TEXT CLUSTERING

CIS 6397-Text mining

By Naga Lakshmi Prasanna Gottipati (Conceptualization, Coding, Software, Report editing), Leena Reddy Devireddy (Formal analysis, Code validation, writing original draft), Sudheer pattamsetti (Coding ,Report review and editing)

GitHub Repo

Abstract:

In our widespread exam of the Corpus documents, we embarked on a radical exploration of the text data to unveil the distribution frequency of words contained within those documents. We harnessed the capabilities of NLTK and Python to conduct a meticulous analysis of the content, with a keen consciousness on figuring out and documenting prevent phrases. This system became an indispensable factor of our take a look at, dropping mild at the intricacies of the textual records handy.

Introduction:

Clustering is a fundamental concept that has received considerable attention from researchers in the fields of pattern recognition, statistics, and machine learning It includes the category of unsupervised learning, where no labeled training data is provided for model construction. Instead, the goal of clustering is to group data points based on their intrinsic similarity. Within a group, data points are more similar than those in different groups. This process is often called unsupervised classification, because it produces results similar to classification algorithms but without the previously defined categories.

At its core, clustering algorithms work to analyze a data set and reveal the unique clusters within it. Areas in this category include psychology, entrepreneurship and retail, computational biology, social media network analysis, and many others Clustering plays an important role in discovering hidden patterns and patterns in data, making it a valuable tool in a variety of industries and industries.

1. **Hierarchical Clustering:**

We used hierarchical clustering as our approach to efficiently cluster information. The main objective was to collect data with shared lexical structures, and to achieve this we used aggregate hierarchical clustering a bottom-up approach This approach starts with many small clusters and gradually accumulates for a deep cluster. It effectively creates a hierarchy of labeled clusters, based on similarities between data points or observations. This process involves repeated clustering or merging, starting with individual data points and extending to entire data sets.

Our specific application involves using word frequency vectors to classify news and blog posts into groups using hierarchical clustering. The main result was the generation of dendrograms, a visual representation of the hierarchy of these clusters, which helped us determine the appropriate number of clusters for further analysis.

2. **K -means Clustering:**

In content analysis of the political blog Daily Kos, K-means clustering, a powerful unsupervised machine learning tool, was used to cluster relevant articles based on their word frequency profiles on. Then, each point is routed to the nearest hub using a distance measurement such as Euclidean distance. Each time the focus area is updated by accounting for all observations. This process of activation and modification is repeated until specific consistency criteria are met, such as a predetermined number of repetitions or a small change in focal areas. When the algorithm converges, each case is divided into one of K groups based on proximity to a particular focal point. In this work, K was set to 7, resulting in seven clusters. These groups were further analyzed for characterization, including common terminology.Additionally, group assignments were compared with the results of the hierarchy, providing valuable insight into the structure of the data.

3. Euclidean Distance

Here, the metric Euclidean distance served primarily to measure the similarity or dissimilarity of two data points in multidimensional space It was used to quantify dissimilarity in two texts or documents by observing the number of words in each category on the Both clustering algorithms were developed . Unlike k-means clustering, which divides data points into a fixed number of clusters by minimizing the squared difference between each point and its cluster mean, hierarchical clustering takes a different approach to data clusters points into nested clusters based on their inter-distances , forming a hierarchical structure that reveals the relationships between data points in a more nuanced way.

2. Methodology

2.1 Data Visualization:

By visualizing the dendrogram, we observed the spread of the different data points, revealing the strong correlations between them. It should be noted that the distance measure used to calculate pairwise distances in this study was the Euclidean distance. Given the size of the data, each variable corresponds to a unique number of terms in each case, the process of calculating these distances is inherently time-consuming Indeed, considering that n(n-1. There are 1)/2 pairs to evaluate a , while the complexity in distance computation is O(n^3). This highlights the statistical challenges posed by high-dimensional data, and highlights the need for effective methods for processing and interpreting such complex data.

2.2 Data Exploration:

In our analysis, we configured both Hierarchical and K-means clustering on the data set as specified in the problem statement. To get meaningful insights from the clusters, we identified the top 6 terms associated with each cluster. However, where the context from these words was not immediately apparent, we also considered using word clouds, which can provide a more visual and convenient representation of word frequencies To be with our legislation clearer and more structured, we assigned specific functions to private enterprise. The results of our clustering efforts and the lexical associations of each cluster are presented in the following section, providing a more general view of our findings.

