Project #2

Contributor:

Yuting Tang 304881011

Xiao PENG   005033608

Date:

Feb 4, 2018

Table of Contents

Introduction

Clustering algorithms are unsupervised methods for finding groups of data points that have similar representations in a proper space. One of the differences between clustering and some other classification is that there is not a priori labeling or grouping of the data points available. In other words, K-means clustering groups the data points into regions that are characterized by a set of cluster centroids iteratively. Then each data point could be assigned to the cluster with the nearest centroid.

In this project, we are requested to learn the proper representations of the data and to evaluate the performances of clustering algorithms. And the dataset we work with is “20 Newsgroups” which is a collection of approximately 20,000 documents partitioned across 20 different newsgroups according the different topics. Therefore, there are 20 classes for the dataset. So we should pretend as if the labels are not available and try to find the groupings of the documents. We could then use the class labels as ground truth to evaluate the performances of the clustering task.

Problem

Problem 1

Requirement: Finding a good representation of the data is fundamental to the task of clustering. Following the steps in Project 1, transform the documents into TF-IDF vectors. Use min df = 3, exclude the stopwords (no need to do stemming). Report the dimensions of the TF-IDF matrix you get.

Result: The shape of the TFxIDF is shown below.

```
The dimension of TF-IDF Vector with min_df=3 is (7882, 27768)
The number of Terms Extracted with min_df=3 is 27768
```

Problem 2

Requirement A: Apply K-means clustering with k = 2 using the TF-IDF data. In Problem 2, we are requested to apply K-means clustering with k = 2 using the TF-IDF data. In other words, we should group subclasses into two main classes, 'Computer technology' and 'Recreation activity'.

Result: Firstly, the contingency matrix is required to evaluate the performance of clustering and the result is as follows.

```
The contingency matrix:
 [[1716 2263]
 [   4 3899]]
```

As we have learnt before, the diagonal contingency matrix could be called the perfect result. However, it could be easily told that we could not reach to an almost diagonal matrix by only implementing K-means algorithms, as more than half of the documents whose true labels are 0 have been predicted as 1.

Requirement B: Compare the clustering results with the known class labels. Report the 5 measures above for the K-means clustering results you get.

In order to make a more concrete analyzation of the clustering result, there are various measures of the purity of a given partitioning of the data points with respect to the ground truth. Due to the reason discussed above, we introduced 5 other metrics to evaluate the performances of the clustering: the homogeneity score, the completeness score, adjusted rand score and adjusted mutual info score. As is illustrated in the project statement, the homogeneity score is a measure of how pure the clusters are. Completeness is satisfied if all the clusters contain only data points that belong to a single class. The V-measure is defined to be the harmonic average of homogeneity score and completeness score. The rand index computes similarity between the clustering labels and ground truth labels. Finally, the mutual info score measures mutual information between the cluster label distributions and the ground truth label distributions. All the above metrics are range from 0 to 1. The difference of these 5 metrics between the contingency metrics we used is that they do not evaluate the accuracy of the match of exact label number, while those scores measure the purity of each cluster.

Result: The result of clustering performance is shown as below.

```
Homogeneity: 0.253
Completeness: 0.335
V-measure: 0.288
Adjusted Rand-Index: 0.180
Adjusted Mutual-Index: 0.253
```

From the result we could tell that all 5 metrics are not satisfactory. The main reason for the bad performance might be that the K-means clustering usually fails to cluster high dimensional datasets. To be more specific, K-means algorithm is based on the distances between data points in space, while the distance in high dimensional space make little sense. As a result, it could barely show the most significant characteristics of each cluster.

Problem 3

Requirement A: Report the plot of the percent of variance the top γ principle components can retain vs. γ, for γ = 1 to 1000.

As is shown in metrics in Problem 2, the high dimensional sparse TF-IDF vectors do not yield a good performance in K-means clustering algorithm. Therefore, we need to find a better representation tailored to how the clustering algorithm works. In Problem 3, we are requested to do more preprocessing and dimensional reduction before we feed the data into K-means algorithm. In this part, truncated SVD and NMF are used to reduce the dimension. In addition, different values of component chosen for the dimension reduction are compared through 5 clustering metrics.

Firstly, we need to find the effective dimension of the data through inspection of the top singular values of the TF-IDF matrix. Thus, we plot the curve to show the percent of variance the top γ principle components can retain with different γ.

Result: The result is as following Figure 1.

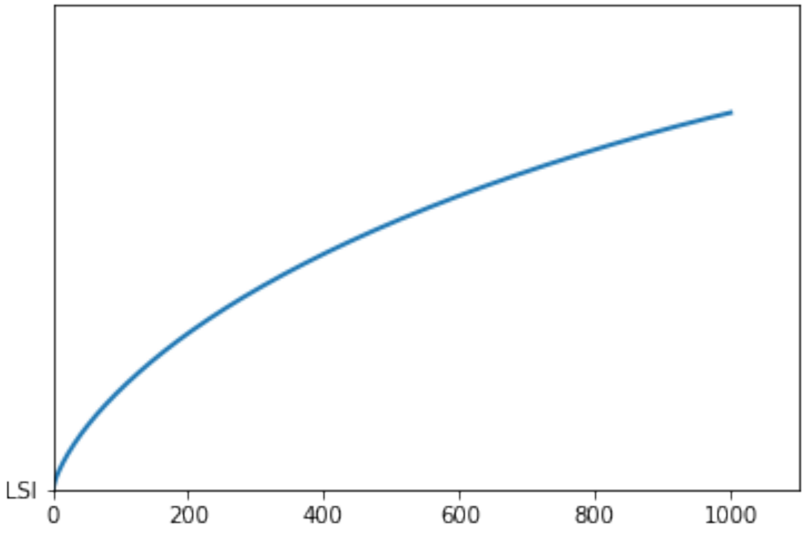

Figure 1: the percent of variance retained v.s. γ

Requirement B: Specifically, try γ = 1,2,3,5,10,20,50,100,300, and plot the 5 measure scores vs. γ for both SVD and NMF; also report the contingency matrices for each γ. Report the best γ choice for SVD and NMF respectively. And how do you explain the non-monotonic behavior of the measures as γ increases?

Then we try to find out the influence the components γ could have on the result, so we try γ = 1, 2, 3, 5, 10, 20, 50, 100, 300.

Result: Following table lists the metrics after LSI and NMF under different dimensions of features.

