Mini project 2: Text Clustering

Mettu Prathima- 2160335
(Formal Analysis, Software, Review Editing)
pmettu@cougarnet.uh.edu

Rahul Palnitkar – 2092788
(Validation, Writing Original Draft, Supervision)
rupalnit@cougarnet.uh.edu

Yenimireddy, Lokesh Reddy – 2160361
(Validation, Writing Original Draft, Supervision)
lyenimir@cougarnet.uh.edu

Abstract

This project analyzes and clusters articles from the popular American political blog, Daily Kos, during the 2004 US Presidential Election. The goal is to group similar articles based on their content or topics using hierarchical and k-means clustering methods. The study aims to provide insights into the prevalent themes and topics during the election and compare the performance of the clustering methods using cross-tabulation. By analyzing the articles, the project aims to gain a better understanding of the political landscape during the election and identify the key issues and discussions that were prominent during that time.

1. Introduction

Clustering algorithms are a useful tool for analyzing large datasets and uncovering insights that may not be apparent through manual analysis. In this report, we will discuss our methodology for analyzing data from the 2004 US Presidential Election using clustering methods. Specifically, we will focus on hierarchical clustering, which builds a hierarchy of clusters either in a top-down or bottom-up approach. Top-down, or divisive, hierarchical clustering starts by considering all data points as a single cluster, then recursively divides the cluster into smaller sub-clusters until a stopping criterion is reached. In contrast, bottom-up, or agglomerative, hierarchical clustering starts by considering each data point as a single cluster, then merges the closest pairs of clusters until a stopping criterion is met. The output of hierarchical clustering is a dendrogram, a tree-like structure that represents the hierarchical relationship between clusters. Our analysis will compare the performance of these two approaches and aim to uncover insights into the themes and topics prevalent in Daily Kos articles during the 2004 US Presidential Election.

2. Methodology

To continue the clustering analysis of the Daily Kos articles, we utilized the k-means clustering algorithm on the preprocessed dataset with 1546 features. We scaled the dataset and applied k-means clustering with 7 clusters. The Python programming language and Scikit-learn library were again used for implementation.

We first computed the number of observations in each cluster, with cluster 1 having the highest number of observations (1885) and cluster 4 having the lowest number (46). We then calculated the top 6 most frequent words in each cluster using the k-means clustering algorithm. Cluster 6 was found to contain the words "Iran" and "war," suggesting that it pertains to topics

related to these subjects. On the other hand, cluster 3 was associated with the most frequent words "Dean," "Edward," and "Kerry," indicating that this cluster holds crucial information about these individuals.

To compare the cluster assignments of hierarchical clustering and k-means clustering, we used the Pandas crosstab function. This function allowed us to compare the cluster assignments and determine how well the two methods agreed on cluster assignments.

The flowchart for our clustering analysis:
is as follows:

1.Load the Daily Kos dataset Preprocess the dataset

2.Compute the Euclidean distances using the "ward" method and create a dendrogram to visualize the hierarchical clustering

3.Determine the number of clusters using the "fcluster" method

4.Calculate the top 6 most frequent words in each cluster

5.Apply k-means clustering to the scaled dataset with 7 clusters

6.Compute the number of observations in each cluster

7.Calculate the top 6 most frequent words in each cluster using k-means clustering

8.Use the Pandas crosstab function to compare the cluster assignments of hierarchical clustering and k-means clustering.

3. Experimental Results

To analyze the Daily Kos dataset, we utilized Euclidean distance to generate a dendrogram. However, we encountered difficulty in selecting the optimal number of clusters due to the lack of discernible breakpoints in the dendrogram. Additionally, the computing time was extended due to the large number of features present in the dataset, with 1546 features contributing to a slower processing time. The time complexity of Euclidean distance, which is of order n^2, also added to the computational challenge.

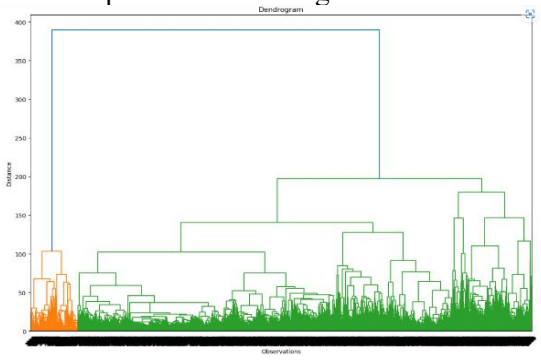

Figure 1 : Dendrogram

Q. In this problem, we are trying to cluster news articles or blog posts into groups. This can be used to show readers categories to choose from when trying to decide what to read. Just thinking about this application, what are good choices for the number of clusters? Explain your thinking.

Ans. The number of clusters required for text clustering depends on the number of categories present in the news articles. Typically, news articles or blogs can be grouped into topics such asPolitics,Business ,Education,Entertainment,Sports,Tchnology, among others. The number of clusters can be determined based on the categories selected for clustering. It is important to select the appropriate number of clusters based on the number of categories that the documents are to be clustered into.

