820:Unsupervised Machine Learning
Spring 2021
Yapei Xiong
yxion013@bu.edu

Assignment#1 –Executive Summary

Overview

Upon receiving the dataset, I conducted the data exploration and did the data cleaning since the war dataset contains numerous missing values and duplicates values by each row that could affect the accuracy of our analysis. Before working on the afterward analysis, I need to clean the data and delete all null and duplicated values. Afterwards, I provided the client with some insights on the clusters and provided recommendations based on our findings.

Exploratory Data Analysis, Data Cleaning, and Preprocessing

The *forums.csv* dataset consists of 2362 rows and 301columns. Each row represents an individual case and the columns are values that mostly between01 and 1. 300 of the columns are numeric while 1 of them contains the textual data –the content of each message. Since there are duplicate text contents, I must delete those duplicated ones. After taking a look at all the columns, I decided to remove the following:

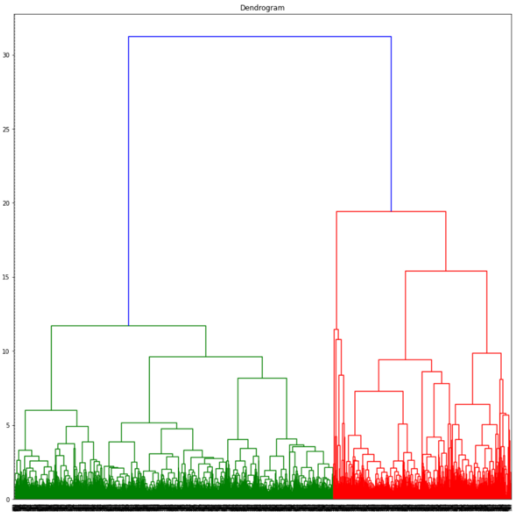
- The column of text; the textual column will affect our following analysis on all numerical models.
- 56 duplicated rows.

That leaves us with 300 numeric columns. Next, I indicate the missing values and duplicated values in the dataset. I found that 56 rows have at least one missing value or were duplicated values. Given that is almost 2.37% of the data, I could just simply drop all duplicated values to avoid the effeteness of them. Therefore, I then can fix the missing values column by column. After dropping all 56 duplicated or missing rows, there remain values of 2306 rows in 300 columns are the available data, I could approximate the values for all of them.

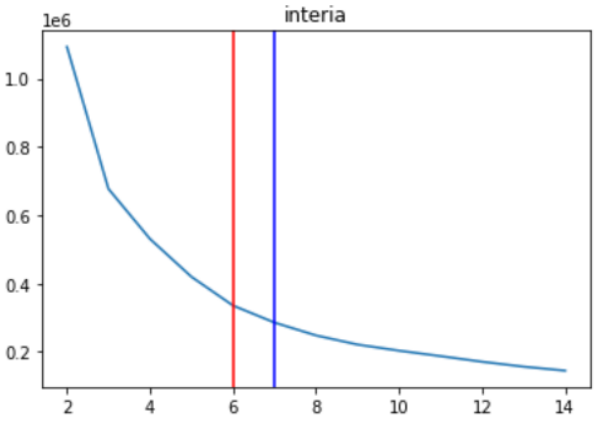
Hierarchical Clustering and K-Means Clustering

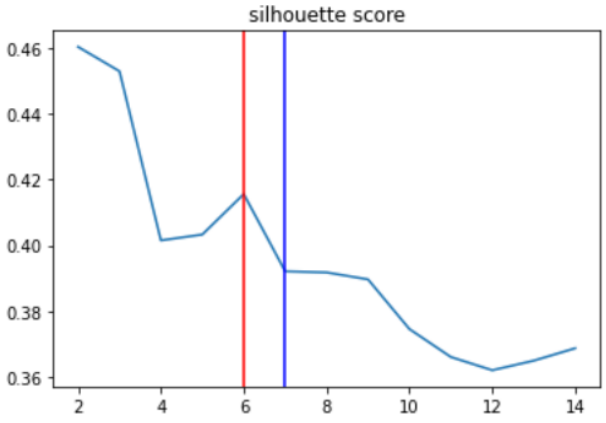
One of us in this analysis is to cluster all available values into like-groups, by doing so, I apply on two clustering techniques - Hierarchical Clustering and K-Means Clustering. By comparing the results of both clustering, I determine which is the best clustering method to match this dataset.

