ECE 219

Project 2

Shengrong Wu & Shuangyu Li

Introduction

In this project, we Implemented clustering on *sklearn.fetch_20newsgroups* that we already explored in Project 1. We would like to evaluate how purely the *a priori* known classes can be reconstructed through clustering. To achieve this, we utilized *sklearn* and *nltk* to implement K-means clustering algorithm on TFIDF matrix, with dimension reduction techniques, such as LSI and NMF. We examined whether and how reducing dimension will affect the performance. In the first part of the project, we worked with a well separable portion of data , and see if we could retrieve the two known classes. In the last part, we would expand dataset into 20 datasets and how purely we can retrieve all 20 original sub-class labels with clustering. Their statistical results are calculated and plotted to compare performance throughout this report.

1.Build the TF-IDF Matrix

We started with the easier task on data from 8 different classes shown in Table 1. We first computed out the TF-IDF matrix, which has a dimension of (7882, 27743). For this part we use min_df = 3, and combine stopping words from different packages but do not perform any stemming. 27743 is the number of distinct words in our TF-IDF matrix. 7882 is the documents size which includes both training and test data set.

Table 1. Subclasses of ‘ Computer technology ’ and ‘Recreational activity ’

Computer technology	Recreational activity
comp.graphics	rec.autos
comp.os.ms-windows.misc	rec.motorcycles
comp.sys.ibm.pc.hardware	rec.sport.baseball
comp.sys.mac.hardware	rec.sport.hockey

2. Apply K-Means Clustering on TF-IDF Matrix

In this part, we use the Kmeans function from sklearn_cluster package to learn the dataset into two clusters. Because of the randomness of the the K-Means method algorithm, we set n_init = 30 to run 30 times and get the final model. Then we evaluated the performance of the clustering by comparing it to the class labels as ground truth.

Problem a) Inspecting the contingency matrix

The contingency matrix shows that the clustering results is consistent with the ground truth class labels.

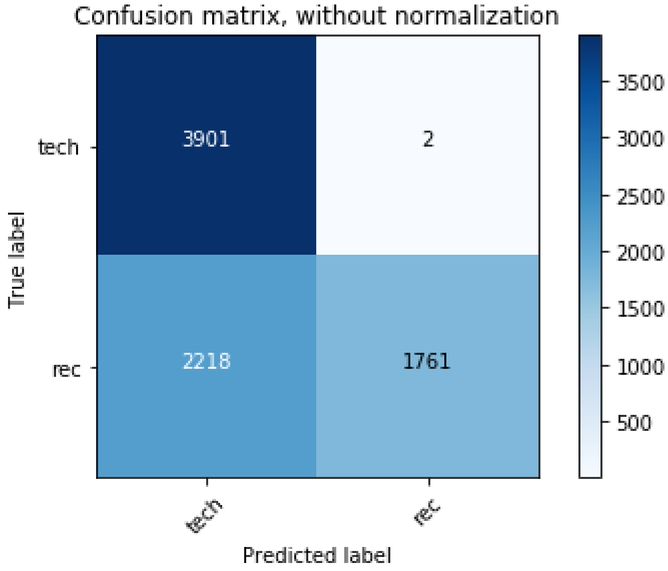


Fig 1: Contingency Matrix with TF-IDF

Problem b) 5 Measures of Purity

Table 2: Measurement for entire TF-IDF

Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Mutual info score
0.264	0.344	0.298	0.191	0.183

3. Preprocessing Data for Clustering

In the last part, the high dimensional sparse TF-IDF vectors yield suboptimal results. In this part, we will try to solve the problem by finding a “better” representation of the data. We will reduce the dimension of the data by performing LSI and NMF, as we did in Project 1.

i) Significant Terms in Truncated SVD

We want to find the effective dimension of the data through inspection of the top singular values of TF-IDF matrix. Here we plot the percent of variance vs. r for r = 1 to 1000. We could see that r = 1000 contains about half the variance of the TF-IDF matrix.

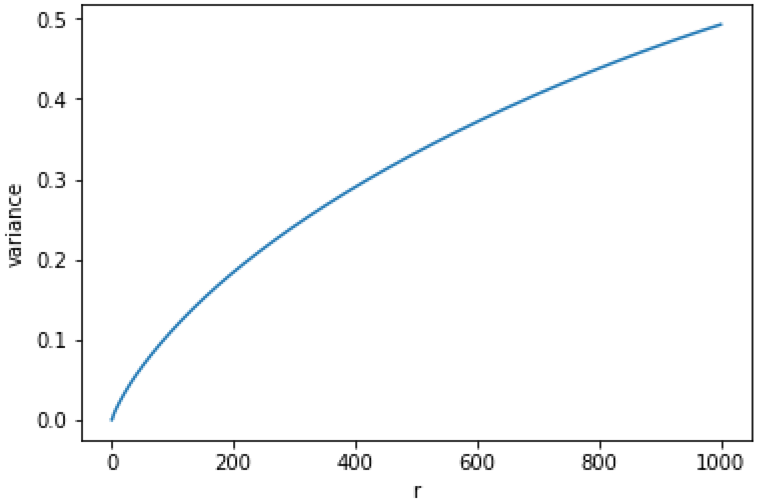


Fig 2: variance vs r of TF-IDF

ii) LSI & NMF

Now we are going to perform LSI and NMF dimension reduction techniques on TF-IDF with different choices of r. The contingency matrices and measure scores for r = [1, 2, 3, 5, 10, 20, 50, 100, 300]are displayed below and plots of score against different r are presented as well. Note that the predicted labels are arbitrarily assigned to each of the clusters by the clustering algorithm. Therefore, the corresponding labels might be swapped in the matrix graph. We could just treat the predicted labels as class 1 and class 2. The best r for NMF is 2; and the best r for LSI is 2.

Table 3: Measurement for LSI, r = 1

Homogeneity	**Completeness**	**V-measure**	**Adjusted Rand-Index**	**Mutual info score**
0.001	0.001	0.001	0.001	0.001

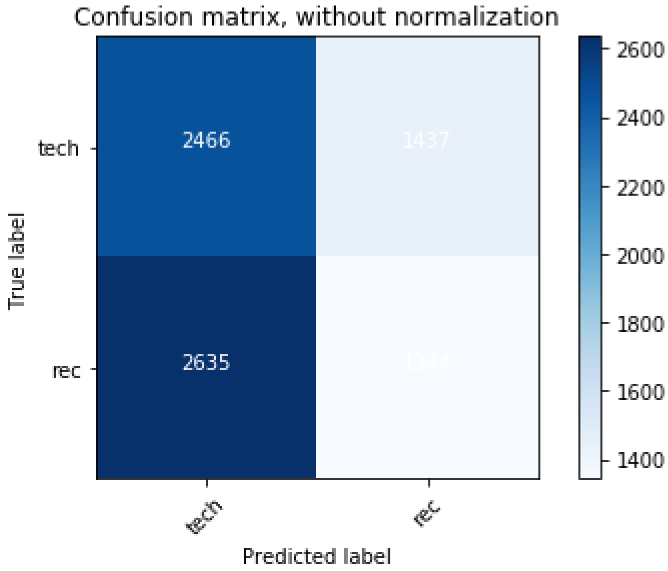


Fig 3: Contingency Matrix for LSI, r = 1

Table 4: Measurement for LSI, r = 2

Homogeneity	**Completeness**	**V-measure**	**Adjusted Rand-Index**	**Mutual info score**
0.334	0.392	0.361	0.292	0.232

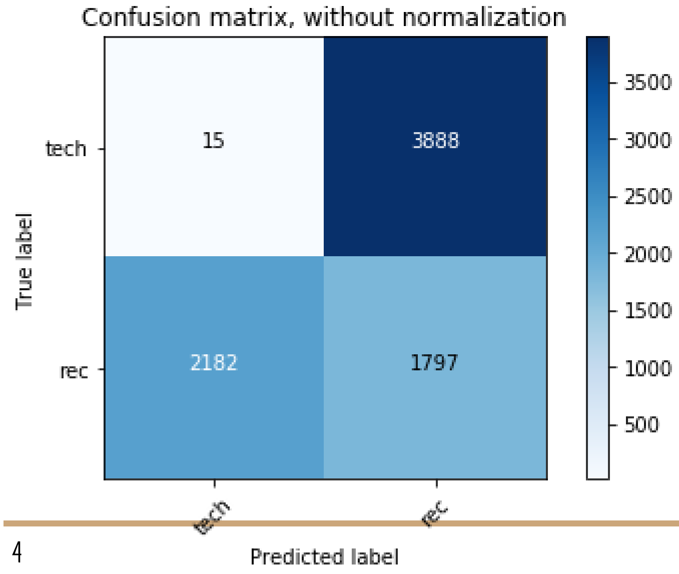


Fig 4: Contingency Matrix for LSI, r = 2

Table 5: Measurement for LSI, r = 3

Homogeneity	**Completeness**	**V-measure**	**Adjusted Rand-Index**	**Mutual info score**
0.227	0.317	0.264	0.147	0.157