γ metrics	1	2	3	5	10	20	50	100	300
Homogeneity	0.000	0.598	0.390	0.221	0.235	0.233	0.242	0.243	0.246
Completeness	0.000	0.599	0.432	0.310	0.321	0.320	0.326	0.328	0.330
V-measure	0.000	0.599	0.410	0.258	0.271	0.270	0.278	0.279	0.282
Adjusted Rand	0.000	0.702	0.377	0.145	0.158	0.156	0.167	0.168	0.171
Adjusted Mutual Info	0.000	0.598	0.390	0.221	0.235	0.233	0.242	0.243	0.246

Table 1: 5 measuring scores of LSI

γ metrics	1	2	3	5	10	20	50	100	300
Homogeneity	0.000	0.593	0.238	0.126	0.096	0.093	0.070	0.072	0.015
Completeness	0.000	0.608	0.317	0.128	0.219	0.216	0.195	0.196	0.136
V-measure	0.000	0.600	0.272	0.127	0.133	0.130	0.103	0.106	0.027
Adjusted Rand	0.000	0.649	0.170	0.165	0.036	0.034	0.021	0.022	0.001
Adjusted Mutual Info	0.000	0.593	0.237	0.126	0.096	0.092	0.070	0.072	0.015

Table 2: Performance of NMF

Below we give the list of the contingency matrix given n components under LSI and NMF respectively.

```
~~~~~~~~~~ 5 measure scores when γ is 1 ~~~~~~~~~~
The contingency matrix:
 [[2200 1703]
 [2323 1656]]
~~~~~~~~~~ 5 measure scores when γ is 2 ~~~~~~~~~~
The contingency matrix:
 [[3684  219]
 [ 421 3558]]
~~~~~~~~~~ 5 measure scores when γ is 3 ~~~~~~~~~~
The contingency matrix:
 [[3874   29]
 [1492 2487]]
~~~~~~~~~~ 5 measure scores when γ is 5 ~~~~~~~~~~
The contingency matrix:
 [[3898    5]
 [2437 1542]]
~~~~~~~~~~ 5 measure scores when γ is 10 ~~~~~~~~~~
The contingency matrix:
 [[3900    3]
 [2372 1607]]
~~~~~~~~~~ 5 measure scores when γ is 20 ~~~~~~~~~~
The contingency matrix:
 [[   3 3900]
 [1600 2379]]
~~~~~~~~~~ 5 measure scores when γ is 50 ~~~~~~~~~~
The contingency matrix:
 [[4    3899]
 [1654 2325]]
~~~~~~~~~~ 5 measure scores when γ is 100 ~~~~~~~~~~
The contingency matrix:
 [[3900    3]
 [2324 1655]]
~~~~~~~~~~ 5 measure scores when γ is 300 ~~~~~~~~~~
The contingency matrix:
 [[3900    3]
 [2307 1672]]
```

```
~~~~~~~~~~ 5 measure scores when γ is 1 ~~~~~~~~~~
The contingency matrix:
 [[2200 1703]
 [2323 1656]]
~~~~~~~~~~ 5 measure scores when γ is 2 ~~~~~~~~~~
The contingency matrix:
 [[ 731 3172]
 [3943   36]]
~~~~~~~~~~ 5 measure scores when γ is 3 ~~~~~~~~~~
The contingency matrix:
 [[13   3890]
 [1674 2305]]
~~~~~~~~~~ 5 measure scores when γ is 5 ~~~~~~~~~~
The contingency matrix:
 [[896  3007]
 [2537 1442]]
~~~~~~~~~~ 5 measure scores when γ is 10 ~~~~~~~~~~
The contingency matrix:
 [[ 712 3191]
 [   2 3977]]
~~~~~~~~~~ 5 measure scores when γ is 20 ~~~~~~~~~~
The contingency matrix:
 [[ 3214 689]
 [   3977 2]]
~~~~~~~~~~ 5 measure scores when γ is 50 ~~~~~~~~~~
The contingency matrix:
 [[3371  532]
 [3977    2]]
~~~~~~~~~~ 5 measure scores when γ is 100 ~~~~~~~~~~
The contingency matrix:
 [[ 3347 556]
 [   3976 3]]
~~~~~~~~~~ 5 measure scores when γ is 300 ~~~~~~~~~~
The contingency matrix:
 [[3790  113]
 [3979    0]]
```

According to the results listed above, we just plot the 5 measure scores v.s. γ to find out the behavior of the measures as γ increases after the data-preprocessing with LSI. The figure follows as below.

