Assignment 8- Text clustering

(a) Test 1:

	No processing	Preprocessing
K-smean	53/60	58/60

No processing:

	Group 1	Group 2	Group 3
Cluster 1	20	0	0
Cluster 2	5	14	1
Cluster 3	1	0	19

Preprocessing:

	Group 1	Group 2	Group 3
Cluster 1	19	1	0
Cluster 2	0	20	0
Cluster 3	0	1	19

Ans 1: Using text preprocessing gives a better result.

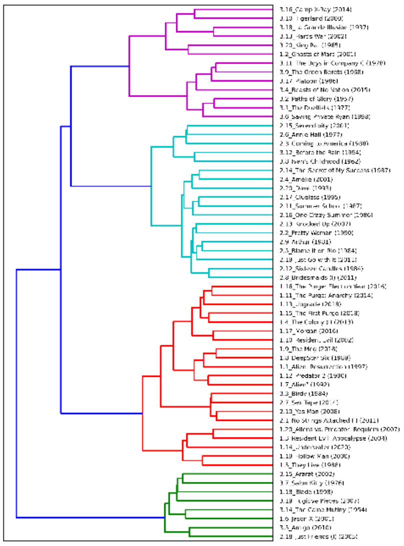
(b) Test 2:

(*Here we choose the **preprocessing text** to do this test.)
Here I set the max features in **100, 300, 500, 700, and None**.

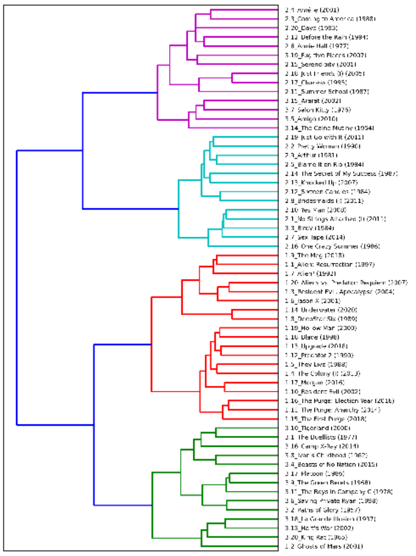
1. If max feature=100, purity=57/60.
2. If max feature=300, purity=54/60.
3. If max feature=500, purity=53/60.
4. If max feature=700, purity=58/60.
5. If max feature=None, purity=58/60.

Ans 2: In conclusion, by setting max feature as None can get the highest purity, as our data can be trained by more features, which gives us better result. Thus, the max feature for test 3 is None.

(c) Test 3:



(a. without preprocessing)

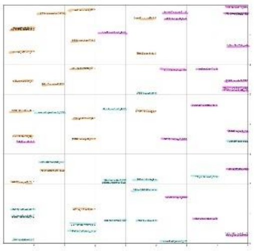


(b. Preprocessing)

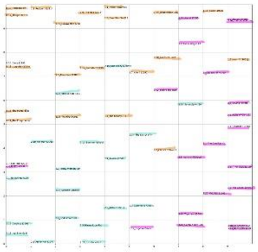
(c1) In my opinion, I think the case with preprocessing gives the better result. To begin with, in the picture a, we can see most of the movies in group 3 are classified in the purple and green cluster, however, these two clusters are the most irrelevant. On the other hand, we can see in the picture 2, we can observe that all the movies in the group 2 are classified in the blue and purple cluster, and these two cluster are the most relevant. Besides, we can see the red cluster contains almost all the movies in group 1(19 movies), and the green cluster contains 13 movies in group 3. Compares to picture a, picture b shows higher similarity within group and lower similarity between group, so **the case with text preprocessing gives better result**.

(c2) By using **K-mean**, the Pros is that we can easily get 3 clusters as we originally set, and we got better result in this case(without processing: 53/60, with processing: 58/60), while the Cons is the K-mean method spend more time(without processing: 21.01 sec, with processing: 18.45 sec).
By using **Hierarchical clustering**, the Pros in this case is less time used (without processing: 0.32 sec, with processing: 3.55 sec). However, the Cons is that the final output is quite different with original data(originally: 3 clusters, result: 4 clusters). Besides, without text preprocessing, the output will be more different than the original data. Though we can get better result after text preprocessing, the result is still not as good as the one of the K-mean. In conclusion, in this case, we should analyze our data by using K-mean, for we do not have large amount of data and we got higher purity by using K-mean method.

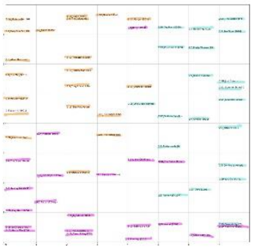
(d) Test 4:



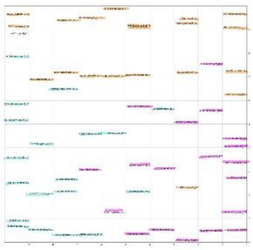
(without Preprocessing 8*8)



(without Preprocessing 10*10)



(with Preprocessing 8*8)



(with Preprocessing 10*10)

(orange: group 1, green: group 2, pink: group 3)

Without preprocessing:

(8*8 neurons): It spend 179.51 sec. I cannot directly separate 3 groups of movies, while I can basically distinguish the distribution of three clusters.
(10*10 neurons): It spend 248.23 sec. I can distinguish 3 groups of movies better than 8*8 neuron, and the distribution of clusters are clearer.

With preprocessing:

(8*8 neurons): It spend 78.53 sec. I can quickly separate 3 groups of movies and easily distinguish the distribution of three clusters.
(10*10 neurons): It spend 125.46 sec. I can also distinguish 3 groups of movies and figure out the distribution of cluster, while 8*8 can also do it well and spend less time in this case.
***Pros of SOM:** By using SOM, we can easily interpret and understand the distribution and similarities of the original data because of the reduction of dimensionality and grid clustering.
***Cons of SOM:** It requires necessary and sufficient data to develop meaningful cluster, or else the weight vector cannot successfully group and distinguish data.

Question 1:

The text preprocessing does affect the cluster performance. Before the preprocessing, we get lower purity and the results of cluster and SOM are not very clear. After the preprocessing, we get higher purity and the results of cluster and SOM become clearer.

Question 2:

When the max feature become larger, the purity also become higher.

Question 3:

Pros of unsupervised cluster method: We do not need do too much to our original data, and the algorithm will help us find the potential rules and association, then do the clustering.
Cons of unsupervised cluster method: It is possible to cause the features that are not that important be over-amplified, leading to the biased and meaningless results.