TEXT CLUSTERING

CIS 6397-Text mining

By Mukesh Reddy Mavurapu (Conceptualization, Coding, Software, Report editing), Jaswanthi Boyapati (Formal analysis, Code validation, writing original draft), Pradeep Reddy Mallepally (Coding, Report review and editing)

GitHub Repo

Abstract

Using Python and Machine learning Libraries, the document outlines instructions for performing Hierarchical clustering and K-means clustering on the given data file. It describes methods to do the text clustering on the data file. We analyze the observations in each cluster formed in both cases and compare the observations assigned to clusters in the two different methods.

Introduction

Clustering is a fundamental concept that has garnered significant attention from researchers in pattern recognition, statistics, and machine learning. It falls under the category of unsupervised learning, meaning it doesn't rely on pre-existing training samples to build a model. Instead, clustering organizes data into clusters or groups, where the elements within each cluster exhibit stronger similarities with one another compared to those in different clusters. This process is often referred to as unsupervised classification since it accomplishes similar outcomes to classification algorithms but without the need for predefined categories. The primary objective of clustering algorithms, in its most basic form, is to analyze a dataset and identify distinct groups or clusters present within it. Clustering finds application in various domains, such as psychology, business and retail, computational biology, social media network analysis, and many others.

Dataset and Preprocessing:

The dataset comprises 3,430 documents and 1,545 features, representing word frequencies, from articles and blog posts on the Daily Kos during the 2004 U.S. Presidential Election. A comprehensive preprocessing step revealed no missing values, eliminating the need for imputation. This clean dataset is now prepared for in-depth analysis, including clustering or text mining, to uncover insights about the content and organization of the political articles and blogs published during this pivotal election year.

A. Hierarchical Clustering

Hierarchical clustering has been employed to create a grouping of these articles. The idea is to include articles with similar wording in one category. Thus, an approach called agglomerative hierarchy clustering was applied. Larger clusters are formed by combining several smaller ones that begin as separate entities. It builds nested clusters from comparable findings or data values. Recursively dividing or merging clusters based on their degree of similarity, starting with individual data points and progressing to the full dataset, is how this is achieved. In this study, a hierarchy of clustering was used to organize news items or blog entries in groups based on the frequency of word vectors. The resulting dendrogram, which showed the clusters' hierarchical structure, helped us choose the appropriate number of clusters for further investigation. After performing the method of clustering, information must be processed to get rid of any unusual patterns to normalize the values.

B. K-means Clustering

K-means clustering, an unsupervised machine learning method, was used to group 46 posts from the political site Daily Kos based on their word usage statistics. The process involves selecting an initial set of cluster centers, assigning each article to the closest center using a measurement metric like Euclidean distance, updating the centers by calculating the mean of assigned observations, and repeating these steps until convergence criteria are met. After the K-means algorithm converges, each article is categorized into one of the clusters based on its proximity to the nearest center. This results in the creation of K clusters, each containing articles with similar word frequency patterns. In this project, K-means clustering with K=7 was applied, leading to the formation of 7 clusters. Subsequently, the characteristics of each cluster, including their most frequently used terms, were explored using the cluster assignments, and these results were compared to the outcomes of hierarchical clustering.

C. Euclidean Distance

The Euclidean distance is a multidimensional measure of the similarity or dissimilarity between two observations or data points. In this project, the distance between pairs of articles or documents was calculated based on the frequency of occurrence of words in each document. The hierarchical clustering and k-means clustering algorithms were then fed this distance matrix. Hierarchical clustering organises data points or

observations into nested clusters based on the distances between them, in contrast to k-means clustering, which groups data points or observations into a fixed number of clusters by minimizing the sum of squared distances between each observation and its cluster center.

Methodology

A. Visualisation of Data

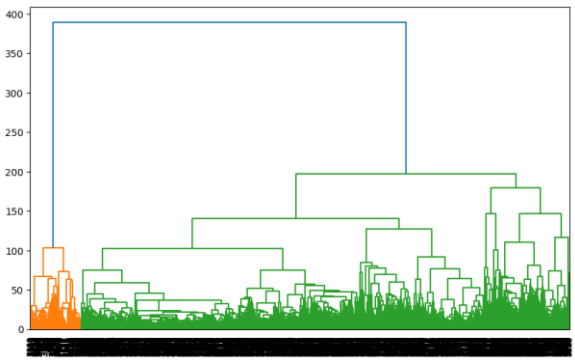
The dendrogram has been plotted, and the data is dispersed throughout. Since there are n(n-1)/2 pairs, the Euclidean distance metric is used to calculate pairwise distances, and the time complexity of the calculation is 0(n3). It will take time to calculate distance because the data is high-dimensional, and the number of variables is the total number of words in each article.