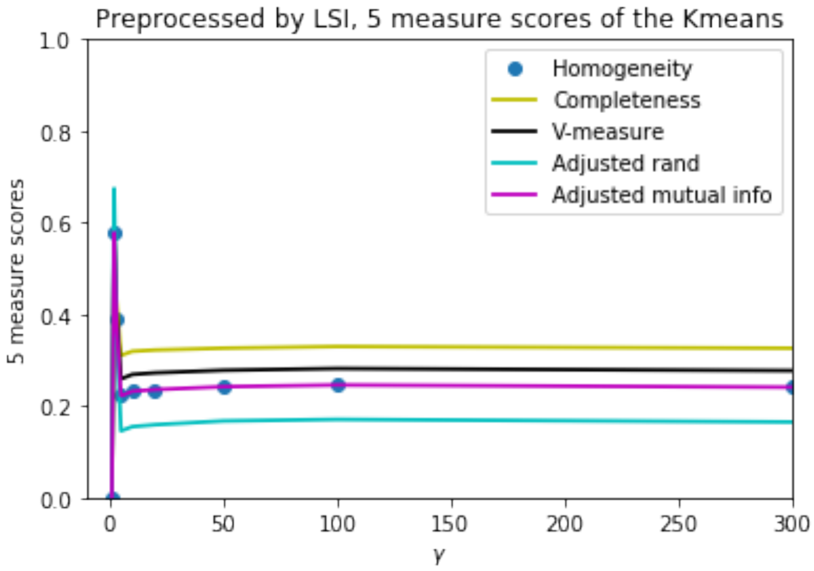

Figure 2: 5 measure scores v.s. γ for LSI

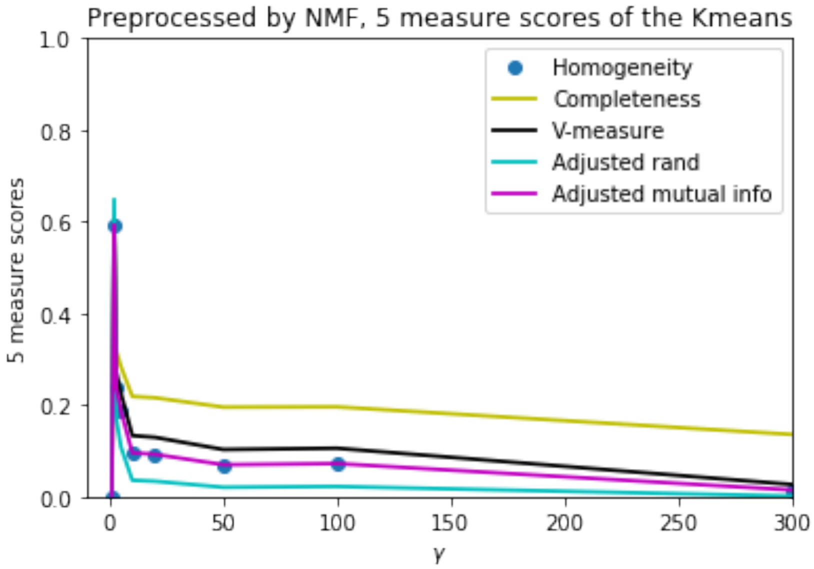

Figure 3: 5 measure scores v.s. γ for NMF

Conclusion: From the results, it could be concluded the preprocessing of data with LSI and NMF has some beneficial effects on the results of clustering and that the optimal value in LSI and NMF are both 2. And the behavior of the measures as γ increases is shown to be non-monotonic.

After discussion, we suppose that the probable reason of the non-monotonic behavior of the measures as γ increases is that there are too much useless terms in the sparse TF-IDF matrix before the dimension reduction which might have a negative effect on the clustering. As we mentioned before, K-means algorithm is based on the distances between data points in space, while the distance in high dimensional space make little sense. Thus, when γ increases over the appropriate value, the matrix after dimension reduction is, to some extent, sparse with regard to the clustering algorithm, which would make the results worse. In addition, the reason why the measures increase as γ increases when γ is relatively small is that too few terms extracted would also confuse the clustering.

Problem 4

In this problem, we need to explore more based on the best γ value that we have found in Problem 3, in another word, we are supposed to reduce the dimension of TFxIDF to 2 and visualize it onto both two-dimensional plane and color coding for the purpose to help understand the data more thoroughly.

Requirement A: Visualize the performance of the case with best clustering results in the previous part your clustering by projecting final data vectors onto two dimension plane and color-coding the classes.

Result: The visualization results of the final clustering vectors at the best γ of 2 are listed as follows. We represent them as two parts, LSI and NMF respectively, as shown in Figure 4 and Figure 5.

—————————————RESULTS of clustering with LSI at γ of 2

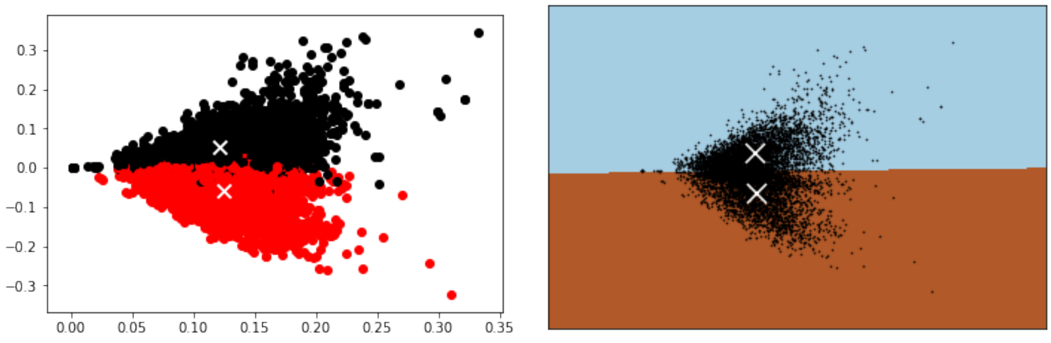

Figure 4: The clustering results of LSI at the best γ value of 2

—————————————RESULTS of clustering with NMF at γ of 2

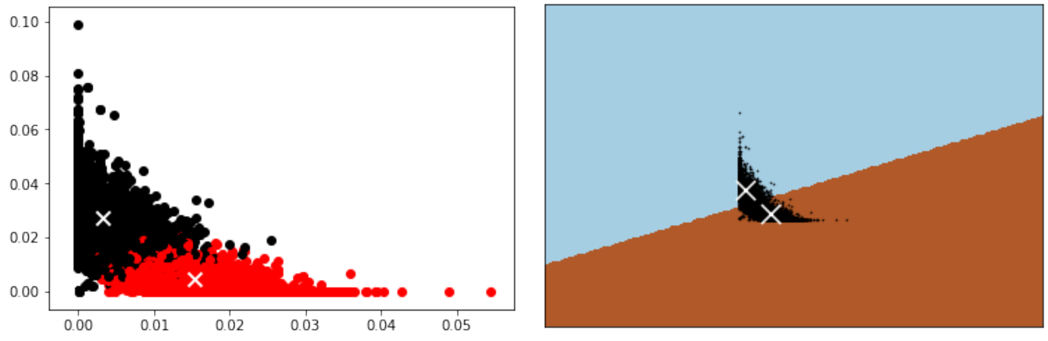

Figure 5: The clustering results of NMF at the best γ value of 2

Conclusion: The 5 measuring scores of these two clustering processes have been mentioned in Problem 3. According to the measuring scores, almost same scores make the performance of these two clustering methods equally satisfying.

Requirement B: Visualize the transformed data as in part (a). Report the new clustering measures including the contingency matrix after transformation. Normalizing features s.t. each feature has unit variance, i.e. each column of the reduced-dimensional data matrix has unit variance (if we use the convention that rows correspond to documents). Applying a non-linear transformation to the data vectors only after NMF. Here we use logarithm transformation as an example. Justify why logarithm transformation may increase the clustering results? Now try combining both transformations (in different orders) on NMF- reduced data.

In this part of the problem to improve the performance of the Kmeans algorithm, we are asked to use various preprocessing methods, according the order of only re-scaling the vectors after LSI and NMF respectively, using logarithm transformation as an example of the non-linear transformation and combining two methods above in two different orders. After coming out of the vectors, we need to visualize the results.

Result: Figure 6-10 are the results of the clustering results with different preprocessing methods.