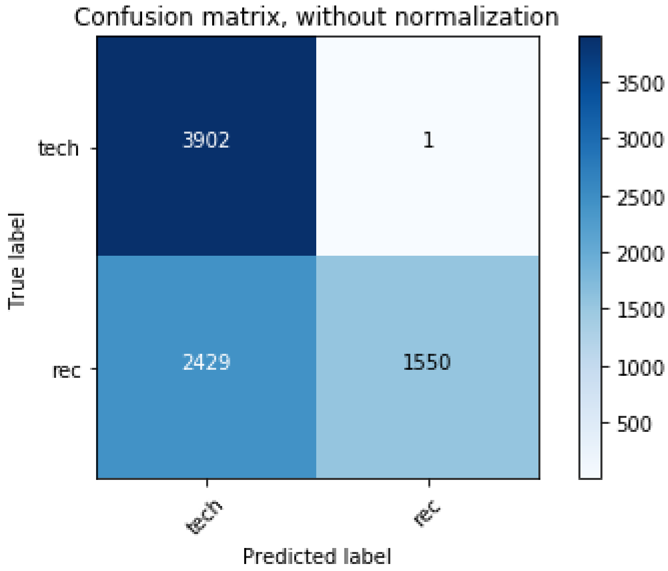


Fig 5: Contingency Matrix for LSI, r = 3

Table 6: Measurement for LSI, r = 5

Homogeneity	**Completeness**	**V-measure**	**Adjusted Rand-Index**	**Mutual info score**
0.234	0.323	0.271	0.155	0.162

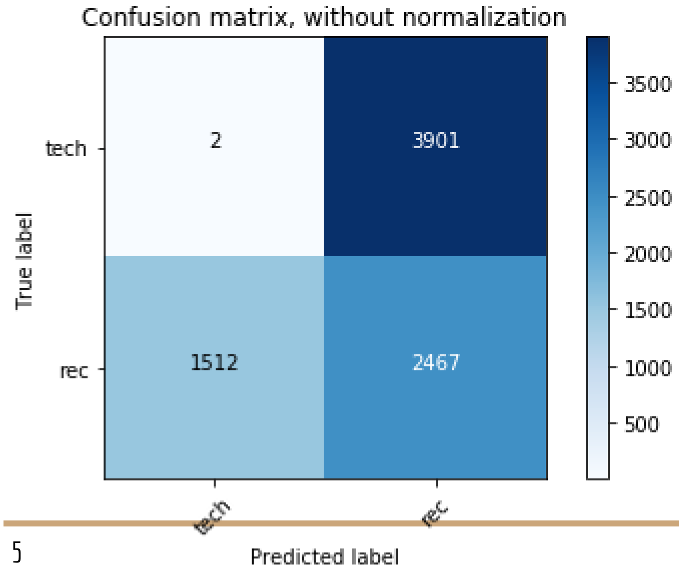


Fig 6: Contingency Matrix for LSI, r = 5

Table 7: Measurement for LSI, r = 10

Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Mutual info score
0.242	0.328	0.278	0.165	0.168

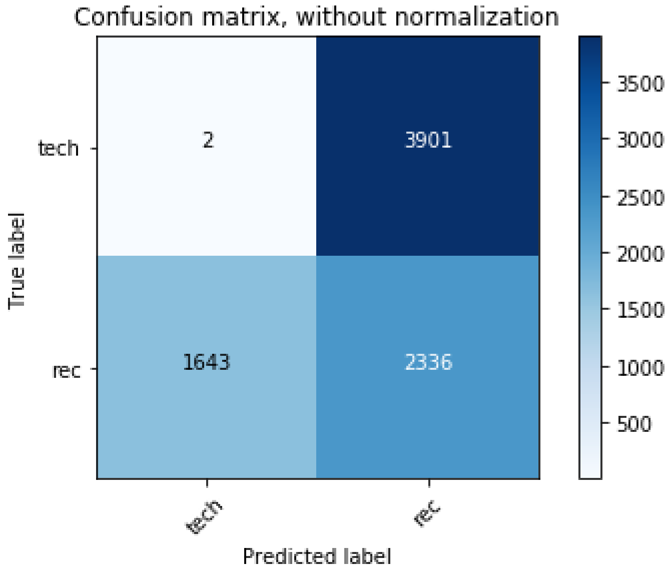


Fig 7: Contingency Matrix for LSI, r = 10

Table 8: Measurement for LSI, r = 20

Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Mutual info score
0.249	0.333	0.285	0.172	0.172

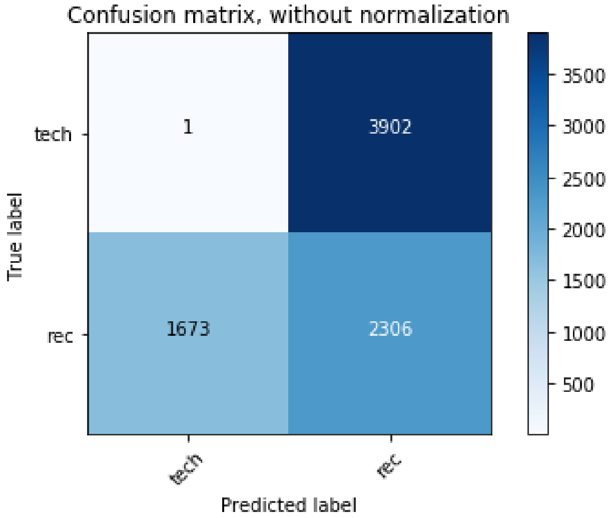


Fig 8: Contingency Matrix for LSI, r = 20

Table 8: Measurement for LSI, r = 50

Homogeneity	**Completeness**	**V-measure**	**Adjusted Rand-Index**	**Mutual info score**
0.250	0.335	0.286	0.171	0.173

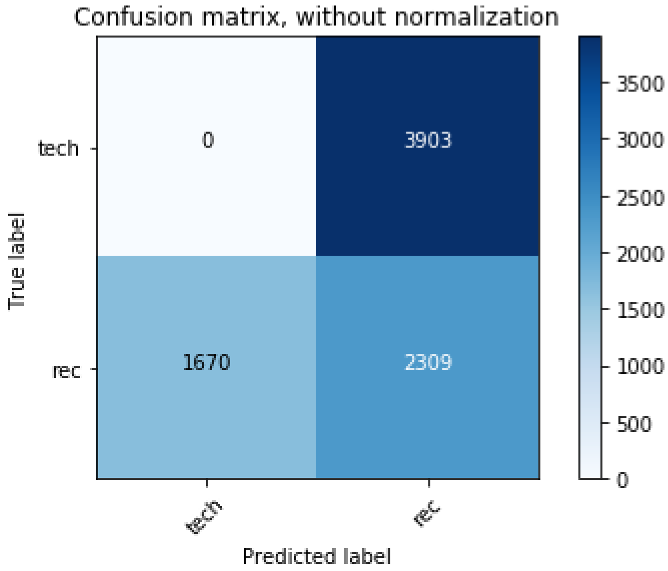


Fig 8: Contingency Matrix for LSI, r = 50

Table 9: Measurement for LSI, r = 100

Homogeneity	**Completeness**	**V-measure**	**Adjusted Rand-Index**	**Mutual info score**
0.247	0.332	0.284	0.170	0.171

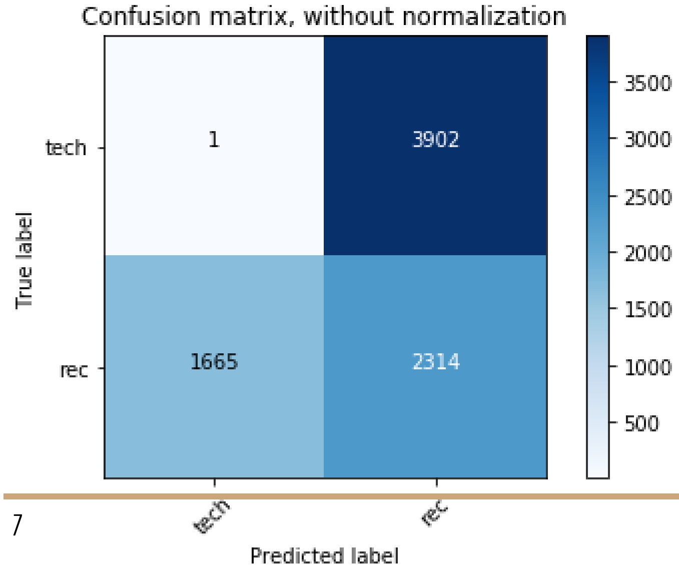


Fig 9: Contingency Matrix for LSI, r = 100

Table 10: Measurement for LSI, r = 300

Homogeneity	**Completeness**	**V-measure**	**Adjusted Rand-Index**	**Mutual info score**
0.247	0.333	0.283	0.168	0.171

