Project 1

STQD6114 – Unstructured Data Analytics

P152419 – Hazim Fitri Bin Ahmad Faudzi

Part 2 – Task 2

Fifty news articles were analyzed using unsupervised clustering to discover underlying thematic groupings. The articles were first transformed into high-dimensional vectors using the Term Frequency–Inverse Document Frequency (TF-IDF) representation, which weights words by their importance in each article. No predefined labels or categories were provided, so the clustering aimed to let the data reveal its own topic clusters. Three clustering algorithms with different approaches were applied which is K-Means , Hierarchical, and DBSCAN. These methods were chosen to compare how each partitions the articles and to identify prominent themes or topics in the news dataset. Each clustering result is discussed below, followed by a comparison of their performance in grouping articles and handling outlier articles.

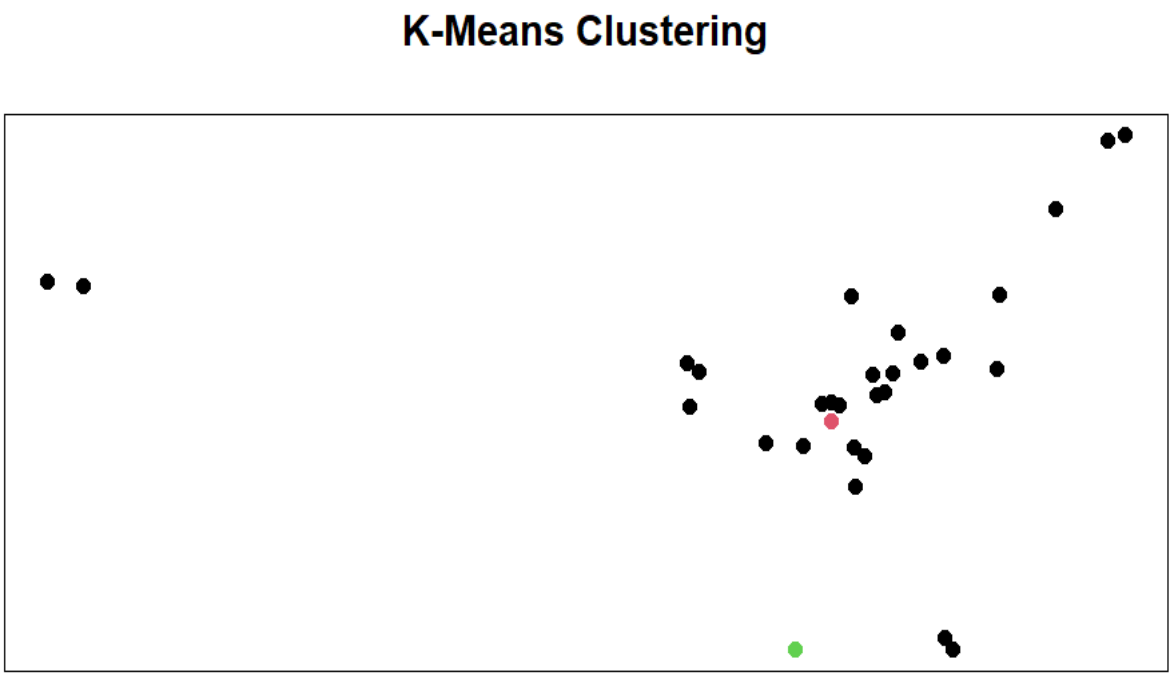

Figure 1 K-Means Clustering

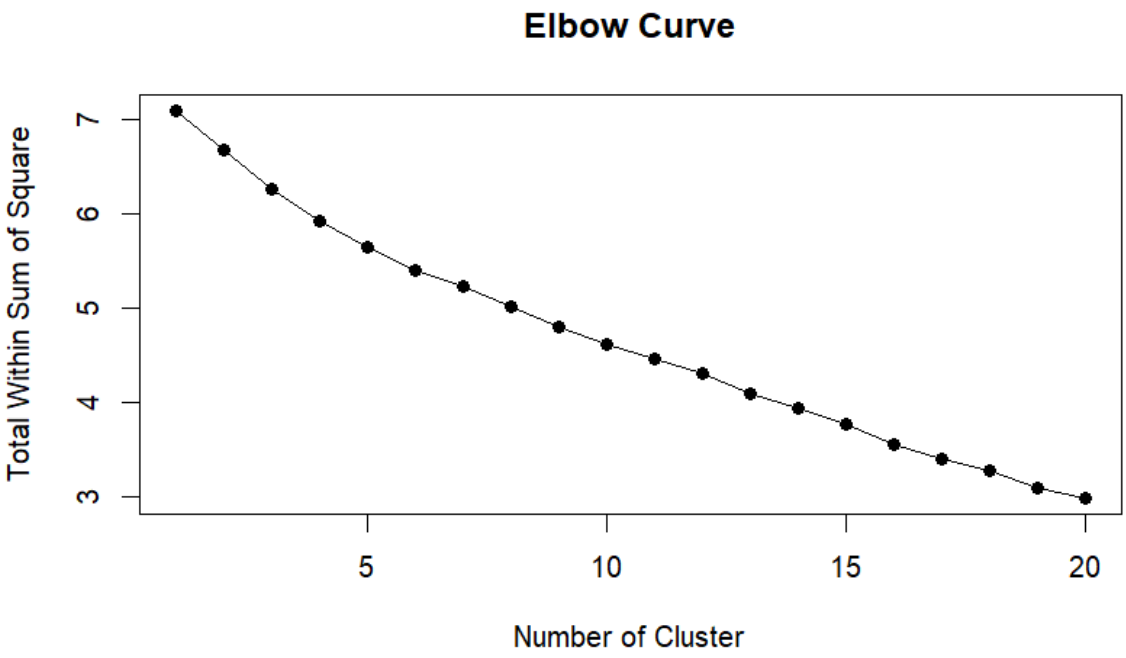

Figure 2 Elbow Curve of Hierarchical Clustering

K-means plot shows 3 clusters which denoted by pink, green, and black centroids. The group does not have any clear boundaries indicating that there might be come theme overlap across clusters. K-means assigns every document to a cluster, so any outliers are still attached to the nearest centroid. Elbow curve can be used to determine the best number of clusters for a K-means clustering method. However, elbow curve from figure 2 shows that the curve is almost linear. This means that the rate of decrease in the total within-sum of squares is the same as the rate of increase in the total number of clusters. To keep model interpretability and not computationally burden, we will keep the number of clusters to three. This will also help when we want to do comparison with other clustering method later.

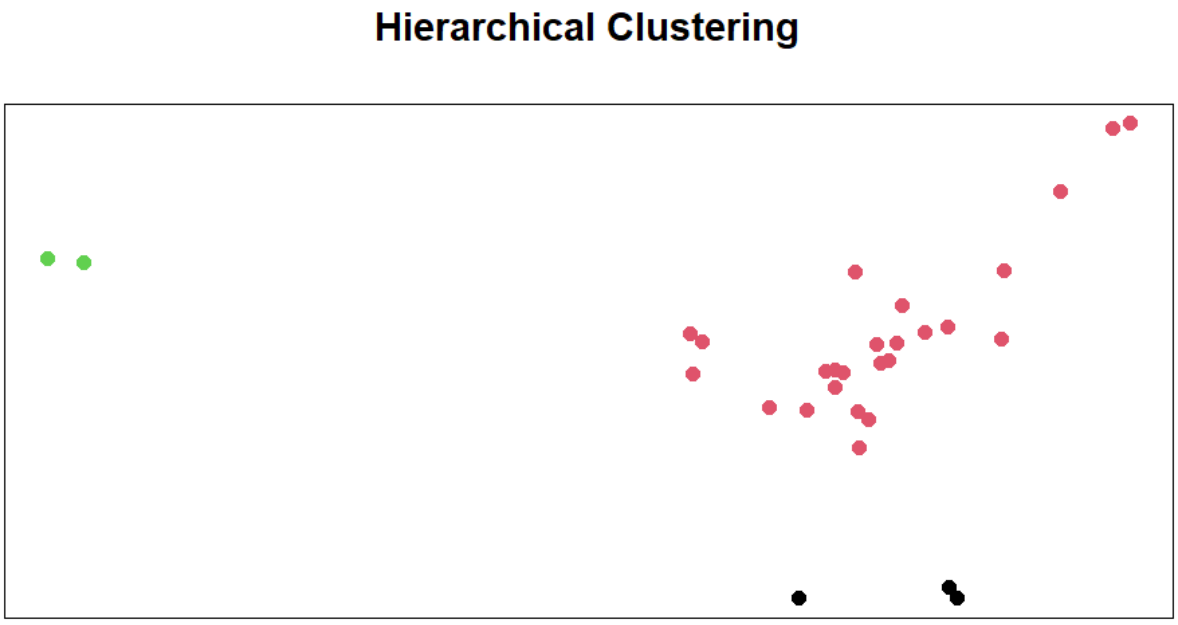

Figure 3 Hierarchical Clustering

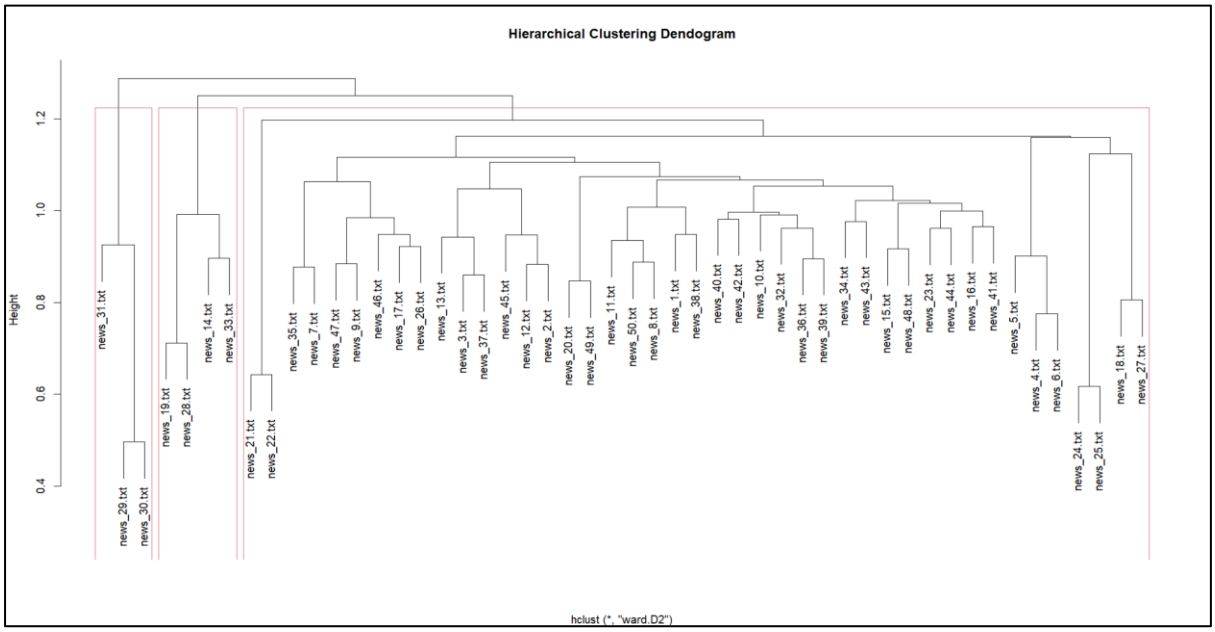

Figure 4 Dendrogram of Hierarchical Clustering

Hierarchical agglomerative clustering also shows 3 clusters which denoted by pink, green, and black centroids. However, hierarchical has a clear separation of clusters. The two points positioned on the left is being grouped into the green cluster while three points at the bottom

has been assigned to the black cluster and the rest is within the large red cluster at the centre. Next, figure 4 shows the dendrogram of of the hierarchical clustering where it can be seen that the smallest cluster is on the left with only three documents. The cluster next to it has just one more news document from the first cluster and the rest of the news are being categorized as a single cluster.

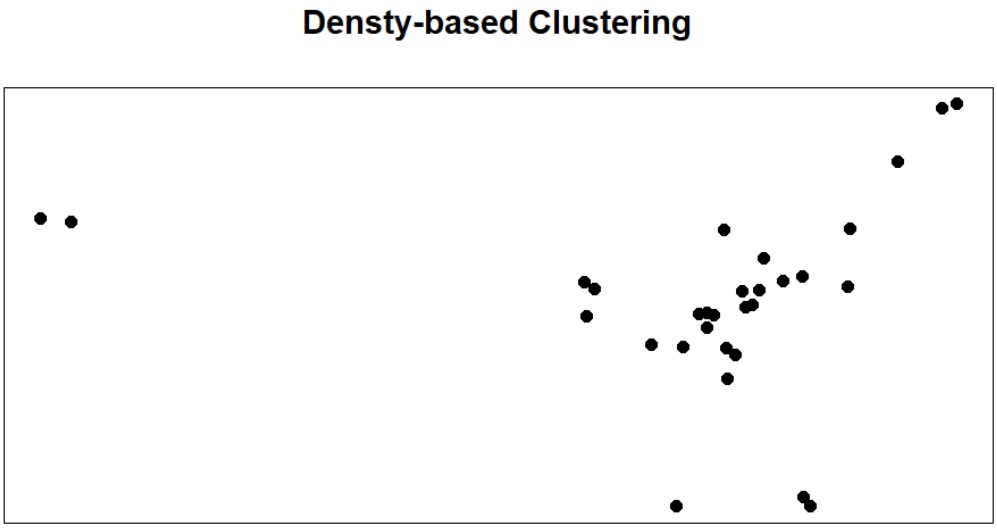

Figure 5 DBSCAN Clustering

From figure 3, we can see that there are no distinct coloured groups, no indicators of separate clusters, and no differentiation between core points. This means that Density-Based (HDBSCAN) clusters all the data points with black means that all data points are in a single cluster. This shows that HDBSCAN clustering is not an appropriate method to use for this kind of data as its clusters lack meaning. The only interpretation that can be made is that maybe HDBSCAN classify all data points as noise.

For comparison of performance between all three clustering methods, K-means shows three clusters but it was poorly separated while HDBSCAN only consists of one cluster, so it lacks meaning. On the other hand, Hierarchical clustering shows three clearly defined clusters.