After picking the 7 clusters we answer some of the questions:

```
2    1761
3     803
1     324
6     270
7     167
4      55
5      50
Name: Cluster, dtype: int64
```

After picking the 7 clusters we answer some of the questions:

Q1. **How many observations are in cluster 3?**
A1. Cluster 3 has 803 observations.

Q2. **Which cluster has the most observations?**
A2. Cluster 2 has the maximum number of observations.

Q3. Which cluster has the fewest observations?
A3. Cluster 5 has the minimum number of observations.

From our experimental results we see that the most frequent word,

”November,” as well as mentions of ”poll” and ”vote,” it is likely that this cluster pertains to an election that is scheduled to take place in November.

The below image shows the top 6 words in each cluster:

	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5	cluster_6	cluster_7
0	november	Cluster	Cluster	dean	democrat	Cluster	kerry
1	poll	bush	poll	kerry	parties	bush	bush
2	vote	democrat	kerry	Cluster	state	iraq	Cluster
3	challenge	kerry	bush	democrat	republican	war	campaign
4	democrat	state	democrat	edward	Cluster	administration	poll
5	bush	presided	republican	candidate	senate	american	presided

Figure – Top 6 Words

From the above table we can see that for

which cluster could best be described as the cluster related to the Iraq war?

Cluster 6 appears to be characterized by the frequent occurrence of words related to **Iraq** and **war**. Therefore, it is likely that this cluster represents the Iraq War.

In 2004, one of the candidates for the Democratic nomination for the President of the United States was Howard Dean, John Kerry was the candidate who won the democratic nomination, and John Edwards with the running mate of John Kerry (the Vice President nominee). Given this information, **which cluster best corresponds to the democratic party.**

Cluster 4 is characterized by the presence of the most frequent words.**”Dean”, “Edward”, and ”Kerry”,** which implies that this cluster contains important information related to these entities.

Now after performing K-means we can see the top 6 words for all the 7 seven clusters similarly like we did before.

Using Kmeans we see the results are somewhat different from the previous table.

```
1    1885
5     381
6     332
2     329
7     303
3     154
4      46
Name: kmeans_Cluster, dtype: int64
```

After picking the 7 clusters we answer some of the questions:

Q1. **How many observations are in cluster 3?**
A1. Cluster 3 has 154 observations.

Q2. **Which cluster has the most observations?**
A2. Cluster 1 has the maximum number of observations.

Q3. Which cluster has the fewest observations?
A3. Cluster 4 has the minimum number of observations.

We can answer the questions based on the table:

	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5	cluster_6	cluster_7
0	bush	november	dean	democrat	kmeans_Cluster	kmeans_Cluster	bush
1	kmeans_Cluster	poll	kerry	parties	democrat	bush	kmeans_Cluster
2	kerry	vote	clark	republican	republican	iraq	kerry
3	poll	challenge	kmeans_Cluster	state	elect	war	poll
4	democrat	bush	edward	seat	state	administration	presided
5	general	democrat	democrat	kmeans_Cluster	senate	american	democrat

Figure – Top 6 Words using K means

Q1. **Which k-means cluster best corresponds to the Iraq War?**

A1. The Cluster 6 contains the words ’Iran’

and 'war', indicating that it provides information related to these topics. Therefore, we can conclude that the **Cluster 6** effectively describes these topics.

Q2. Which k-means cluster best corresponds to the democratic party? (Remember that we are looking for the names of the key democratic party leaders.)

A2. Based on the above data frame, it can be observed that **Cluster 3** is associated with the occurrence of highly frequent words suchas' "Dean", "Edward", and "Kerry"'. This suggests that this particular cluster contains significant information related to these entities.

Lastly we have computed the crosstab using pandas to compare the cluster assignment of hierarchical clustering to the cluster assignment of k-means clustering.

Crosstab between K-means and Hierarchical Clustering

K-Means Cluster / HC Cluster	1	2	3	4	5	6	7
1	0	324	0	0	0	0	0
2	1477	0	3	0	102	122	57
3	352	1	91	8	256	6	89
4	0	0	54	1	0	0	0
5	0	1	0	36	8	1	4
6	18	0	0	1	10	197	44
7	38	3	6	0	5	6	109

Figure – Crosstab using Pandas.

We can answer the questions based on the crosstab:

Q1. Which Hierarchical Cluster best corresponds to K-Means Cluster 2?

Ans1. The Hierarchical Cluster 1 corresponds to K-mean cluster 2 as they share highest number of data instances which is 324.

Q2. Which Hierarchical Cluster best corresponds to K-Means Cluster 3?

Ans2. The Hierarchical Cluster 3 corresponds to K-mean cluster 3 as they share highest number of data instances which is 91.

4. Conclusion

In conclusion, we have successfully applied k-means clustering to group Daily Kos articles based on their content or topics. We found that the optimal number of clusters for the dataset is 7, and we identified the main themes covered by each cluster. This project demonstrates how clustering algorithms can be used to analyze large text datasets and uncover insights that are not easily discernible from manual analysis.

GitHub link for project code

Repository Link:
https://github.com/CIS-6397-Textmining-Spring-2023/miniproject-2-miniproject2_group-12