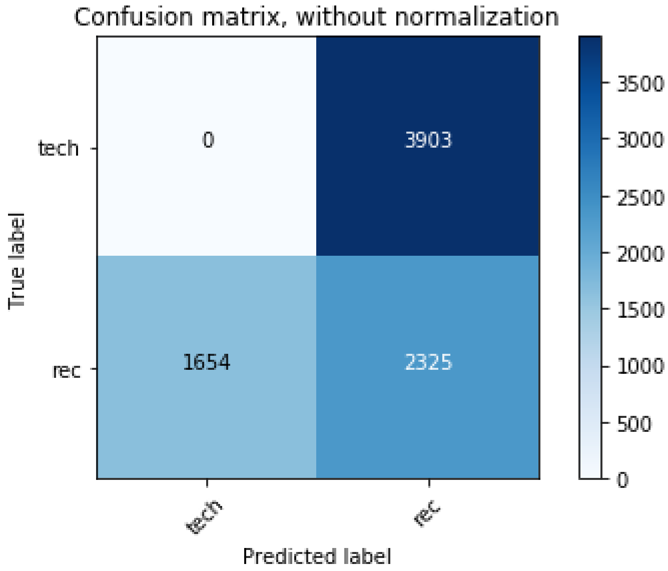


Fig 10: Contingency Matrix for LSI, r = 300

Table 11: Measurement for NMF, r = 1

Homogeneity	**Completeness**	**V-measure**	**Adjusted Rand-Index**	**Mutual info score**
0.001	0.001	0.001	0.001	0.001

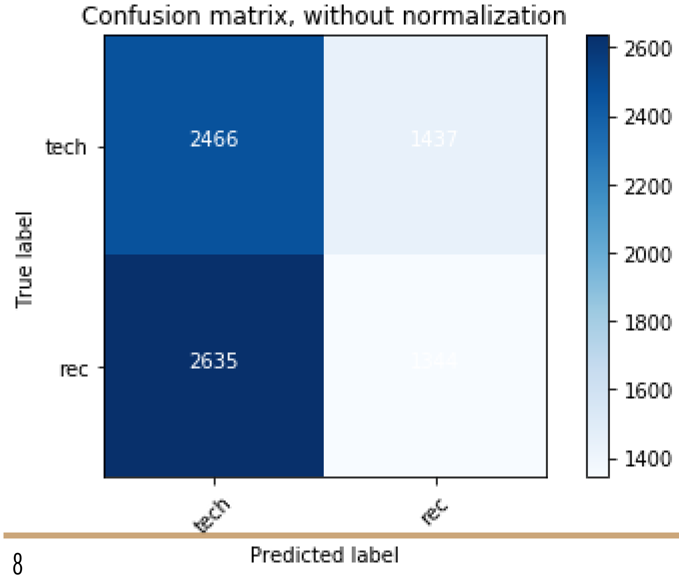


Fig 11: Contingency Matrix for NMF, r = 1

Table 12: Measurement for NMF, r = 2

Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Mutual info score
0.484	0.512	0.498	0.498	0.336

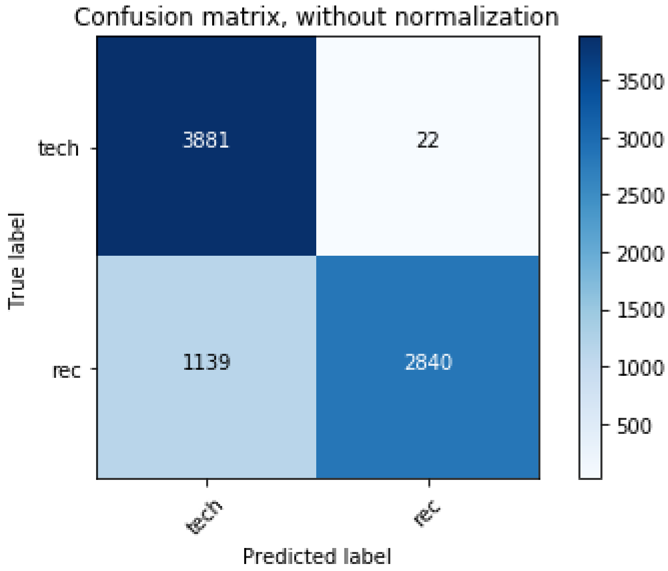


Fig 12: Contingency Matrix for NMF, r = 2

Table 13: Measurement for NMF, r = 3

Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Mutual info score
0.231	0.321	0.269	0.150	0.160

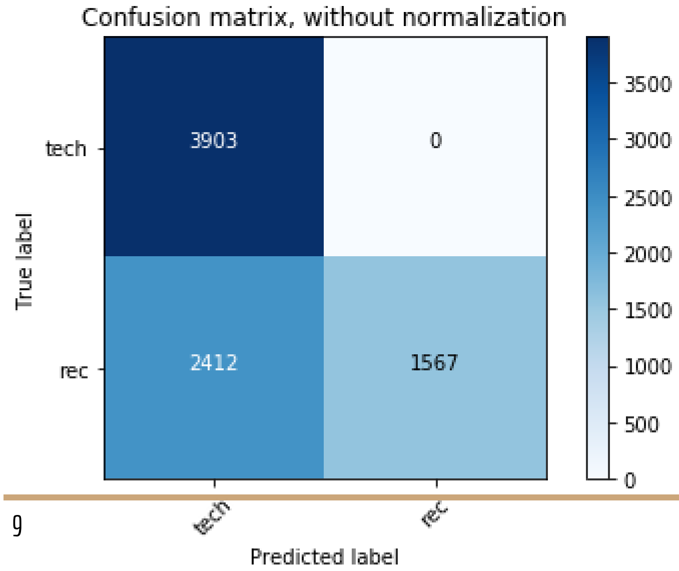


Fig 13: Contingency Matrix for NMF, r = 3

Table 14: Measurement for NMF, r = 5

Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Mutual info score
0.236	0.305	0.266	0.182	0.163

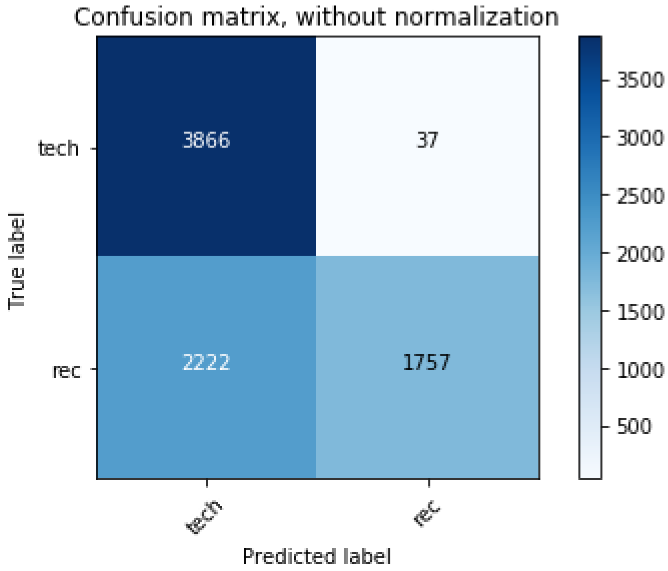


Fig 14: Contingency Matrix for NMF, r = 5

Table 15: Measurement for NMF, r = 10

Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Mutual info score
0.213	0.292	0.247	0.149	0.148

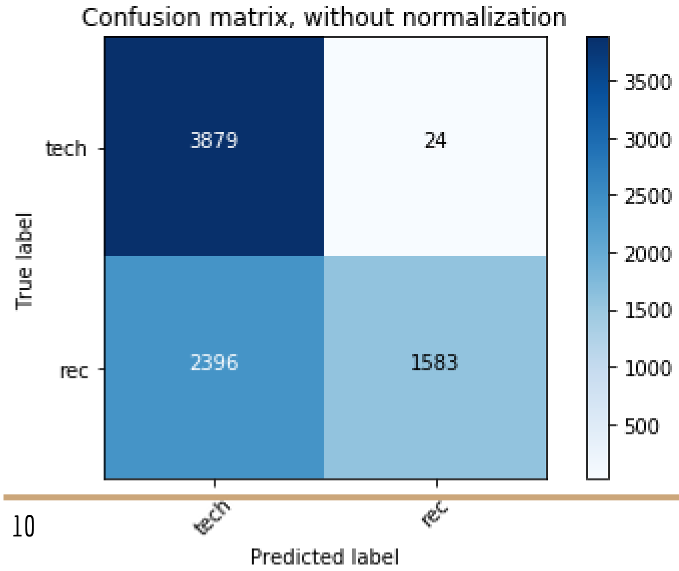


Fig 15: Contingency Matrix for NMF, r = 10

Table 16: Measurement for NMF, r = 20

Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Mutual info score
0.001	0.015	0.001	0	0

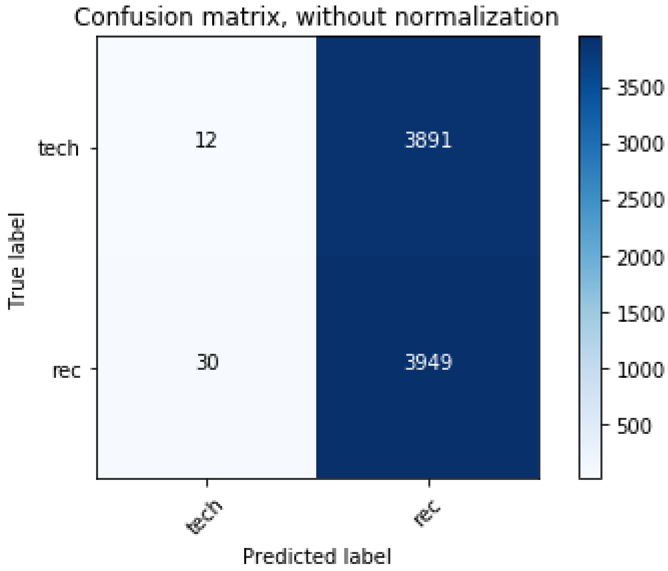


Fig 16: Contingency Matrix for NMF, r = 20

Table 16: Measurement for NMF, r = 50

Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Mutual info score
0.005	0.107	0.009	0	0.003

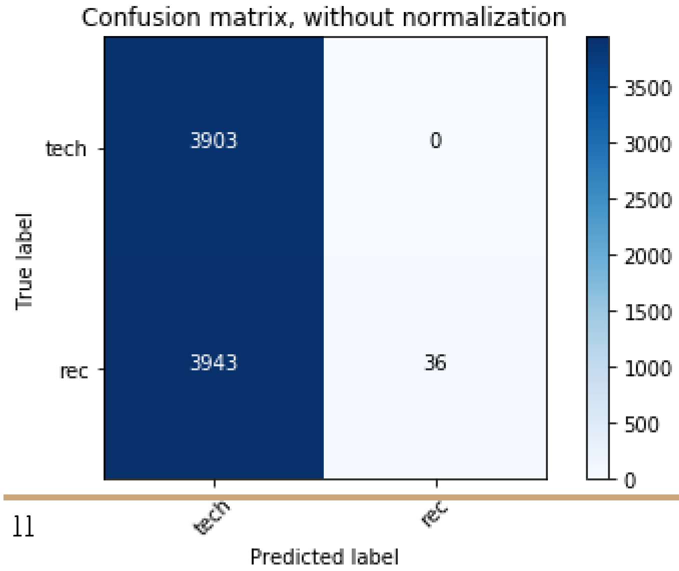


Fig 16: Contingency Matrix for NMF, r = 50

Table 17: Measurement for NMF, r = 100

Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Mutual info score
0.012	0.130	0.022	0.001	0.008

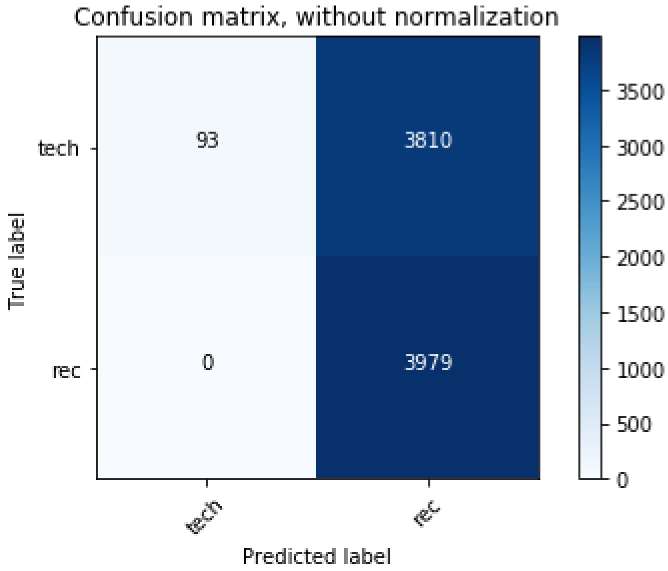


Fig 17: Contingency Matrix for NMF, r = 100

Table 18: Measurement for NMF, r = 300

Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Mutual info score
0.003	0.098	0.005	0	0.002

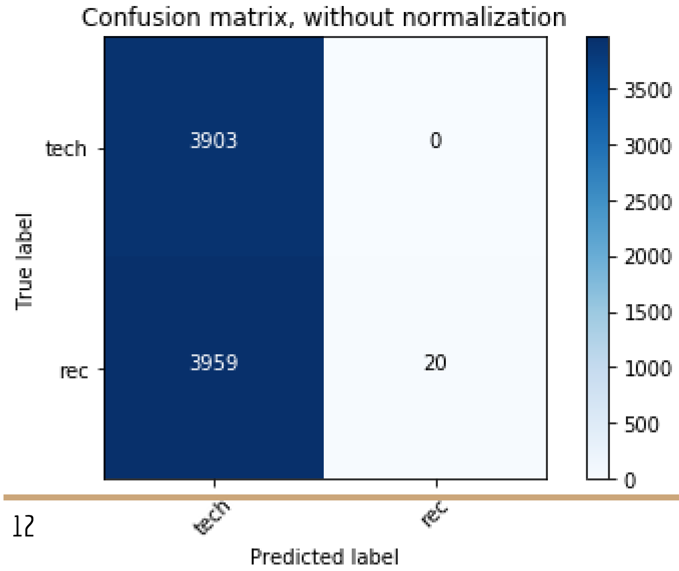


Fig 18: Contingency Matrix for NMF, r = 300

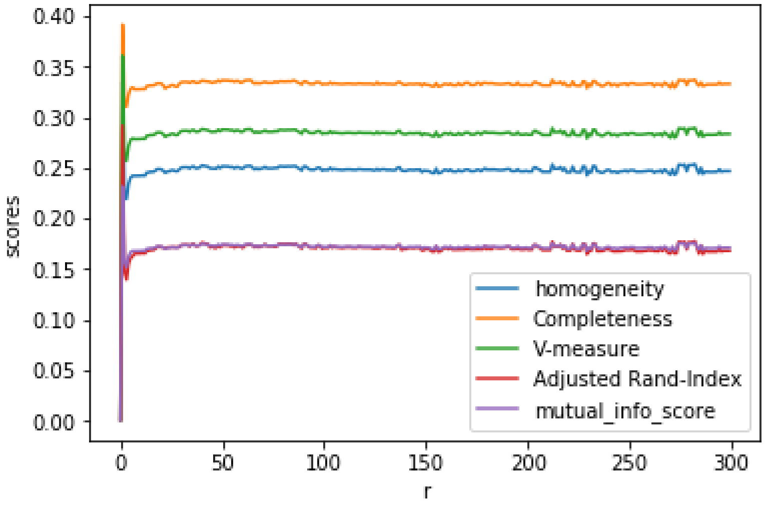


Fig 19: Scores vs r, LSI

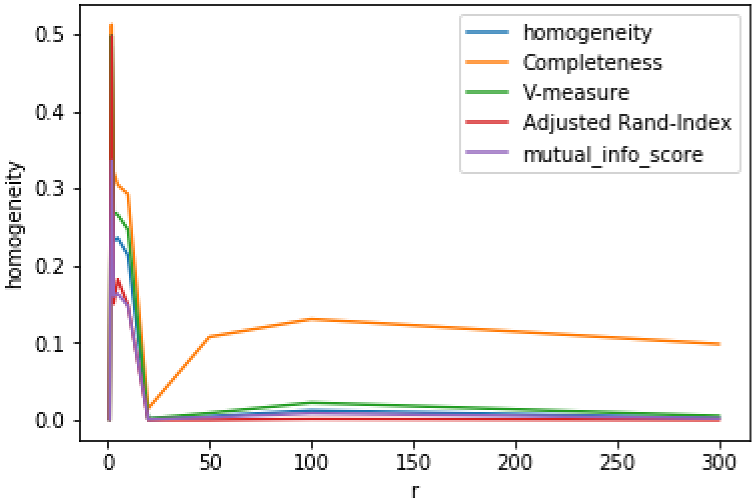
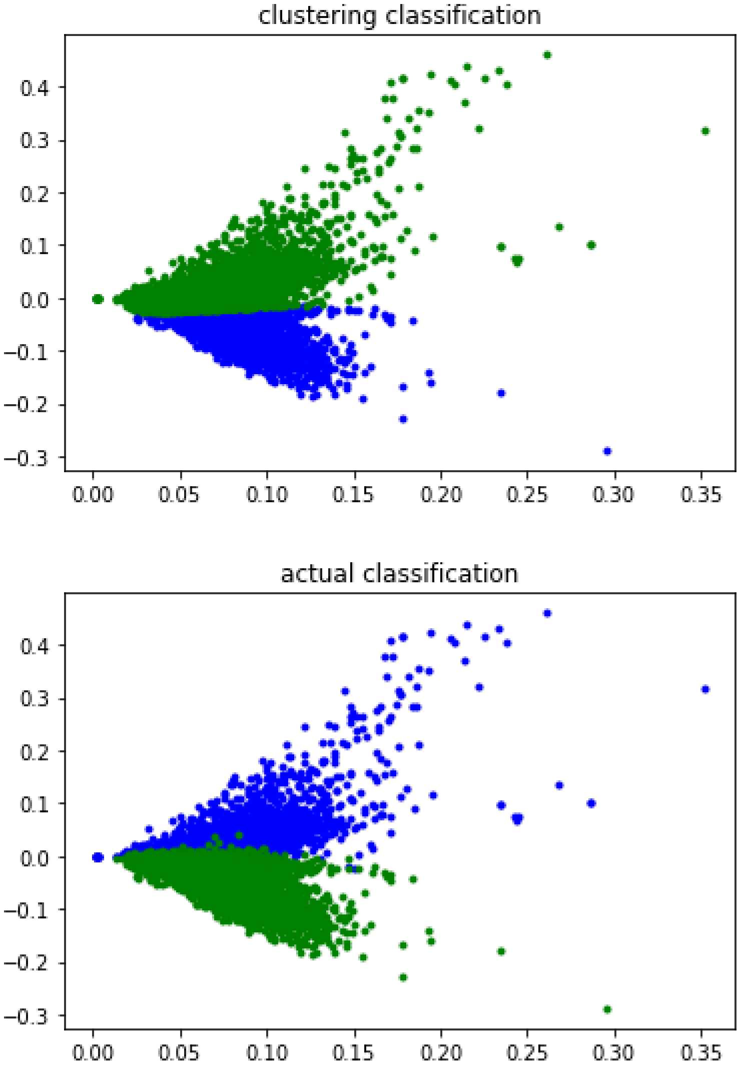


Fig 20: Scores vs r, NMF

4. Visualizing with Methods of Improvement

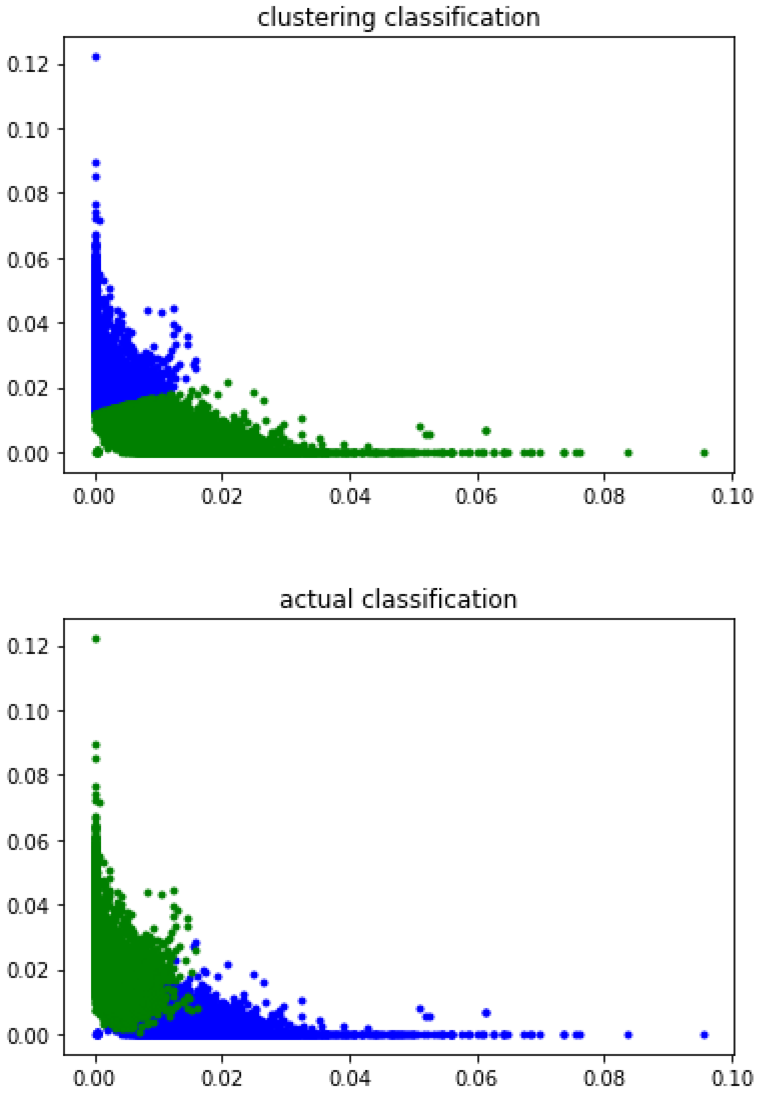
a) Visualizing on 2 dimensional plane

From the projection, we could easily tell whether the clustering perform well. In this case, LSI did clustering better than NMF. We could also see that NMF data are converged to a very specific region and therefore harder to differentiate. Again, note that the colors might be swapped because the clustering algorithm does not have the knowledge of actual labels.


clustering classification

actual classification

Fig 21: 2-D Visualization of Data Clustering for LSI, r=2


clustering classification

actual classification

Fig 22: 2-D Visualization of Data Clustering for NMF, r = 2

b) Three methods of improvement

In this part, we tried improving the clustering performance by normalizing and log transforming the data vectors. The normalization is applied on both LSI and NMF data, while the log transformation, and a mixture of both techniques, were only applied on NMF-reduced data. The model performed much better because the non-linear transformation like log could bring the data points to the normal distribution, compensating the inaccurate assumption in the previous calculations.

LSI with Normalization

Table 19: Measurement for LSI, r = 2 with Normalization

Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Mutual info score
0.142	0.202	0.166	0.109	0.098

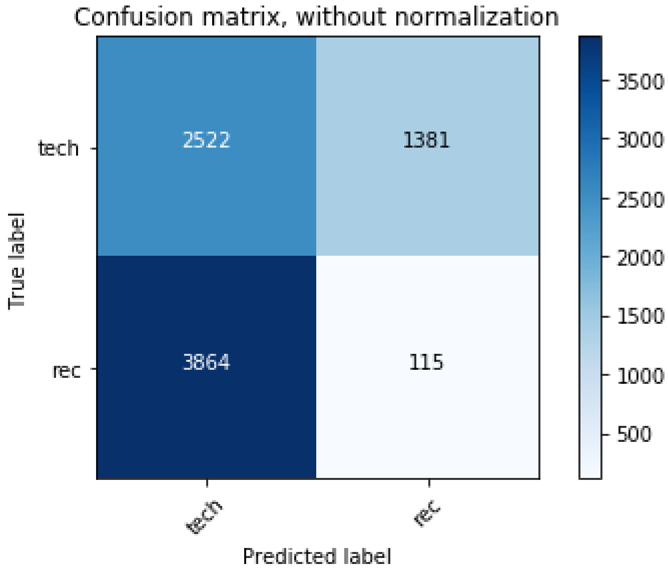
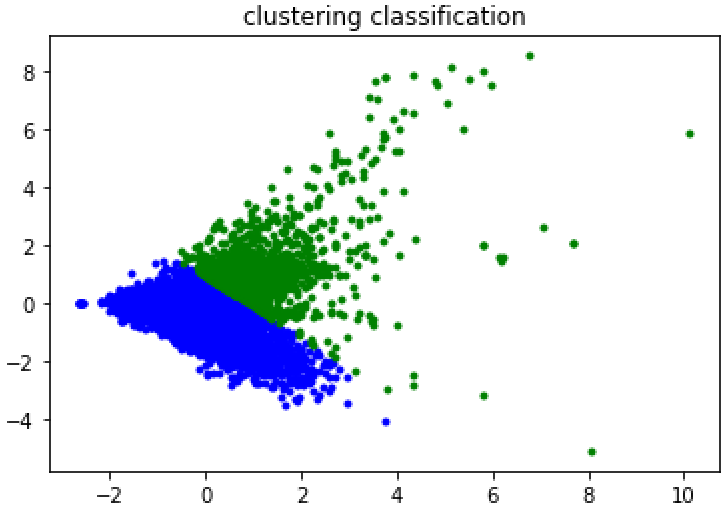