1. **I used a dendrogram to visualize the hierarchical relationship.** Since the dataset volume is huge, I tried multiple types of linkage methods and distance metrics. I choose to take the ward linkage method as the one that best reveals the hierarchical clustering in the plot on the right. This method plots more simple and concise clusters for the majority of messages. According to this plot, the dataset can be segmented into 2 or 3 clusters. However, since the values are numerous, each cluster could not have sufficient distance apart from each other.

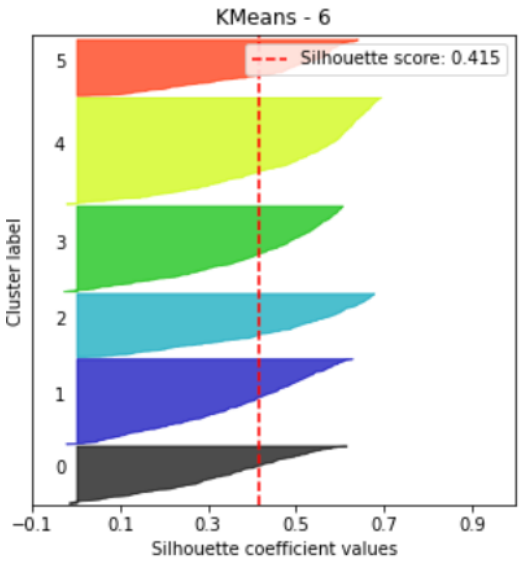


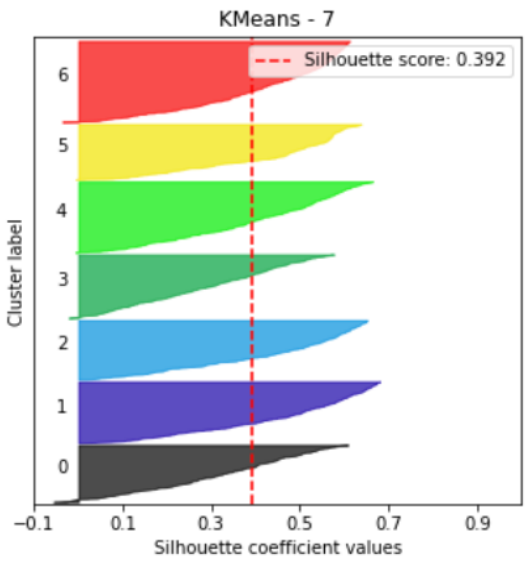
2. **Preparing for K-Means clustering, I reordered the intertie and the silhouette score.** I reduced the dimensions of dataset by applying PCA and T-NSE techniques. According to the inertia graph, the number of K cluster increases, the inertia decreases. Meanwhile, matching to the silhouette score graph and inertia graph, the most optimal numbers of cluster are 6 and 7 (the red line is when k=6 and blue line is when k=7), since the inertia was low and the silhouette scores are high on both cluster number.





3. Inspected the silhouette score for the dataset on both k numbers. We then can figure out the silhouette score. According to the graphs on next page, both graphs show that all values of the dataset are mostly equally been segmented into 6 or 7 clusters. When cluster is 6, the silhouette score is 0.415 and when cluster is 7, the silhouette score is 0.392. The larger silhouette score is better, because the larger silhouette score means that the clusters are denser and more nicely separated. Therefore, the cluster of 6 is better than the cluster of 7 in the K-Means clustering. It may demonstrate that the K-Means clustering method is the better clustering method for the company to categorize the dataset.





Conclusion and Recommendations

Lastly, after I did the data cleaning, reduced dataset dimensions by using PCA and T-NSE, categorized each message into a cluster, and checked the summary statistics and characteristics of each cluster, I look for establishing the optimal number of clusters for this dataset for the company.

Consequently, my recommendation for this project is that, the K-Means clustering is the best method to categorize the dataset and the optimal number of clusters is 6. In this way, based on this analysis the Hooli company can have the forum product to be categorized in 6 categories.