—————————————RESULTS of clustering with LSI and only scaling

```
Homogeneity: 0.256
Completeness: 0.286
V-measure: 0.270
Adjusted Rand-Index: 0.274
Adjusted Mutual-Index: 0.255
The contingency matrix:
 [[2236 1667]
 [210  3769]]
```

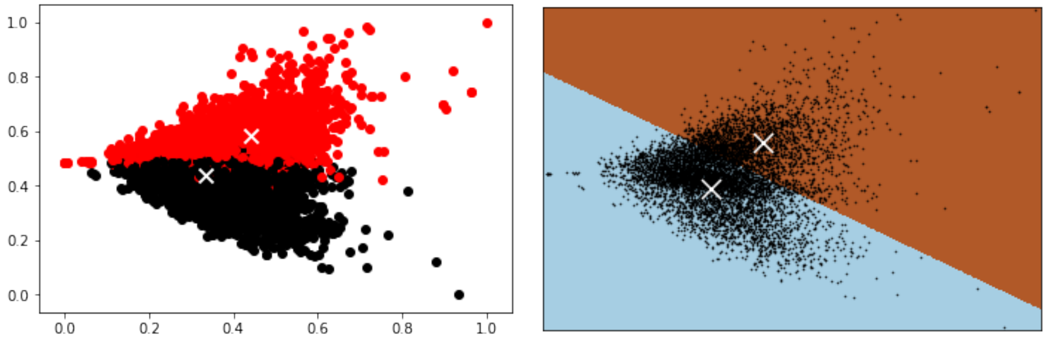

Figure 6: The clustering result with LSI and only scaling at γ of 2

——————————————RESULTS of clustering with NMF and only scaling

```
Homogeneity: 0.677
Completeness: 0.678
V-measure: 0.678
Adjusted Rand-Index: 0.774
Adjusted Mutual-Index: 0.677
The contingency matrix:
 [[326  3577]
 [ 3831 148]]
```

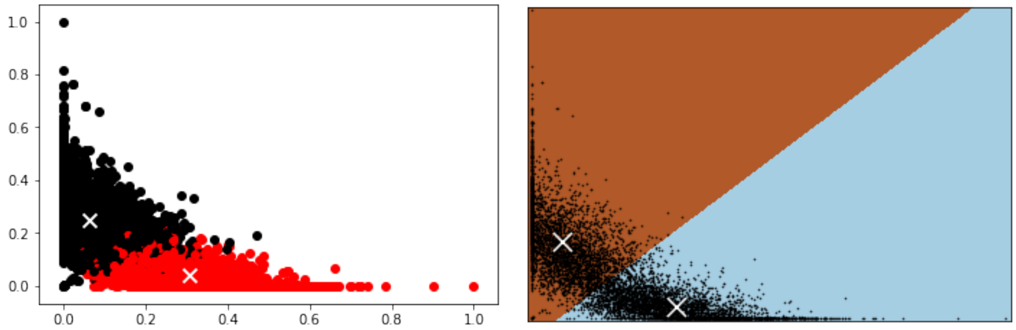

Figure 7: The clustering result with NMF and only scaling at γ of 2

——————————————RESULTS of clustering with NMF and only logarithm

```
Homogeneity: 0.706
Completeness: 0.707
V-measure: 0.707
Adjusted Rand-Index: 0.801
Adjusted Mutual-Index: 0.706
The contingency matrix:
 [[3761  142]
 [ 272 3707]]
```

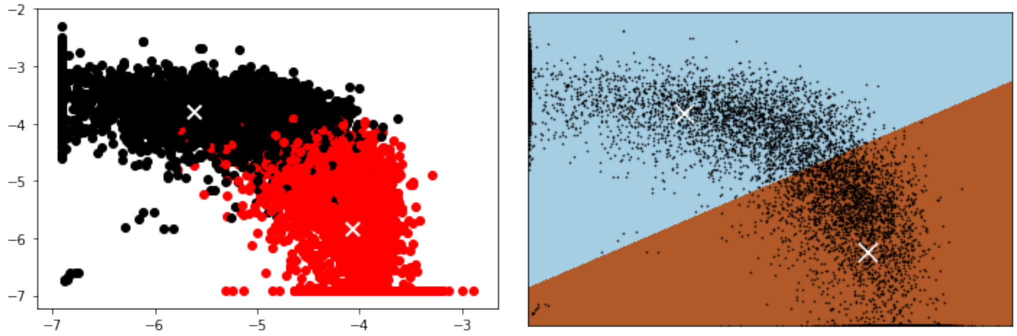

Figure 8: The clustering result with NMF and only logarithm at γ of 2

—————————————RESULTS of clustering with NMF and scaling then logarithm

```
Homogeneity: 0.275
Completeness: 0.354
V-measure: 0.310
Adjusted Rand-Index: 0.201
Adjusted Mutual-Index: 0.275
The contingency matrix:
 [[   3903 0]
 [2172 1807]]
```

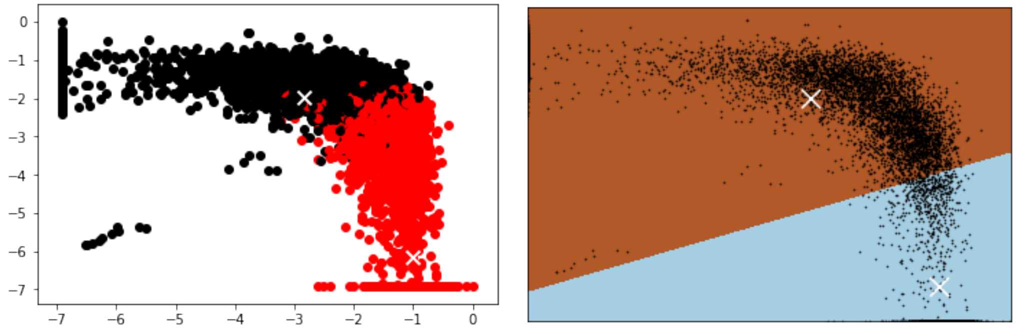

Figure 9: The clustering result with NMF and scaling before logarithm at γ of 2

—————————————RESULTS of clustering with NMF and logarithm then scaling

```
Homogeneity: 0.709
Completeness: 0.709
V-measure: 0.709
Adjusted Rand-Index: 0.806
Adjusted Mutual-Index: 0.709
The contingency matrix:
 [[ 210 3693]
 [3787  192]]
```

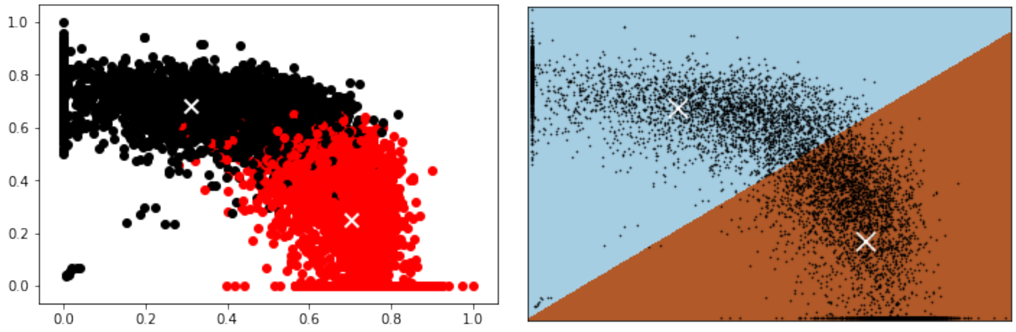

Figure 10: The clustering result with NMF and logarithm before scaling at γ of 2

Conclusion: The 5 measuring scores of each situation are organized into the table so that we can observe more easily and draw conclusion more intuitively.