Fig 23: Contingency Matrix for LSI, r = 2 with Normalization



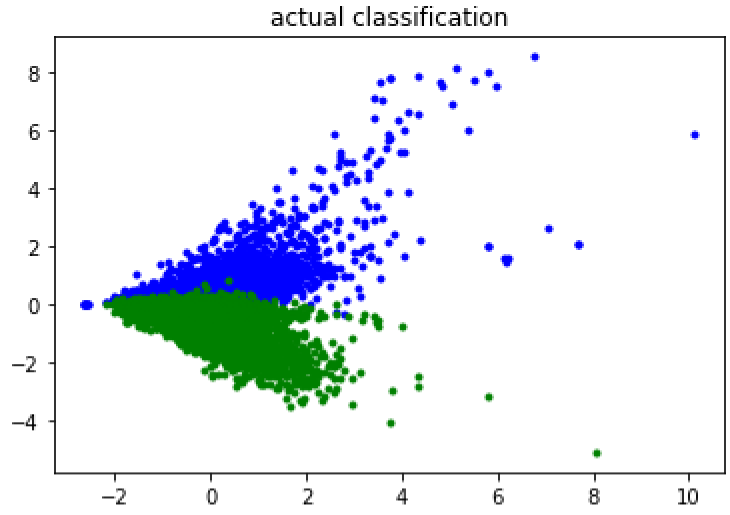


Fig 24: 2-D Visualization of Data Clustering for LSI, r = 2 with Normalization

NMF with Normalization

Table 20: Measurement for NMF, r = 2 with Normalization

Homogeneity	**Completeness**	**V-measure**	**Adjusted Rand-Index**	**Mutual info score**
0.751	0.751	0.751	0.839	0.520

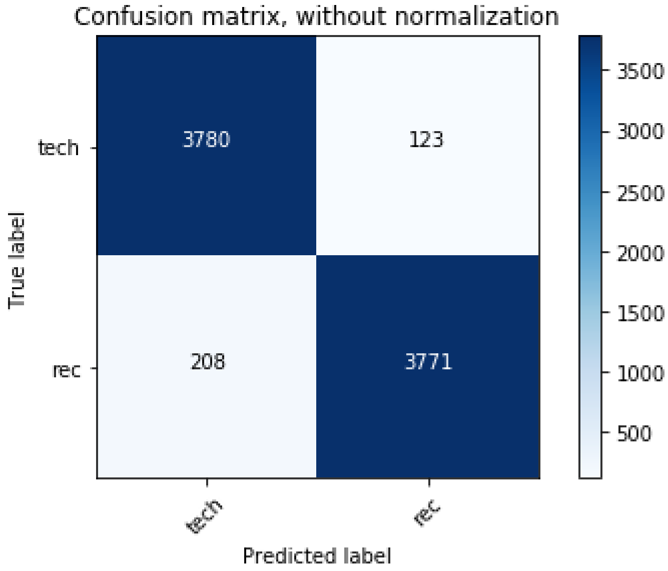


Fig 25: Contingency Matrix for NMF, r = 2 with Normalization

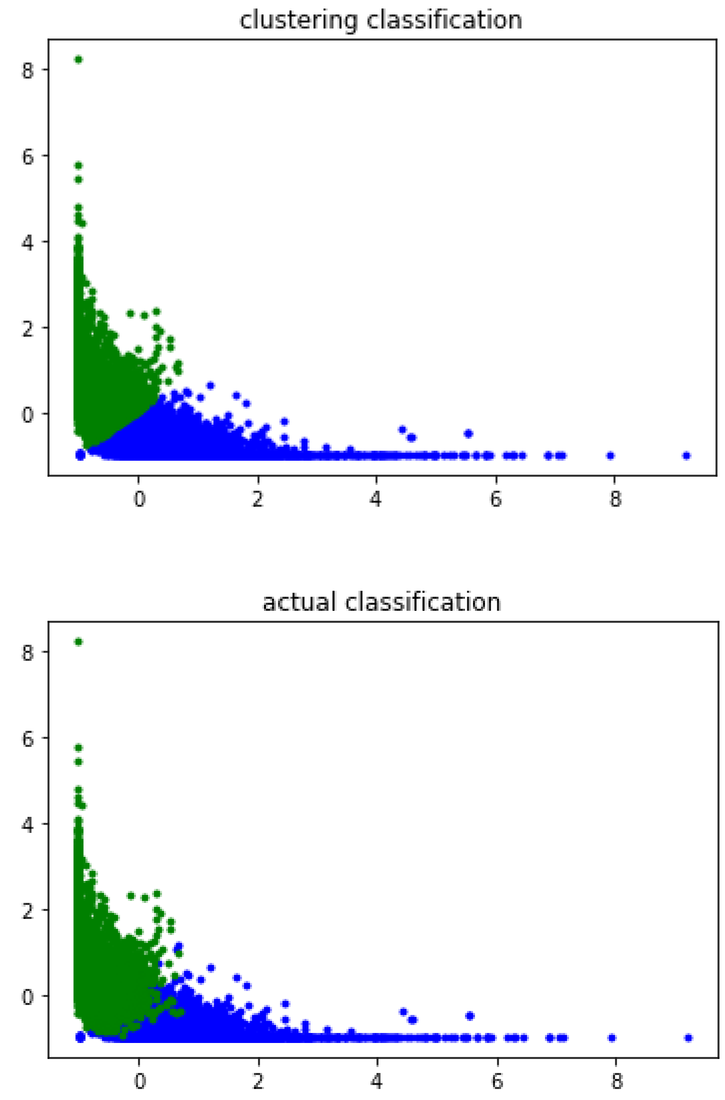


Fig 26: 2-D Visualization of Data Clustering for NMF, r = 2 with Normalization

NMF with Log transformation

Table 21: Measurement for NMF, r = 2 with Log Transformation

Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Mutual info score
0.749	0.749	0.749	0.839	0.519