3 Experimental Results:

3.1 Data analyzing using Hierarchical Clustering:

After examining the value counts for hierarchical clustering, it became evident that Cluster 2, housing 1761 articles, stood out as the most extensive cluster. Following this, Cluster 1 contained 324 articles, and Cluster 3 held 803 articles. Smaller in size, clusters 6, 7, 4 and 5 contained 270, 167, 55, and 50 articles, respectively. These findings indicate that Cluster 2 encompasses the majority of the articles within the dataset. Although smaller in size, the articles in the other clusters may exhibit differences when compared to those in the largest cluster. Further exploration, such as examining the topics and keywords associated with each cluster, could provide more insights into the nature of the articles that constitute each cluster."

```
Out[11]: 2    1761
         3     803
         1     324
         6     270
         7     167
         4      55
         5      50
         Name: Cluster, dtype: int64
```

3.2 Top 6 words using Hierarchical Clustering:

In cluster 1:

```
november     10.376543
poll          4.851852
vote          4.376543
challenge     4.104938
democrat      2.858025
bush          2.858025
dtype: float64
```

In cluster 2:

```
bush          1.546281
democrat      0.659852
kerry         0.607609
state         0.542873
presided      0.526973
republican    0.519591
dtype: float64
```

In cluster 3:

```
poll          2.429639
kerry         2.012453
bush          1.922790
democrat      1.823163
republican    1.328767
elect         1.165629
dtype: float64
```

In cluster 4:

```
dean          12.309091
kerry          5.345455
democrat       3.545455
edward         2.818182
candidate      2.727273
gephardt       2.672727
dtype: float64
```

In cluster 5:

```
democrat      12.38
parties        6.34
state          5.74
republican     5.64
senate         3.30
seat           3.14
dtype: float64
```

In cluster 6:

```
bush              4.777778
iraq              3.425926
war               2.470370
administration    2.225926
american          1.633333
presided          1.488889
dtype: float64
```

In cluster 7:

```
kerry       8.101796
bush        7.574850
campaign    1.862275
poll        1.736527
presided    1.616766
democrat    1.389222
dtype: float64
```

3.3 Data analyzing using k-means Clustering:

The majority of articles were assigned to Cluster 7, with a total of 1894 articles. Following this were clusters 4 (383 articles), 1 (364 articles), and 6 (330 articles) in terms of value counts for the k-means clusters. Clusters 3, 2, and 5 had fewer articles, with 255, 158, and 46 articles, respectively. These findings suggest that the dataset possibly contains distinct sets of articles differentiated by their content. A subsequent analysis of these clusters could reveal patterns related to the subjects or themes covered within the articles.

```
7    1894
4     383
1     364
6     330
3     255
2     158
5      46
Name: kmeans_Cluster, dtype: int64
```

3.4 Top 6 words using k-means Clustering:

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

3.5 Top 6 words using Hierarchical clustering & K-means Clustering:

The results obtained from these two clustering techniques are remarkably consistent. While cluster indexes may vary, the word compositions of most clusters are strikingly similar. This suggests a degree of agreement between the hierarchical and k-means clustering outcomes, as indicated by the Crosstab result. Our next step is to compute the mean frequency values for each word within Cluster 1. Following this calculation, we will identify and present the top 6 words with the highest frequency in that cluster. This process is repeated twice, utilizing both the hierarchical clustering method and the k-means clustering method. For example, the top 6 words in Cluster 1, obtained through hierarchical clustering, are: 'democrat,' 'republican,' 'elect,' 'state,' 'senate,' and 'parties'.

3.6 Comparison of KMeans and Hierarchical Clustering:

By Using Cross tab:

K-Means Cluster	1	2	3	4	5	6	7
HC Cluster							
1	0	0	0	0	0	324	0
2	91	4	75	96	0	0	1495
3	248	94	3	104	8	1	345
4	0	54	0	0	1	0	0
5	9	0	0	4	36	1	0
6	14	0	174	60	1	0	21
7	2	6	3	119	0	4	33

4 Conclusion:

Our main objective in this project was to optimize the news and blog posts from Daily Kos political blog. We achieved this by using hierarchical and K-means clustering methods, each resulting in 7 different clusters. For the two clustering methods, we identified and presented the top 6 terms characterizing each cluster. Through this analysis, we better identified important themes and topics in the Daily Kos political blog, which allowed us to segment relevant articles based on content. This new organization may be valuable for facilitating further research on specific topics or topics by providing readers with information in a more structured way. Our clustering methods provide a powerful tool for classifying and analyzing content in political media.

5 GitHub:

https://github.com/CIS-6397-Textmining-Spring-2023/miniproject2-miniproject2_group8-1