	Figure 4	Figure 5	Figure 6	Figure 7	Figure 8	Figure 9	Figure 10
	None		scaling		log	scaling & log	log & scaling
	LSI	NMF	LSI	NMF	NMF	NMF	NMF
Homogeneity	0.598	0.593	0.256	0.677	0.706	0.275	0.709
Completeness	0.599	0.608	0.286	0.678	0.707	0.354	0.709
V-measure	0.599	0.600	0.270	0.678	0.707	0.309	0.709
Adjusted Rand	0.702	0.649	0.274	0.774	0.801	0.201	0.806
Adjusted Mutual Info	0.598	0.593	0.255	0.677	0.706	0.275	0.709

Table 3: The 5 measuring scores of each situation mentioned above

As shown in Table 3, we can draw several conclusions by comparing different sets of data. Firstly, comparing the data in Figure 4 with Figure 6, as well as Figure 5 with Figure 7, the scaling brings significant improvement to the preprocessing method with NMF while fatal depletion to the method with LSI, which reveals the crucially significance and specifity of the scaling method to NMF. So in the following part of the problem, we only apply the scaling method with NMF instead of both LSI and NMF.

Also, the data processed by LSI can be negative but those processed by NMF only have positive numbers in the vectors, thus we can only apply the logarithm with the NMF too. The numbers in Figure 5 and Figure 8 simply give us the facts that logarithm transformation can also help promotion of the clustering performance from “fair” to “good”. According to the result, we can find that the logarithm has better promoting effect in comparison with the scaling.

Therefore we may wonder what it can achieve if we combine the two promoting methods together. Naturally, we will probe into the combination in which order we can accomplish better clustering results. Observing the data coming from Figure 9 and Figure 10, the huge discrepancy makes it easy for us to conclude that the scaling can affect perfectly only when it’s applied after the NMF rather than before.

Question: *Can you justify why logarithm transformation may increase the clustering results?* ***Answer:*** *I believe that the improvement is brought by the unique property of the logarithm function. To be more specific, the derivative of the logarithm function log(x) is indirect proportion to the reciprocal of x. Since the values of TFxIDF vectors are ranging between 0 and 1, the logarithm function log(x) decreases more and more rapidly as x is approaching 0. As we suppose the data is divided into two parts with two centroids, so when we apply the logarithm function to the dimension reduced TFxIDF vectors, the centroids have been drawn farther away, which can effectively distance the data that surrounding its own centroid. Intuitively, we can simply observe that, with two centroids being farther and the all the data surrounding its own centroid, it’s more effortless for us to cluster the data.*

Problem 5

Requirement: Expand Dataset into 20 categories. Try different dimensions for both truncated SVD and NMF dimensionality reduction techniques and the different transformations of the obtained feature vectors as outlined above.

What we are supposed to do in this problem is repeating what we have done in Problem 1 to Problem 4 with the whole 20 groups data, and use the results of 20 groups data to verify the conclusion we have obtained above. So firstly, we do the K-means algorithm directly with the TFxIDF of 20 groups data. Then we use 5 measuring scores as the coefficients to locate the best γ for TFxIDF respectively processed by LSI and NMF. With the best γ, we explore the impact of normalization and on-linear transformation on the clustering results by applying each of them and both with different orders.

Result: The results of this problem are listed as follows.

```
The dimension of TF-IDF Vector with min_df=3 is (18846, 52295)
The number of Terms Extracted with min_df=3 is 52295
```

```
Homogeneity: 0.309
Completeness: 0.363
V-measure: 0.334
Adjusted Rand-Index: 0.125
Adjusted Mutual-Index: 0.307
The contingency matrix:
 [[ 38 144 13 2 95 0 0 156 1 14 70 27 0 0 0 88 5 144 2 0]
 [ 0 186 0 1 96 81 1 1 0 2 81 18 1 0 1 383 17 2 9 93]
 [ 0 111 0 0 64 498 5 0 0 6 31 12 11 0 1 128 11 1 10 96]
 [ 0 109 0 0 113 48 2 0 0 6 26 6 186 2 2 172 31 0 28 251]
 [ 0 221 0 0 72 8 1 0 0 18 21 22 72 0 0 202 15 1 27 283]
 [ 0 121 0 0 145 167 1 0 3 21 88 1 0 0 0 417 3 0 6 15]
 [ 0 381 0 3 131 21 5 0 0 15 5 5 42 12 39 182 7 1 42 84]
 [ 0 135 0 3 161 2 0 0 0 18 18 41 0 0 483 76 12 3 38 0]
 [ 0 128 0 0 206 0 1 0 0 12 100 9 0 0 442 71 10 1 16 0]
 [ 0 334 0 4 144 0 1 0 0 8 2 24 0 384 1 63 21 0 8 0]
 [ 0 108 0 0 37 0 0 0 0 51 2 2 0 730 1 57 0 0 11 0]
 [ 0 72 0 28 154 8 27 0 501 9 51 5 0 0 0 118 9 0 7 2]
 [ 0 195 0 0 167 8 12 0 1 6 51 11 5 1 30 435 13 10 8 31]
 [ 0 287 0 2 188 2 2 2 0 9 20 6 0 0 1 446 7 9 9 0]
 [ 0 91 0 2 70 1 182 0 0 11 20 40 0 0 1 552 12 4 1 0]
 [ 1 127 2 2 91 1 0 540 0 11 17 9 0 0 1 190 3 0 2 0]
 [ 0 98 1 527 146 0 0 0 4 4 14 32 0 0 3 52 4 6 15 4]
 [ 0 118 633 0 64 0 0 2 0 2 5 30 0 0 1 63 2 0 20 0]
 [ 0 122 105 87 179 0 1 7 2 0 22 23 0 0 1 173 1 0 52 0]
 [ 70 102 2 44 87 0 4 141 0 6 17 11 0 1 1 105 2 16 19 0]]
```

```
@@@ Do dimension reduction with LSI at min_df=3 @@@

