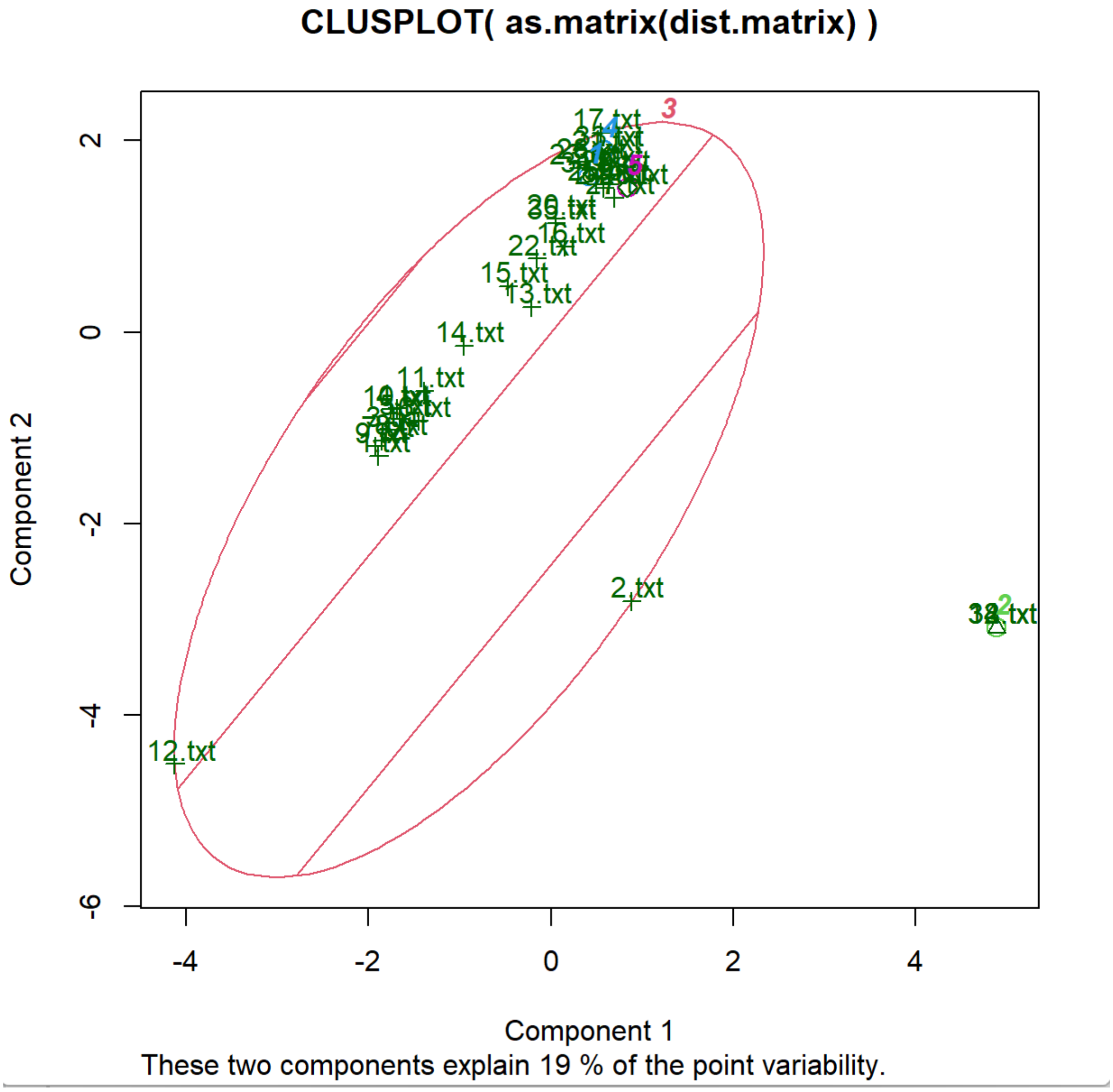


Fig.1 k-means

We specify the expected number of clusters for clustering to be 5. It can be seen that the distance between the 3 clusters is relatively large, and the error may be large. The 1, 4, and 5 clusters are densely distributed and close to each other. The 2 clusters are far away from other clusters, and the number is dense, which may be more accurate. The 3 clusters contain 1, 4, and 5 clusters, which means that we can also only divide into two clusters.

