in the Ward and Complete methods appear at similar heights, creating a relatively symmetrical tree structure. However, the Average linkage method presents a different pattern, with clusters on the right side connecting at a higher point than those on the left side, introducing an asymmetry in the dendrogram. Notably, the Single linkage method produced a considerably different outcome, characterized by a chaotic arrangement of numerous connections concentrated at the bottom of the dendrogram. This disorderly structure suggests a higher sensitivity to small differences between data points, resulting in a less cohesive clustering outcome.