~~~~~~~~~~ 5 measure scores when γ is 1 ~~~~~~~~~~
Homogeneity: 0.028
Completeness: 0.031
V-measure: 0.029
Adjusted Rand-Index: 0.006
Adjusted Mutual-Index: 0.025
~~~~~~~~~~ 5 measure scores when γ is 2 ~~~~~~~~~~
Homogeneity: 0.210
Completeness: 0.224
V-measure: 0.217
Adjusted Rand-Index: 0.065
Adjusted Mutual-Index: 0.208
~~~~~~~~~~ 5 measure scores when γ is 5 ~~~~~~~~~~
Homogeneity: 0.309
Completeness: 0.325
V-measure: 0.317
Adjusted Rand-Index: 0.122
Adjusted Mutual-Index: 0.306
~~~~~~~~~~ 5 measure scores when γ is 8 ~~~~~~~~~~
Homogeneity: 0.340
Completeness: 0.377
V-measure: 0.357
Adjusted Rand-Index: 0.138
Adjusted Mutual-Index: 0.338
~~~~~~~~~~ 5 measure scores when γ is 10 ~~~~~~~~~~
Homogeneity: 0.339
Completeness: 0.380
V-measure: 0.358
Adjusted Rand-Index: 0.137
Adjusted Mutual-Index: 0.337
~~~~~~~~~~ 5 measure scores when γ is 15 ~~~~~~~~~~
Homogeneity: 0.319
Completeness: 0.376
V-measure: 0.345
Adjusted Rand-Index: 0.129
Adjusted Mutual-Index: 0.317
~~~~~~~~~~ 5 measure scores when γ is 20 ~~~~~~~~~~
Homogeneity: 0.288
Completeness: 0.383
V-measure: 0.329
Adjusted Rand-Index: 0.093
Adjusted Mutual-Index: 0.286
~~~~~~~~~~ 5 measure scores when γ is 50 ~~~~~~~~~~
Homogeneity: 0.296
Completeness: 0.375
V-measure: 0.331
Adjusted Rand-Index: 0.109
Adjusted Mutual-Index: 0.294
~~~~~~~~~~ 5 measure scores when γ is 100 ~~~~~~~~~~
Homogeneity: 0.348
Completeness: 0.481
V-measure: 0.404
Adjusted Rand-Index: 0.110
Adjusted Mutual-Index: 0.346
```

