[MS4610 Course Project]

Hierarchical Clustering (Bonus Project Report)

Problem Statement

Given an Excel document about a collection of product reviews on a set of products collected from sources. As a student, you have to devise an unsupervised machine learning algorithm that clusters product reviews into various aspects (for example, price, utility, smell, etc.). It is a text mining-based problem. It is expected from the students that they design a 2-level hierarchical clustering.

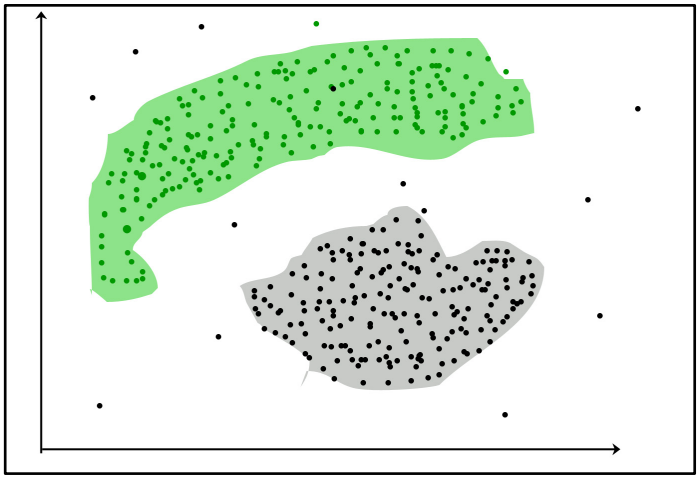
Terminologies

Sentimental analysis:

The process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.

Clustering:

It is basically a type of unsupervised learning method.
Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.



(https://media.geeksforgeeks.org/wp-content/uploads/clusteringg.jpg)

Types

Density-Based Methods: These methods consider the clusters as the dense region having some similarities and differences from the lower dense region of the space.

Partitioning Methods: These methods partition the objects into k clusters and each partition forms one cluster.

Hierarchical Based Methods: The clusters formed in this method form a tree-type structure based on the hierarchy. New clusters are formed using the previously formed one. It is divided into two category

- **Agglomerative** (bottom-up *approach*)
- **Divisive** (top-down *approach*)

Our Approach

Looking at the columns we can conclude that date of review, author of review, and similar other attributes **do not affect a review** and hence can be removed. Furthermore, sentiment, its score and confidence has already been given. Since all useful information has been extracted from reviews, we decided to remove these features like the original review statement as well.
Since confidence intervals are in order, therefore we label encoded them since algorithms can handle integer values easily.

Level 1 Clustering

Looking at the unique values of Aspect attribute we can safely conclude that aspect attributes form the first level of clustering.
We now have divided the rows with unique value of aspect attribute in separate data frames. All these data frames are stored in a list.

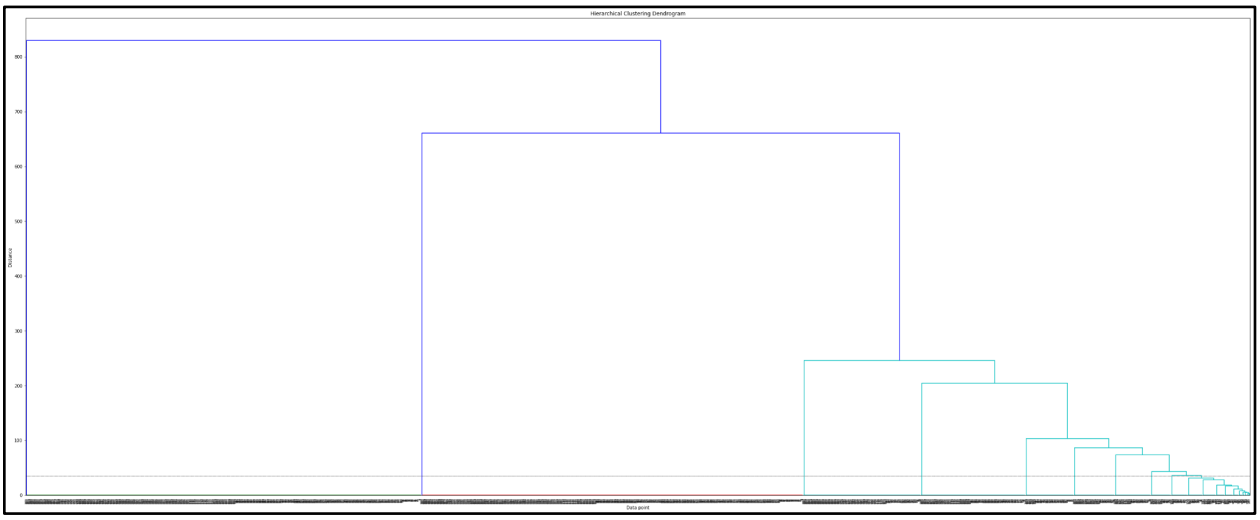
Level 2 Clustering

We can notice that values in the context aspect group represent sub features within a cluster. So, the new Clustering in each Data frame is done based on this context aspect.

Algorithm

As the columns like Aspect and Context Aspect are text, we need to convert them to a format that the computer understands.

- ➔ For each row, we first tokenize it using **nltk.WordPunctTokenizer().** We also make use of stopwords function to remove unwanted words like it, this, etc.
- ➔ We can use either **Countvectorizer()** or **TfidfVectorizer()** to form one hot encoded dummy column.
- ➔ We then use **Cosine Similarity** to find the similarity between the columns of above found vectorized dummy columns.
- ➔ We can use the **linkage** of the similarity matrix to view the dendrograms. The **dendrograms** show us the pictorial hierarchical representation, looking at which we can choose the distance to form the required number of clusters.



- ➔ In our approach we have used max_dist value as 35, which gives us **10 clusters** on the sample dataset provided, for level one clustering.
- ➔ After performing the above steps, a **csv file** containing the level 1 cluster labels is generated.
- ➔ After making 10 separate dataframes and performing above similar steps, we can get separate csv files for viewing each level 2 clustering.
- ➔ Finally, we append the all the above data frames and get our final result in **Final_Clusters.csv**

Conclusion

- ➔ After performing the above steps, we can see that for level 1 hierarchical clustering with respect to Aspect, **we get 10 different labels**, like one for pain and spasm, other for effectiveness, etc.
- ➔ In the further clustering with respect to Aspect Context, we get further clustering in each Cluster label in level 1. Like for **pain and spasm** we get further clusters like one for all pains, like **shoulder pain**; another for **knee pain**, etc.