B. Examining Data

We utilized K-means and Hierarchical clustering techniques on the dataset in alignment with the problem statement. We also identified the top six words for each cluster. To enhance code readability and organization, we defined individual functions. The subsequent section presents the results, denoted as "1."

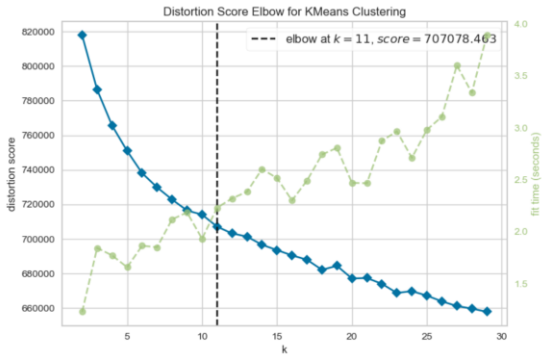
Q1:

Euclidean distances are determined by finding the square root of the sum of the squared discrepancies between the feature values of each pair of data points. This process may be time-consuming. Specifically, when dealing with sizable datasets with a substantial number of features, computing Euclidean distances can present computational challenges. As the dataset's size grows, the number of pairwise distances that need to be calculated increases exponentially, resulting in a considerable increase in computation time. Moreover, if the dataset exhibits high dimensionality or encompasses a large number of features, the computation becomes even more intensive due to the increased number of operations required.



Q2:

I would consider 5 clusters as the dataset naturally falls into common categories like "Election Coverage," "Foreign Policy", "Social Issues", "Economic Issues" and "Technology". The "elbow method" is a commonly used approach to figure out how many clusters are best for a K-means clustering analysis. Essentially, it involves creating a graph that shows how the variation in the data (explained by the clusters) changes with different numbers of clusters. We're on the lookout for a specific point in this graph, often referred to as the "elbow," where the rate of decrease in variation starts to slow down noticeably. This "elbow" point is usually a good estimate for the optimal number of clusters. Using Elbow method, we got the most feasible k value as 11 =.



In practical terms, we start by considering a range of possible cluster numbers, say from 1 to 10. Then, we perform the K-means clustering process for each of these numbers. For every clustering, we calculate a value called "inertia," which essentially measures how compact the clusters are.

Next, we plot the inertia values against the number of clusters. The key insight we're seeking is to identify the point in the plot where the inertia levels off. This is the "elbow" point. At this juncture, adding more clusters doesn't significantly improve the clustering quality, so we usually choose the number of clusters corresponding to this point.

Q3:

```
Cluster 1: 1761 observations
Cluster 2: 167 observations
Cluster 3: 324 observations
Cluster 4: 803 observations
Cluster 5: 50 observations
Cluster 6: 270 observations
Cluster 7: 55 observations
The cluster with the most observations is cluster 1
The cluster with the fewest observations is cluster 5
```

Cluster 3 Has 324 Observations
Cluster 1 is the largest cluster comprising 1761 observations.
Cluster 5 is the smallest cluster with 50 observations in it.

Q4:

The following step involves extracting and examining the frequency of words within cluster 1 of the Data Frame. The results are saved in a dictionary, which can be employed for extended analysis or for gaining insights into the word attributes within that specific cluster. This process computes and records the word frequencies for cluster 1 of the Data Frame, enabling a detailed exploration of the word occurrences within this cluster or for supplementary research purposes.

```
Top 6 words in Cluster 1:
bush          1.546281
democrat      0.659852
kerry         0.607609
state         0.542873
presided      0.526973
republican    0.519591
dtype: float64
```

Mean Frequency Distribution (Pending)

Q5:

Cluster 6, distinguished by keywords like "iraq" and "war," is the cluster that most comprehensively represents the topic related to the Iraq War. The presence of these specific terms in the cluster indicates a direct association with the Iraq War, making it the most relevant cluster for describing this particular aspect of the topic you inquired about.