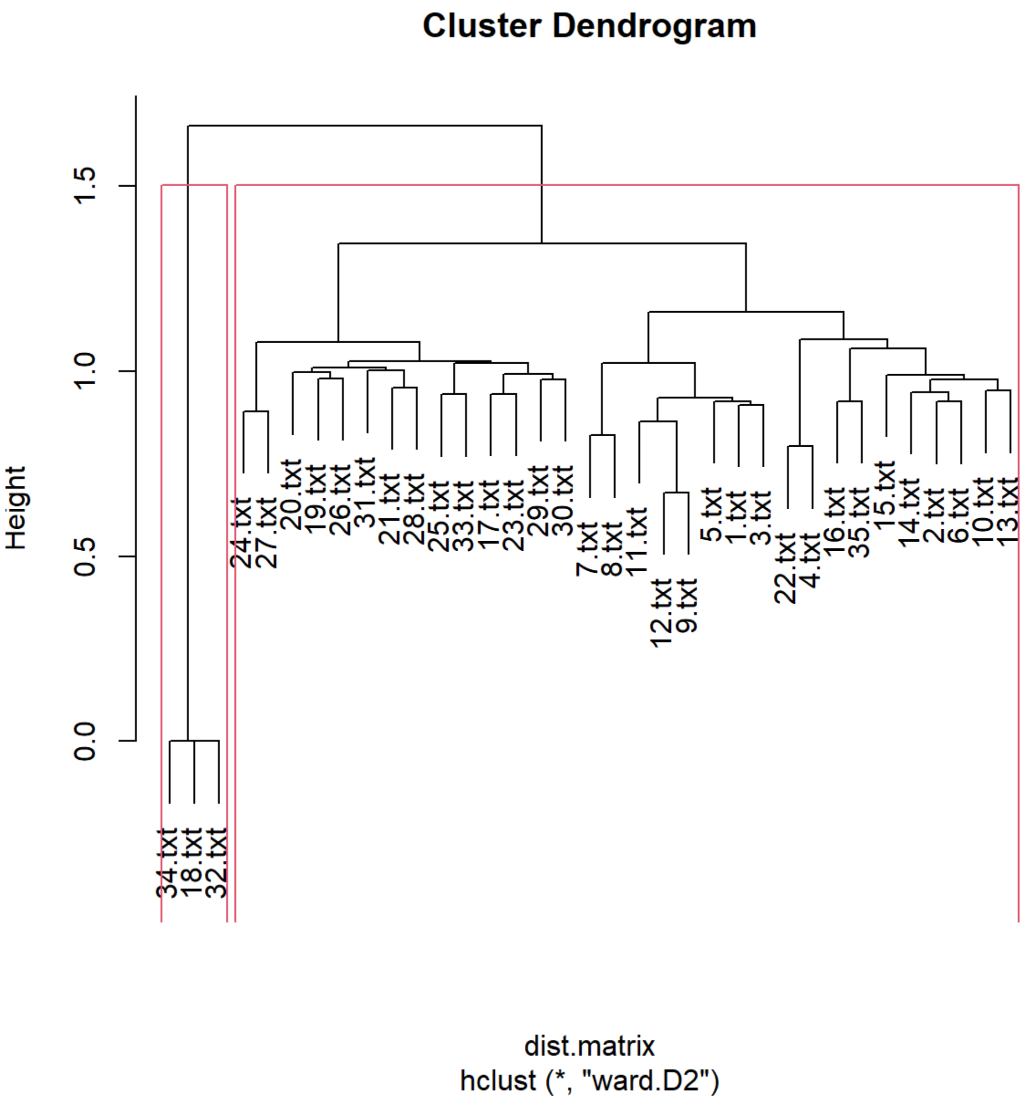


Fig.2 hierarchical

It can be seen from this figure that when height=1.5, the text is divided into 2 clusters, and when height=1.2, the text is divided into 3 clusters.

