BA820: Unsupervised Machine Learning
Prof. Brock Tibert – 11.05.2022
Alexander von Schwerdtner

Assignment 1 – Executive Summary

Overview

To reimagine the platform on Hooli's recently acquired web-based user forum I conducted a cluster analysis using two techniques, hierarchical and k-means clustering, to provide a recommendation on how to categorize the user forum. I first cleaned and pre-processed the data before I performed the cluster analysis. I discovered that the data could potentially be segmented into 4 groups. Finally, I provided Hooli with insights on my analysis and provided recommendations based on my findings.

Exploratory Data Analysis, Data Cleaning & Pre-processing

The dataset *'forums'* consists of 2,304 rows and 301 columns. Each row represents a user's post from the forums message board. The '*text'* column is the only object containing the text of the post. The remaining 300 columns are numeric values representing details of the message as a set of numbers. After a closer look at the dataset, there were a couple of data cleaning and pre-processing steps that I ran:

1. Set the '*text'* column to the index so that cluster analysis is possible (*".set_index()")*
2. Made sure there is no missing *data ("isna().sum().sum()")*
3. Dropping 58 duplicates from the dataset *("drop_duplicates(inplace=True)")*
4. No scaling was needed as the data seems to be on the same scale after looking at the dataset using describe().T syntax and looking at the means and standard deviations of the columns

'Principal component analysis' on a dataset is only required when variables are highly correlated. *Figure 1* shows that the columns are not highly correlated. Even though it was evident that the dataset did not have to be scaled, I used the PCA method to see if it would influence the cluster analysis. I found that running PCA on the dataset did not change the outcome of clusters significantly.

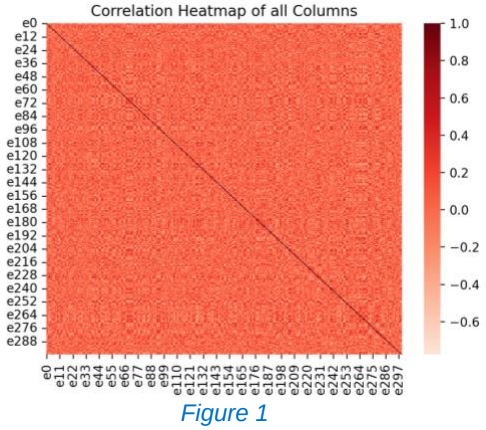

Figure 1

Hierarchical Clustering and K-Means Clustering

Since the business objective is to categorize the forum post data based on the theme of the discussion, I tried out two clustering techniques – *'Hierarchical Clustering'* and *'K-Means Clustering'* – to examine and analyze which method would yield better results. The following will run through the approach I took to categorize the forum posts:

BA820: Unsupervised Machine Learning
Prof. Brock Tibert – 11.05.2022
Alexander von Schwerdtner

1. **Running hierarchical clustering on 4 methods**

   To see which hierarchical cluster method would yield the best results I ran four different linkage methods: single, complete, average and ward. To visualize the hierarchical relationship, I used a dendrogram and plotted the four methods as visible in *figure 2.*

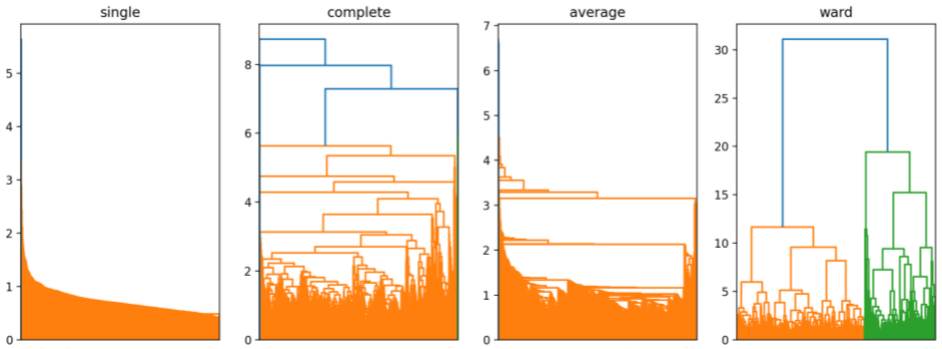

Figure 2

2. **Choosing the *'ward'* distance approach**

   After running the 4 distance approaches, I decided to stick with the Ward method as it seems to yield the best results in terms of clustering. Looking at the length of the vertical lines (visualizing the distances) I found that the dataset could potentially be segmented into 4 clusters, ensuring that the individual clusters are at a sufficient distance apart from each other. The black dashed line represents the scenario where I segment the data into 4 clusters. As visible in the dendrogram one of the clusters will end up being very small compared to the other clusters.

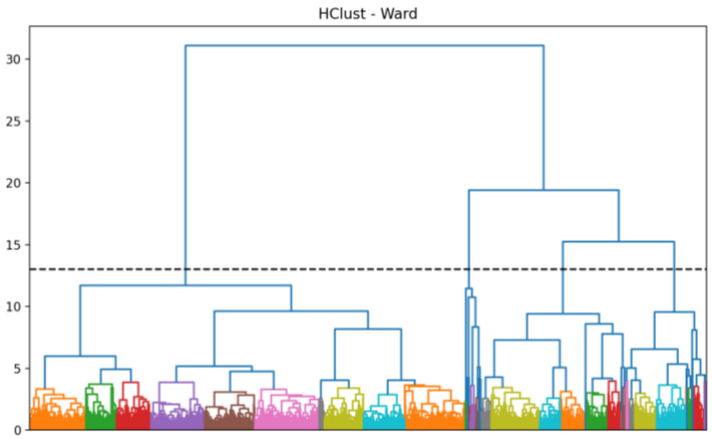

Figure 3

3. **Running K-Means recoding the inertia and average silhouette scores for different numbers of clusters (k=2-20)**

   The graph on the left of *figure 4* demonstrates how the inertia decrease as the number of clusters K increases, while the graph on the right shows how the average silhouette score decreases as K increases. From these two graphs I concluded that the optimal number of clusters is also 4 (vertical black dashed line in both graphs), where the inertia is reaches its elbow and the average silhouette score is still moderately high before it drops.