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

We can indeed choose Cluster 7 as the answer to your question. Cluster 7 primarily contains terms related to the 2004 Democratic presidential nomination, such as "dean," "Kerry," "democrat," and "candidate." It is closely associated with the Democratic Party's activities and candidates during that election cycle. While Cluster 7 may not directly represent the Iraq War, it does pertain to the Democratic Party and the 2004 Democratic presidential nomination. If your question is specifically about the Democratic Party's role in the 2004 election, then Cluster 7 would be a valid choice.

```
Top 6 words in Cluster 7 :
dean          12.309091
kerry          5.345455
democrat       3.545455
edward         2.818182
candidate      2.727273
gephardt       2.672727
dtype: float64
```

Q6:

```
Cluster 1: 339 observations
Cluster 2: 1937 observations
Cluster 3: 330 observations
Cluster 4: 368 observations
Cluster 5: 153 observations
Cluster 6: 39 observations
Cluster 7: 264 observations
The cluster with the most observations is cluster 2
The cluster with the fewest observations is cluster 6
```

cluster 3 has 330 observations

cluster 2 contains the Highest number of observations with a count of 1937 in it.

cluster 6 contains the Fewest number of observations with a count of 39 in it.

Q7:

Top 6 most frequent words according to mean frequency in each cluster and tabulate in a ranking order but this time using K-means

1 democrat republican state elect parties vote
2 bush kerry poll democrat general e lect
3 november poll vote challenge bush democrat
4 bush kerry poll presided campaign democrat
5 dean kerry clark edward democrat poll
6 democrat parties republican state senate seat
7 iraq bush war administration american iraqi

From the above Table

Cluster 7 best describes the cluster related to war as the words Iraq, Iraqi and war are frequently occurring in the observations of the cluster.

Cluster 5 best corresponds to democratic party as words like dean, Kerry, democrat and Edward which are related to democratic party are frequent in this cluster observations.

Q8:

K-Means Cluster	0	1	2	3	4	5	6
Hierarchical Cluster							
0	0	3	91	1487	0	58	122
1	3	4	2	38	0	114	6
2	324	0	0	0	0	0	0
3	1	91	249	359	7	90	6
4	1	0	9	0	35	4	1
5	0	0	12	18	1	44	195
6	0	54	0	0	1	0	0

From the above cross tab visualization

We can observe that the number of observations tagged under cluster 2 by K means model have most overlap with the cluster 2 of the hierarchical model. The best hierarchical cluster to imply cluster 2 in K is cluster 2 itself, with 1509 corresponding observations.

We can see that the observations in cluster 3 of the k-means model have a large overlap with the first cluster of the hierarchical model. Therefore, the best hierarchical cluster corresponding to cluster 3 of K means that the cluster is cluster 1.

Experimental Results

A. Using Hierarchical Clustering for Data Analysis

After obtaining the value counts for hierarchical clustering, Cluster 1, which contained 1761 articles, was found to be the largest cluster. Cluster 3, which contained 324 articles, and Cluster 4, which contained 803 articles, came next. There were 270, 167, 55, and 50 articles in clusters 6, 2, 7, and 5, respectively. These findings imply that Cluster 1 comprises the majority of the dataset's articles. Even though they are smaller, the articles in the other clusters might be different from those in the largest cluster. Additional investigation, such as looking at the subjects and keywords connected to each cluster, might reveal more details about the kinds of articles that make up each cluster.

```
Cluster 1: 1761 observations
Cluster 2: 167 observations
Cluster 3: 324 observations
Cluster 4: 803 observations
Cluster 5: 50 observations
Cluster 6: 270 observations
Cluster 7: 55 observations
The cluster with the most observations is cluster 1
The cluster with the fewest observations is cluster 5
```