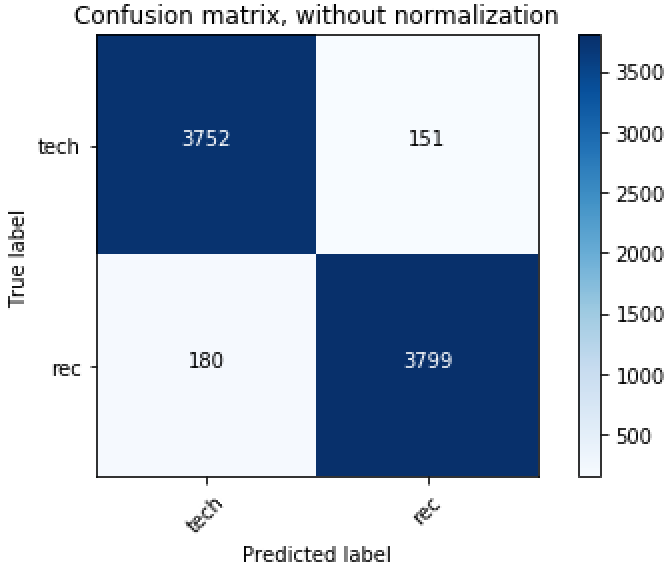


Fig 27: Contingency Matrix for NMF, r = 2 with Log Transformation

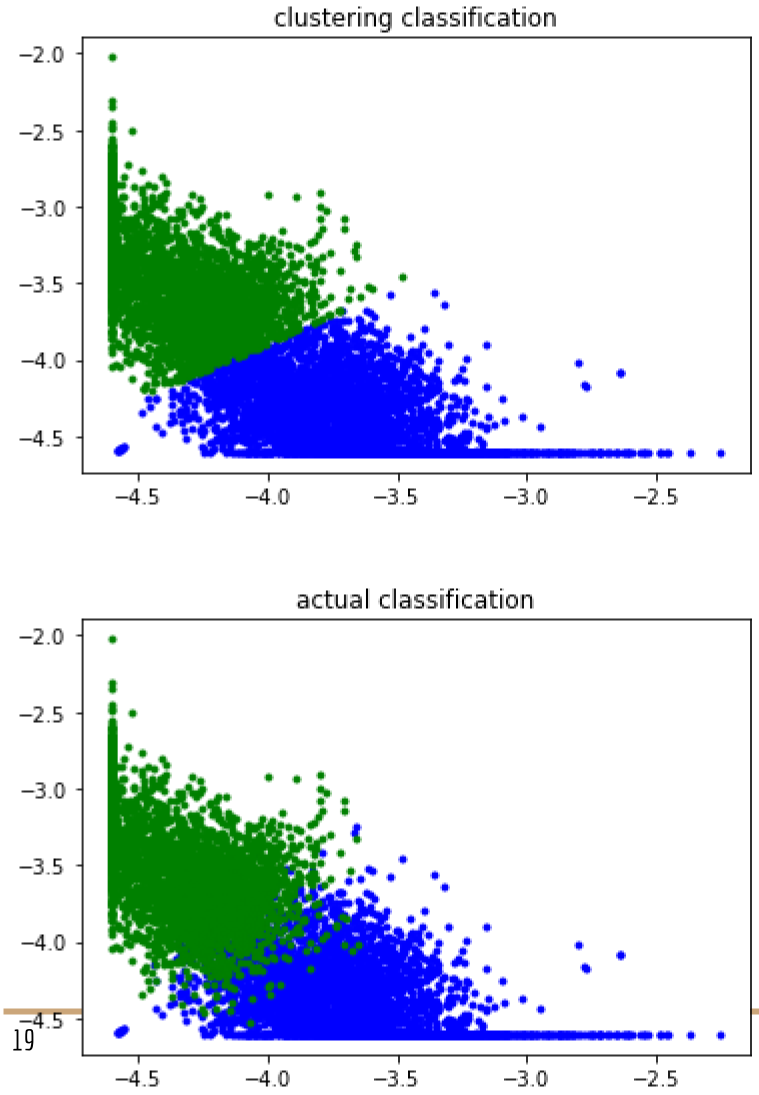


Fig 28: 2-D Visualization of Data Clustering for NMF, r = 2 with Log Transformation

NMF with Normalization -> Log transformation

Table 22: Measurement for NMF, r = 2 with Normalization -> Log Transformation

Homogeneity	**Completeness**	**V-measure**	**Adjusted Rand-Index**	**Mutual info score**
0.748	0.748	0.748	0.837	0.518

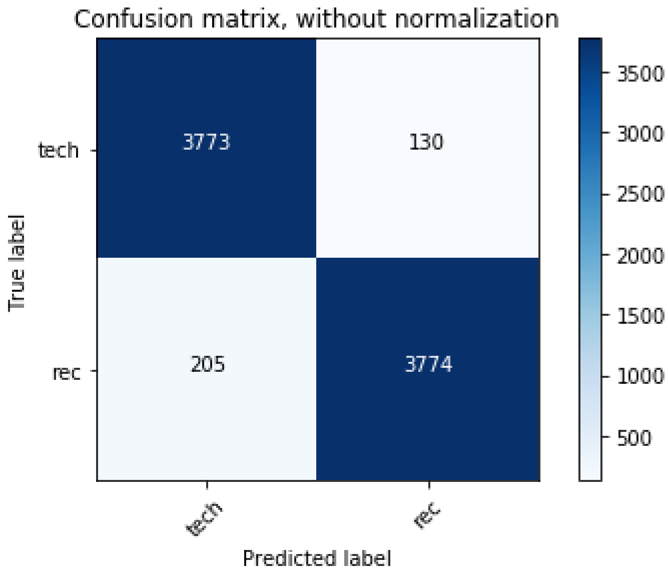
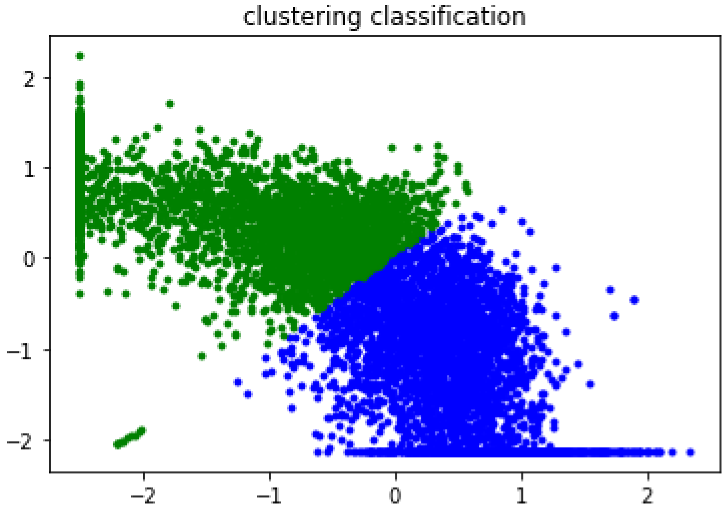


Fig 29: Contingency Matrix for NMF, r = 2 with Normalization -> Log Transform



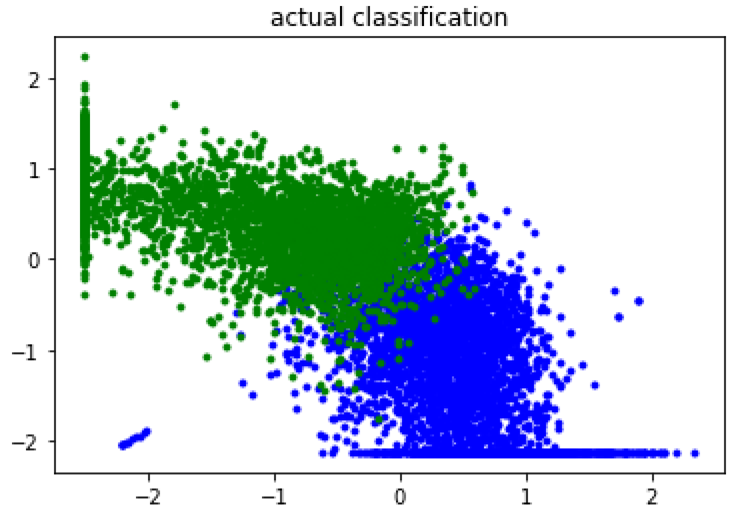


Fig 30: 2-D Visualization of Data Clustering for NMF, r = 2 with Normalization -> Log Transform

NMF with Log transformation -> Normalization

Table 23: Measurement for NMF, r = 2 with Log Transformation -> Normalization

Homogeneity	**Completeness**	**V-measure**	**Adjusted Rand-Index**	**Mutual info score**
0.716	0.720	0.718	0.796	0.496