```
@@@ Do dimension reduction with NMF at min_df=3 @@@
~~~~~~~~~~ 5 measure scores when γ is 1 ~~~~~~~~~~
Homogeneity: 0.028
Completeness: 0.031
V-measure: 0.029
Adjusted Rand-Index: 0.006
Adjusted Mutual-Index: 0.025
~~~~~~~~~~ 5 measure scores when γ is 2 ~~~~~~~~~~
Homogeneity: 0.165
Completeness: 0.177
V-measure: 0.171
Adjusted Rand-Index: 0.048
Adjusted Mutual-Index: 0.163
~~~~~~~~~~ 5 measure scores when γ is 5 ~~~~~~~~~~
Homogeneity: 0.279
Completeness: 0.294
V-measure: 0.286
Adjusted Rand-Index: 0.104
Adjusted Mutual-Index: 0.276
~~~~~~~~~~ 5 measure scores when γ is 8 ~~~~~~~~~~
Homogeneity: 0.292
Completeness: 0.325
V-measure: 0.308
Adjusted Rand-Index: 0.112
Adjusted Mutual-Index: 0.290
~~~~~~~~~~ 5 measure scores when γ is 10 ~~~~~~~~~~
Homogeneity: 0.304
Completeness: 0.340
V-measure: 0.321
Adjusted Rand-Index: 0.117
Adjusted Mutual-Index: 0.302
~~~~~~~~~~ 5 measure scores when γ is 15 ~~~~~~~~~~
Homogeneity: 0.257
Completeness: 0.301
V-measure: 0.277
Adjusted Rand-Index: 0.088
Adjusted Mutual-Index: 0.255
~~~~~~~~~~ 5 measure scores when γ is 20 ~~~~~~~~~~
Homogeneity: 0.257
Completeness: 0.329
V-measure: 0.289
Adjusted Rand-Index: 0.086
Adjusted Mutual-Index: 0.255
~~~~~~~~~~ 5 measure scores when γ is 50 ~~~~~~~~~~
Homogeneity: 0.217
Completeness: 0.302
V-measure: 0.252
Adjusted Rand-Index: 0.046
Adjusted Mutual-Index: 0.214
~~~~~~~~~~ 5 measure scores when γ is 100 ~~~~~~~~~~
Homogeneity: 0.177
Completeness: 0.317
V-measure: 0.227
Adjusted Rand-Index: 0.027
Adjusted Mutual-Index: 0.175
```

```
~~ Do the scaling after LSI, 5 measure scores when γ is 10 ~~
Homogeneity: 0.300
Completeness: 0.329
V-measure: 0.314
Adjusted Rand-Index: 0.129
Adjusted Mutual-Index: 0.298
The contingency matrix:
 [[239 127 0 98 2 1 0 12 54 56 0 0 38 92 29 11 2 0 38 0]
 [ 1 2 0 35 56 0 0 367 0 147 7 0 5 60 183 9 12 0 65 24]
 [ 0 0 0 10 289 0 2 374 0 82 1 0 2 31 83 8 7 0 53 43]
 [ 0 1 0 23 40 0 101 128 0 91 0 0 0 37 119 27 11 3 101 300]
 [ 0 0 0 37 10 0 21 72 0 201 5 0 4 96 141 33 8 1 77 257]
 [ 0 1 0 9 81 0 0 455 0 84 10 0 0 43 196 4 20 0 84 1]
 [ 1 1 0 29 9 0 16 36 0 312 1 0 1 80 207 44 10 12 93 123]
 [ 0 5 0 237 0 0 0 9 0 162 2 0 39 84 131 40 21 0 256 4]
 [ 7 31 0 178 0 0 0 23 0 122 1 0 11 83 125 14 21 2 378 0]
 [ 0 5 0 91 0 0 0 4 0 167 1 0 9 132 104 7 3 358 113 0]
 [ 1 0 0 8 0 0 0 0 0 67 0 0 3 57 105 10 3 710 35 0]
 [ 0 30 0 116 8 488 0 47 0 24 9 0 49 47 44 4 18 0 105 2]
 [ 1 0 0 169 9 2 1 72 0 194 14 0 5 86 204 7 38 1 145 36]
 [ 5 5 0 206 3 0 0 8 2 180 1 75 25 76 252 6 30 0 116 0]
 [ 0 1 0 59 0 0 0 9 0 70 284 0 14 34 95 1 375 0 44 1]
 [457 2 0 63 1 0 0 8 211 63 2 0 17 23 115 1 2 0 31 1]
 [ 1 109 1 221 0 5 0 1 0 31 2 0 291 85 33 15 5 0 110 0]
 [ 3 3 547 90 0 0 0 2 2 63 0 0 56 69 63 21 0 0 21 0]
 [ 5 68 1 206 0 2 0 0 2 50 3 0 184 55 35 47 22 1 94 0]
 [149 63 0 93 0 0 0 0 71 39 0 1 42 54 46 18 5 0 47 0]]
```

```
~~ Do the scaling after NMF, 5 measure scores when γ is 10 ~~
Homogeneity: 0.279
Completeness: 0.297
V-measure: 0.288
Adjusted Rand-Index: 0.105
Adjusted Mutual-Index: 0.277
The contingency matrix:
 [[ 2 74 46 0 70 98 1 45 0 55 9 0 54 61 0 56 1 60 167 0]
 [174 2 76 260 13 99 15 0 0 71 0 0 152 33 0 4 0 16 1 57]
 [200 0 54 272 7 36 4 0 5 21 0 0 65 15 1 2 0 7 0 296]
 [105 1 140 176 17 92 8 0 163 25 0 0 148 45 7 1 0 18 0 36]
 [104 2 80 112 27 231 11 0 60 23 0 0 197 91 4 1 0 14 0 6]
 [133 0 97 363 15 44 26 0 0 68 0 0 126 26 0 0 3 15 0 72]
 [ 73 4 126 35 16 310 8 0 36 5 0 3 177 137 14 1 0 22 0 8]
 [ 2 38 202 3 128 215 8 0 0 22 0 0 148 77 5 77 0 65 0 0]
 [ 0 9 249 0 119 142 15 0 0 85 0 0 127 51 14 49 0 136 0 0]
 [ 0 3 100 1 47 200 3 0 0 2 0 124 60 62 354 5 0 33 0 0]
 [ 1 1 34 0 8 80 3 0 0 2 0 387 31 35 409 2 0 6 0 0]
 [ 11 36 99 22 48 63 26 0 0 46 12 0 87 23 1 25 428 60 0 4]
 [ 45 2 166 56 45 193 30 0 3 46 0 0 301 58 9 6 1 21 0 2]
 [ 4 68 137 1 75 205 18 1 0 17 2 0 272 103 3 56 0 21 5 2]
 [ 1 24 72 4 26 117 482 0 0 22 3 0 184 24 1 12 0 15 0 0]
 [ 1 22 64 1 7 97 3 200 0 14 2 0 120 10 0 11 0 8 436 1]
 [ 0 251 40 0 61 54 3 0 0 12 138 0 25 43 0 185 3 95 0 0]
 [ 1 315 4 0 29 64 0 2 0 5 337 0 42 58 0 78 0 4 1 0]
 [ 1 163 35 0 45 77 13 0 0 118 64 0 51 52 2 111 2 35 6 0]
 [ 1 48 57 0 37 64 4 67 0 17 16 0 62 46 0 59 0 26 124 0]]
```

```
~~ Do the non-linear only, 5 measure scores when γ is 10 ~~
Homogeneity: 0.366
Completeness: 0.371
V-measure: 0.368
Adjusted Rand-Index: 0.200
Adjusted Mutual-Index: 0.364
The contingency matrix:
 [[ 0 0 93 99 46 2 179 3 8 251 5 4 44 51 0 2 0 0 11 1]
 [ 30 7 11 8 85 4 1 113 43 0 26 144 65 0 339 9 21 7 0 60]
 [ 45 4 6 5 25 3 0 61 20 1 14 206 19 1 468 7 38 5 0 57]
 [258 2 12 0 30 5 0 41 26 0 14 53 12 0 65 7 287 3 0 167]
 [157 5 8 3 82 11 0 37 23 0 14 20 22 0 35 9 278 2 1 256]
 [ 17 3 9 3 48 15 1 125 22 1 34 314 19 1 341 10 2 10 0 13]
 [ 98 31 17 1 182 4 1 7 34 0 12 37 37 0 26 44 238 5 1 200]
 [ 27 6 259 16 99 5 1 19 196 2 25 68 99 33 5 54 6 8 7 55]
 [ 32 6 202 63 58 1 12 51 170 1 25 49 66 54 0 164 16 5 3 18]
 [ 1 485 10 2 52 0 0 2 55 0 5 5 95 0 0 277 0 0 0 5]
 [ 0 754 1 0 12 0 0 2 9 0 3 5 15 0 0 190 0 0 0 8]
 [ 7 2 13 10 10 366 0 33 26 2 12 13 17 3 6 3 0 452 6 10]
 [ 53 10 32 9 107 38 3 61 151 1 46 51 73 9 36 35 25 18 1 225]
 [ 6 7 168 25 146 3 4 11 157 20 37 40 237 44 19 22 1 8 23 12]
 [ 5 4 27 29 43 3 0 12 78 2 618 17 104 3 8 14 0 5 5 10]
 [ 1 4 7 45 30 0 395 7 21 412 10 1 32 10 5 12 0 2 1 2]
 [ 2 0 178 27 18 10 2 2 48 19 8 13 78 251 0 7 0 13 232 2]
 [ 0 2 86 19 26 4 0 1 17 28 0 2 55 105 1 2 0 1 591 0]
 [ 1 2 117 168 28 10 6 10 48 21 22 1 124 107 1 7 0 8 94 0]
 [ 0 2 57 28 34 0 180 0 15 165 6 6 38 66 1 4 0 4 21 1]]
```

```
~~ Do the scaling after non-linear, 5 measure scores when γ is 10 ~~
Homogeneity: 0.349
Completeness: 0.355
V-measure: 0.352
Adjusted Rand-Index: 0.170
Adjusted Mutual-Index: 0.347
The contingency matrix:
 [[ 80 4 54 0 56 6 1 0 137 3 2 191 10 34 42 4 0 173 0 2]
 [ 2 277 1 4 104 143 1 177 0 0 9 1 32 15 23 20 142 2 14 6]
 [ 0 292 2 1 53 59 1 293 0 0 3 0 22 4 14 10 196 0 34 1]
 [ 1 108 1 7 42 86 4 69 0 0 4 0 38 8 30 14 244 1 321 4]
 [ 19 128 1 3 35 156 4 44 0 0 11 0 37 13 28 12 120 0 350 2]
 [ 3 235 0 6 109 50 7 206 1 0 3 2 26 2 28 30 270 0 0 10]
 [ 21 150 1 23 7 110 4 13 2 0 63 1 36 21 109 12 91 0 306 5]
 [ 64 41 57 7 29 49 2 2 1 0 20 1 160 306 164 21 42 4 15 5]
 [ 24 18 55 28 118 26 0 0 3 0 27 1 186 210 197 24 37 13 21 8]
 [ 8 20 1 198 4 20 0 0 0 0 476 1 47 139 62 7 9 0 0 2]
 [ 1 2 0 441 2 1 0 1 1 2 494 4 11 17 17 3 2 0 0 0]
 [ 18 16 14 2 38 38 357 5 0 7 1 3 38 15 14 16 18 0 0 391]
 [ 18 116 4 14 66 286 9 25 7 0 16 0 137 50 54 49 77 3 35 18]
 [167 64 73 7 28 134 1 4 16 2 17 29 152 135 97 36 19 1 1 7]
 [ 33 29 6 6 29 66 0 7 3 1 5 10 75 36 30 629 12 2 1 7]
 [ 19 11 1 3 13 15 0 1 331 0 5 362 16 9 1 9 1 199 0 1]
 [234 1 331 3 15 13 7 0 1 99 0 15 48 73 32 11 4 5 0 18]
 [323 7 138 0 5 8 3 1 0 347 6 32 12 34 19 0 1 3 0 1]
 [158 9 165 2 125 35 3 0 14 56 5 30 47 70 19 25 2 5 0 5]
 [ 64 4 53 2 14 9 1 0 93 4 1 119 14 43 20 5 5 174 0 3]]
```

```
~~ Do non-linear transformation after the scaling, 5 measure
scores when γ is 10 ~~
Homogeneity: 0.289
Completeness: 0.293
V-measure: 0.291
Adjusted Rand-Index: 0.132
Adjusted Mutual-Index: 0.287
The contingency matrix:
 [[ 1 67 27 0 1 4 25 2 106 0 2 1 193 59 0 1 0 4 109 197]
 [189 1 117 2 105 114 1 0 0 34 122 0 0 40 5 13 48 74 106 2]
 [292 2 57 0 216 110 1 0 0 57 61 0 0 47 5 3 41 56 37 0]
 [ 43 8 42 1 154 124 2 0 2 275 46 0 0 39 12 11 142 73 8 0]
 [ 31 4 71 7 40 101 16 7 3 270 33 0 0 28 2 9 262 67 8 4]
 [191 0 62 2 197 185 0 0 0 5 122 4 0 57 1 25 20 54 63 0]
 [ 49 1 46 42 37 131 25 10 3 282 5 4 2 4 9 13 150 130 30 2]
 [ 12 26 22 9 22 83 197 14 73 2 10 5 4 60 8 22 308 39 69 5]
 [ 7 113 3 15 25 57 216 4 35 8 10 9 15 75 31 19 157 86 108 3]
 [ 14 1 33 348 5 3 279 1 0 2 1 2 0 33 112 4 28 70 50 8]
 [ 2 0 11 454 0 4 75 0 0 0 3 0 0 15 318 3 1 86 19 8]
 [ 29 10 86 8 21 31 20 36 18 7 46 485 1 67 1 25 28 20 37 15]
 [ 45 5 154 8 47 61 17 3 4 37 48 8 4 71 7 30 276 108 51 0]
 [ 41 14 127 19 8 43 92 29 82 0 10 8 16 79 1 23 84 17 254 43]
 [ 20 13 102 1 7 22 28 4 13 2 21 6 1 97 2 456 55 29 93 15]
 [ 9 23 45 3 3 2 7 0 13 0 8 2 351 109 8 7 0 5 126 276]
 [ 2 26 51 1 13 7 92 130 317 0 0 65 2 65 5 7 40 3 38 46]
 [ 6 15 26 6 0 10 43 477 180 0 2 4 1 41 6 0 8 1 17 97]
 [ 5 155 60 5 1 7 71 74 126 0 2 25 11 62 5 14 25 1 59 67]
 [ 4 38 9 6 7 9 27 11 87 0 1 3 168 64 2 4 3 3 98 84]]
```