Cluster counts

```
{0: 1761, 1: 167, 2: 324, 3: 803, 4: 50, 5: 270, 6: 55}
```

B. Applying K-means clustering to data analysis

Most articles were assigned to cluster 4 (1902 articles), which was followed by clusters 3 (363 articles), 7 (330 articles), and 1 (329 articles), according to the value counts for the k-means clusters. There were fewer articles in Clusters 6, 2, and 5, with 310, 152, and 44 articles, respectively. These findings imply that the dataset may contain different groups of articles that are distinguished by the words they contain. Subsequent examination of the clusters could identify trends in the subjects or themes covered in the articles.

```
Cluster 1: 329 observations
Cluster 2: 152 observations
Cluster 3: 363 observations
Cluster 4: 1902 observations
Cluster 5: 44 observations
Cluster 6: 310 observations
Cluster 7: 330 observations
The cluster with the most observations is cluster 4
The cluster with the fewest observations is cluster 5
```

C. Top 6 words using Hierarchical clustering & K-means Clustering.

The outcomes of the two cluster groups created with these techniques are the same. The words were the same for most clusters, although the cluster indexes may differ. There seems to be some agreement between the hierarchical and k-means clustering results, based on the Crosstab result. We will begin by computing the mean frequency values for each of the words within cluster 1. After this calculation, we will identify and present the top 6 words that occur with the highest frequency in that cluster. That's how we got the top 6 words in each cluster. We do the procedure two times by including the hierarchical clustering method and the K-means clustering method. For instance we can the top 6 words in cluster 1 using Hierarchical clustering are bush, democrat Kerry, state, presided, republican

Top 6 words using Hierarchical clustering:

```
Top 6 words in Cluster 1 :
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 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 :
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 4 :
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 5 :
democrat      12.38
parties        6.34
state          5.74
republican     5.64
senate         3.30
seat           3.14
dtype: float64
```

```
Top 6 words in Cluster 6 :
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 7 :
dean          12.309091
kerry          5.345455
democrat       3.545455
edward         2.818182
candidate      2.727273
gephardt       2.672727
dtype: float64
```

Top 6 words using K-means clustering:

```
Top 6 words in Cluster 1 :
november     10.370821
poll          4.844985
vote          4.428571
challenge     4.118541
bush          3.030395
democrat      2.869301
dtype: float64
```

```
Top 6 words in Cluster 2 :
dean        7.710526
kerry       5.184211
clark       3.046053
edward      2.901316
democrat    2.559211
poll        2.302632
dtype: float64
```

```
Top 6 words in Cluster 3 :
democrat      2.933884
republican    2.837466
elect         1.947658
state         1.947658
parties       1.652893
vote          1.592287
dtype: float64
```

```
Top 6 words in Cluster 4 :
bush        1.198212
kerry       0.824921
poll        0.720820
democrat    0.608833
general     0.512618
elect       0.473186
dtype: float64
```

```
Top 6 words in Cluster 5 :
democrat      14.681818
parties        6.454545
republican     5.886364
state          5.045455
seat           4.113636
senate         4.000000
dtype: float64
```

```
Top 6 words in Cluster 6 :
bush        9.032258
kerry       5.661290
poll        2.506452
presided    1.880645
democrat    1.438710
campaign    1.400000
dtype: float64
```

```
Top 6 words in Cluster 7 :
bush              3.687879
iraq              3.587879
war               2.603030
administration    2.093939
american          1.593939
presided          1.351515
dtype: float64
```

D. Comparison of KMeans and Hierarchical Clustering

Using Cross tab:

K-Means Cluster	0	1	2	3	4	5	6
Hierarchical Cluster							
0	0	3	91	1487	0	58	122
1	3	4	2	38	0	114	6
2	324	0	0	0	0	0	0
3	1	91	249	359	7	90	6
4	1	0	9	0	35	4	1
5	0	0	12	18	1	44	195
6	0	54	0	0	1	0	0

```
Hierarchical Cluster that best corresponds to K-Means Cluster 2: 3
Hierarchical Cluster that best corresponds to K-Means Cluster 3: 0
```

GitHub link for the Project

Repository link: Link

Conclusion:

In this project, we used k-means clustering and hierarchical clustering to group blog posts or news articles from the Daily Kos political blog. We used both hierarchical and KMeans methods to extract the top six words from each of the seven clusters that were provided in the problem statement. Overall, we were able to identify significant themes and subjects in the Daily Kos political blog and group-related articles based on their content using our clustering analysis. This could be useful for conducting further research on topics or issues or for organizing and presenting articles to readers.

References

https://scikitlearn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec