STQD6114 Task 2

PANZHANGYU P136922

K-mean:


```
summary(kmeans_result)
           Length Class  Mode
cluster        50  -none- numeric
centers     15657  -none- numeric
totss           1  -none- numeric
withinss        3  -none- numeric
tot.withinss    1  -none- numeric
betweenss       1  -none- numeric
size            3  -none- numeric
iter            1  -none- numeric
ifault          1  -none- numeric
```

In this case, the fact that the bar with x=1 is much larger than that with 2 and 3 may indicate that one of the k-means clusters contains the majority of documents, while the other clusters are relatively small.

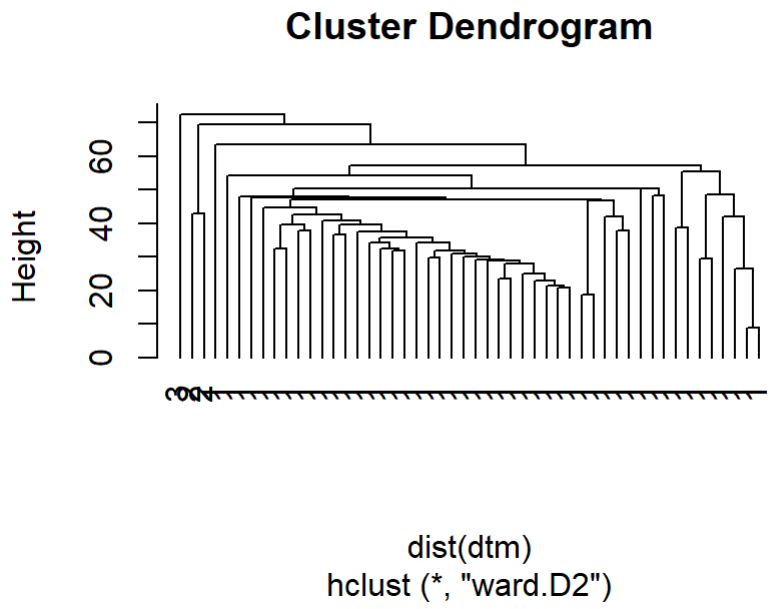
The summary information about kmeans_result shows some important statistics

about the clustering results, including the size of each cluster (size), the total sum of squared errors (tot.withinss), and the number of iterations of the algorithm (iter).

Hierarchical clustering:

```
> table(hierarchical_clusters)
hierarchical_clusters
 1  2  3
47  2  1
```

The results of hierarchical clustering show that there are 3 clusters with sizes of 47, 2, and 1. Specifically, 47 documents are classified into the first cluster, 2 documents are classified into the second cluster, and only 1 document is classified into the third cluster.



HDBScan:

```
> table(dbscan_result$cluster)

 0
50
```

All data points are assigned to a single cluster with cluster label 0. This means that under the current parameter settings, the HDBScan algorithm does not assign data points outside of any cluster, but treats them all as noise points.

If the density variation between data is not obvious enough, the HDBScan algorithm may regard all data points as noise points.