Conclusion: Below in Table 4-6 are the results after organization.

γ metrics	1	2	5	8	10	15	20	50	100
Homogeneity	0.028	0.210	0.309	0.340	0.339	0.319	0.288	0.296	0.348
Completeness	0.031	0.224	0.325	0.377	0.380	0.376	0.383	0.375	0.481
V-measure	0.029	0.217	0.317	0.357	0.358	0.345	0.329	0.331	0.404
Adjusted Rand	0.006	0.065	0.122	0.138	0.137	0.129	0.093	0.109	0.110
Adjusted Mutual Info	0.025	0.208	0.306	0.338	0.337	0.317	0.286	0.294	0.346

Table 4: 5 measuring scores of LSI

γ metrics	1	2	5	8	10	15	20	50	100
Homogeneity	0.028	0.165	0.279	0.292	0.304	0.257	0.257	0.217	0.177
Completeness	0.031	0.177	0.294	0.325	0.340	0.301	0.329	0.302	0.317
V-measure	0.029	0.171	0.286	0.308	0.321	0.277	0.289	0.252	0.227
Adjusted Rand	0.006	0.048	0.104	0.112	0.117	0.088	0.086	0.046	0.027
Adjusted Mutual Info	0.025	0.163	0.276	0.290	0.302	0.255	0.255	0.214	0.175

Table 5: 5 measuring scores of NMF

	None		scaling		log	scaling & log	log & scaling
	LSI	NMF	LSI	NMF	NMF	NMF	NMF
Homogeneity	0.339	0.304	0.299	0.279	0.366	0.289	0.349
Completeness	0.380	0.340	0.329	0.297	0.371	0.293	0.355
V-measure	0.358	0.321	0.314	0.288	0.368	0.291	0.352
Adjusted Rand	0.137	0.117	0.128	0.105	0.200	0.132	0.170
Adjusted Mutual Info	0.337	0.302	0.297	0.277	0.364	0.287	0.347

Table 6: The 5 measuring scores of each situation

Based on the performance of the clustering, the best gamma of 20 groups data is 10 for both LSI and NMF. And according to the data organized in the tables, we can confirm that most of the conclusions coming from 8 groups data are a little bit different for the 20 groups data. The logarithm transformation can give the best clustering results, while the scaling has negative effect on the results.