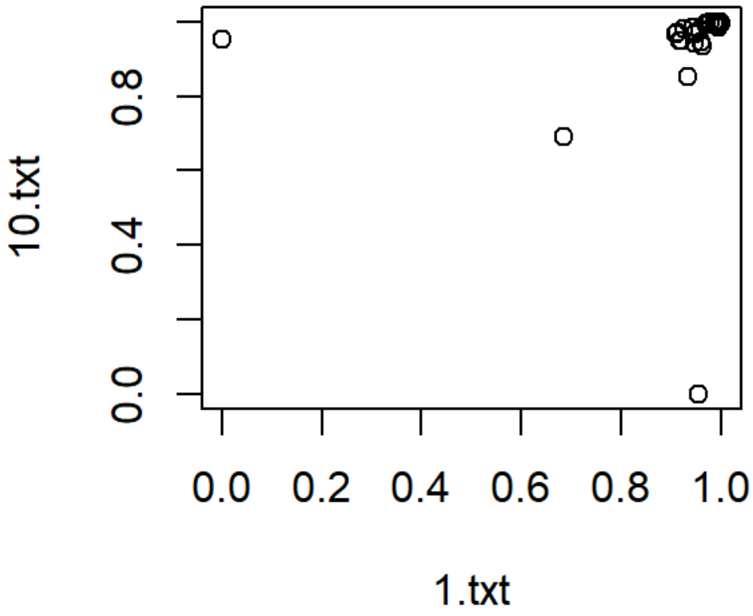
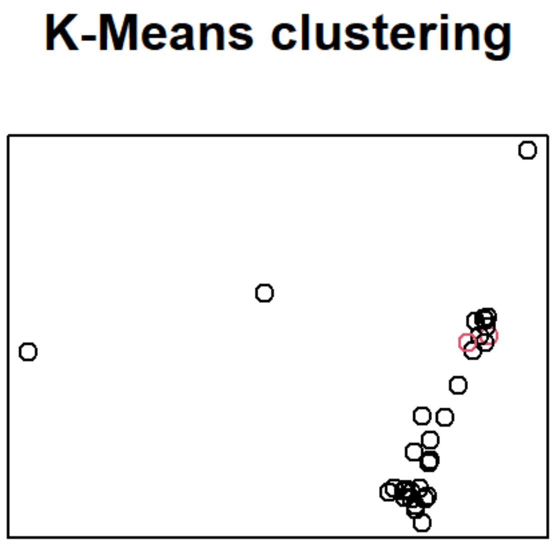
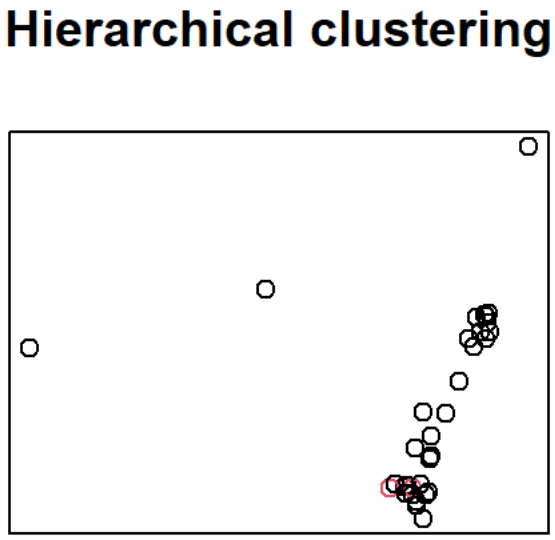


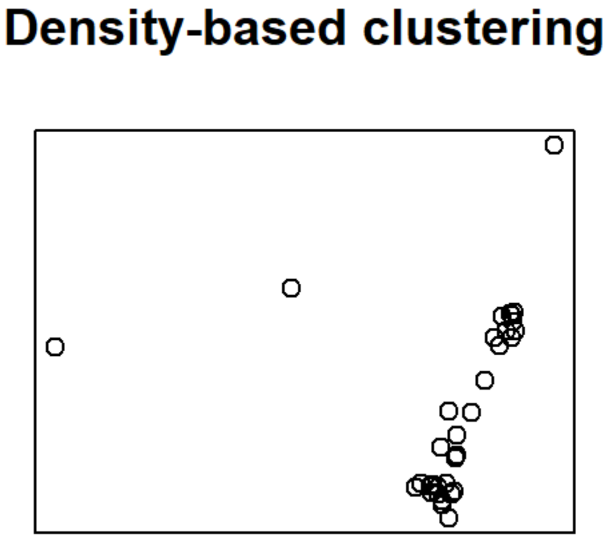
Fig.3 HDBScan

The densest clusters are in (1.0,0.9).



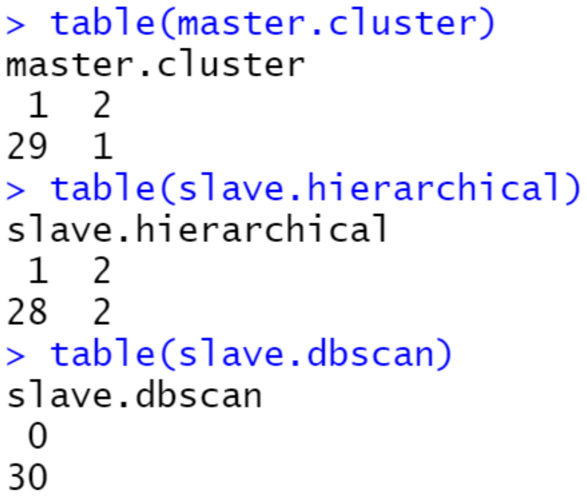


Hierarchical clustering



Density-based clustering

As can be seen from these figures, the k-mean method is better because it has the densest clusters and the least noise points.



```
> table(master.cluster)
master.cluster
 1  2
29  1
> table(slave.hierarchical)
slave.hierarchical
 1  2
28  2
> table(slave.dbscan)
slave.dbscan
 0
30
```

For "master.cluster", there is 1 cluster marked 1 and 29 clusters marked 2.
For "slave.hierarchical", there are 2 clusters marked 1 and 28 clusters marked 2.
For "slave.dbscan", there are 30 noise points, i.e. no clusters marked as 0.
According to these results, we can preliminarily judge that the "master.cluster" method divides the data into two clusters, while the "slave.hierarchical" method divides the data into two clusters, but more samples are divided into noise points. The "slave.dbscan" method did not form any clusters, all samples were marked as noise points.