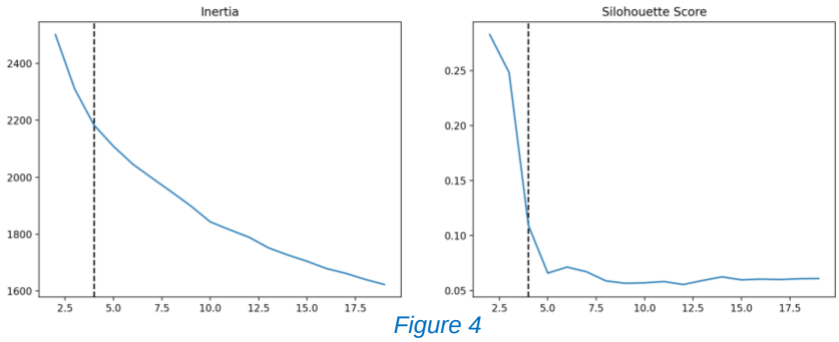

Figure 4

BA820: Unsupervised Machine Learning
Prof. Brock Tibert – 11.05.2022
Alexander von Schwerdtner

4. Comparing techniques

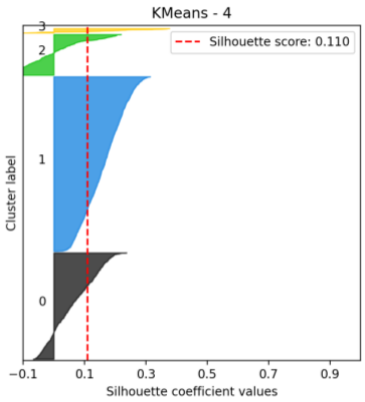

Figure 5

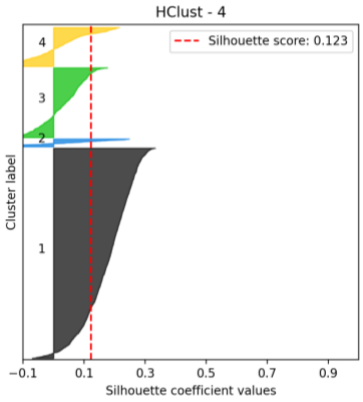

Figure 6

Figures 5 and 6 show that for both clustering techniques, most of the forum posts have a positive silhouette score within the clusters. It is worth noting that cluster 1 in the hierarchal approach has the majority of posts in its cluster with only a few negative silhouette scores and that for the K-means approach cluster 1 also has the majority of posts in its cluster with no negative silhouette scores overall. This could represent a common theme within the posts in the specific clusters. Looking at the overall silhouette score we can see that the hierarchical approach yields a slightly higher score with 0.123 compared to K-means with 0.111

5. Examining the distribution of posts in clusters based on the two methods

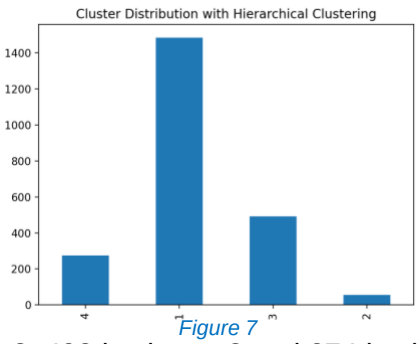

Figure 7

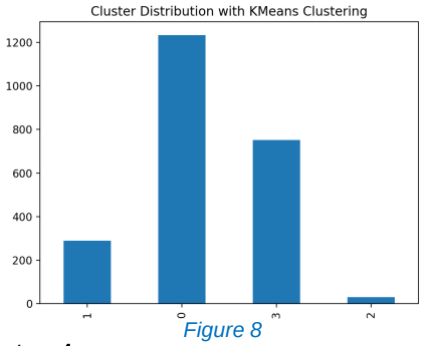

Figure 8

Having a closer look at both approaches I found that both result with one cluster that is highly populated with posts (observations) and one that only has few (*figure 7&8)*. In the example of the Hierarchical clustering the distribution of posts throughout the clusters is 1483 in cluster 1, 55 in cluster 2, 492 in cluster 3 and 274 in cluster 4.

Conclusion & Recommendations

After I looked at the distribution, I was curious to see if I could correlate different posts within the clusters. I profiled the clusters onto two copies of the original dataset and looked at individual posts within each cluster. I found several reoccurring themes within the specific clusters. For example, for the hierarchical cluster 1 a reoccurring theme were posts about space and the military, whereas for cluster 4 reoccurring themes were posts about software and electronics. To further dig into my analysis, I looked at the features of individuals clusters using looking at the mean, min, max, standard deviations, and specific percentiles to see whether I found similarities of columns. When plotting a correlation matrix on all observations of a given cluster, I found that observations where now somewhat correlated with one another.

Based on my analysis I recommend Hooli to categorize and segment their forum into 4 categories as a starting point. I would suggest using the segments based on the hierarchal technique as it has a higher silhouette score. From here I believe Hooli will have the ability to categorize their forum further in the future.