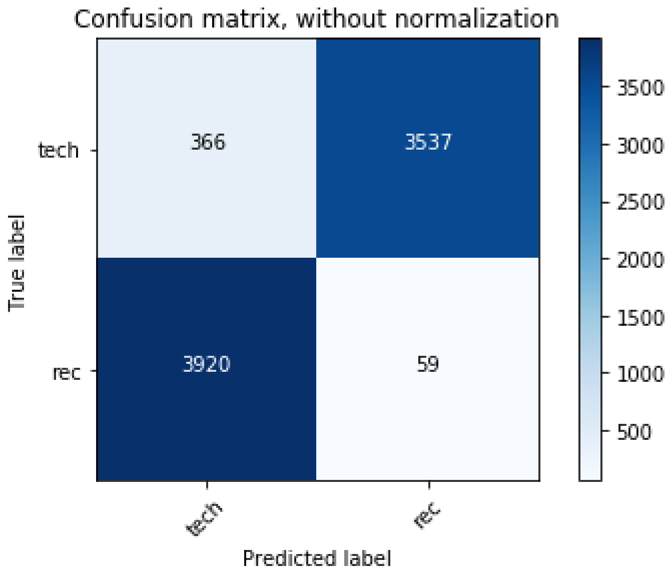


Fig 31: Contingency Matrix for NMF, r = 2 with Log Transform -> Normalization

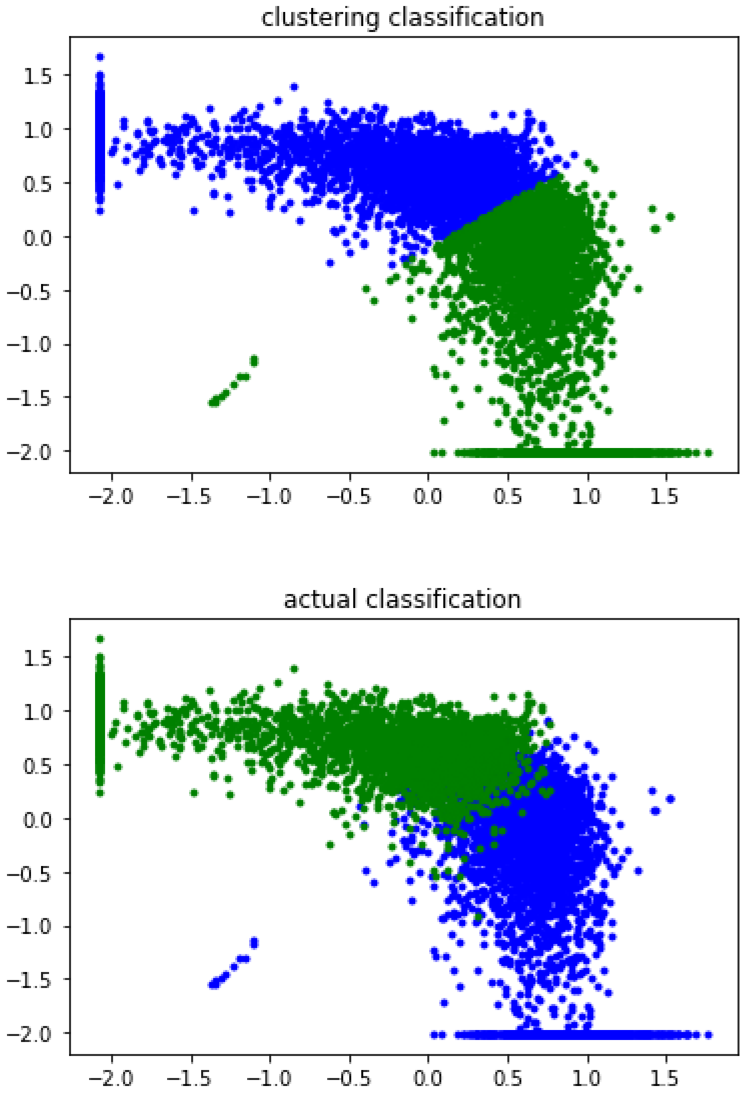


Fig 32: 2-D Visualization of Data Clustering for NMF, r = 2 with Log Transform -> Normalization

5. Clustering on 20 Clases

So far, we have been dealing with clustering the data points into two classes. In this part, we explore clustering the whole *sklearn.fetch_20newsgroups* into 20 subclasses. The TF-IDF matrix is of dimension (18846, 52268). Similar as above with two classes, we want to find the optimal reduced dimension r for the dataset.

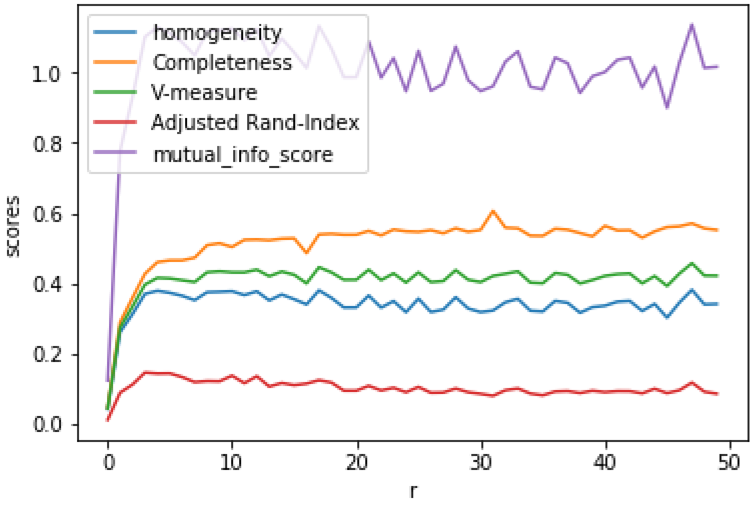


Fig 33: Scores vs r, LSI multiclass

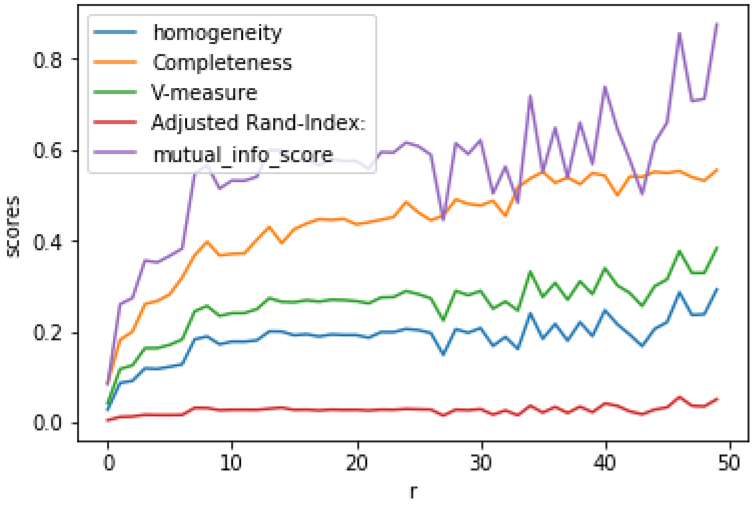


Fig 34: Scores vs r, NMF multiclass

Per the graph, we set the r_LSI = 5, and r_NMF = 8. The choice of r is made such that the score is reasonably high with a low dimension for the sake of computation. The results are recorded below for different optimization methods. However, the confusion matrix for this one is trivial because the cluster are not assigned in order.

NMF with Normalization

Table 24: Measurement for NMF, r = 8 with Normalization

Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Mutual info score
0.343	0.445	0.387	0.121	1.025

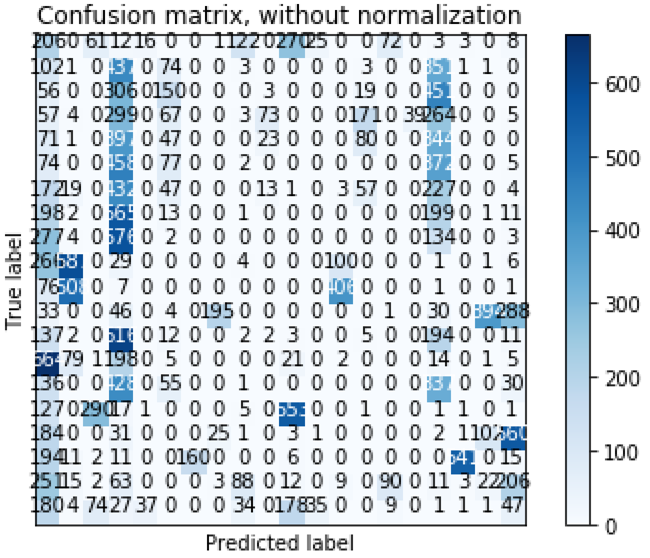


Fig 35: 2-D Visualization of Data Clustering for NMF, r = 8 with Normalization

NMF with Log Transformation

Table 25: Measurement for NMF, r = 8 with Log Transformation

Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Mutual info score
0.361	0.426	0.391	0.130	1.079

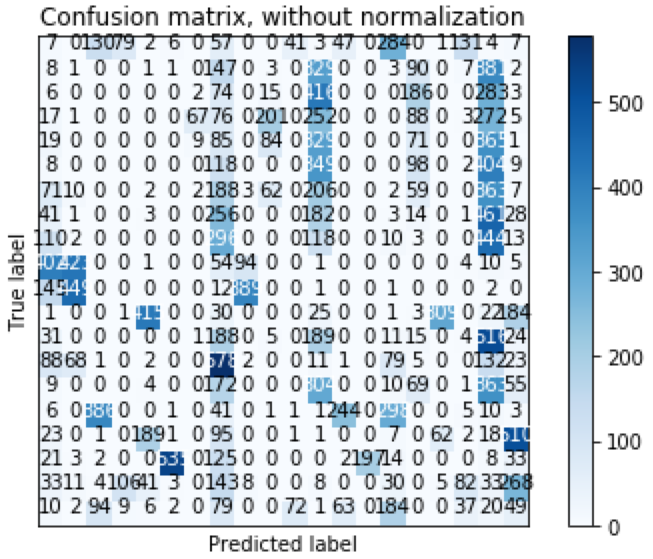


Fig 36: 2-D Visualization of Data Clustering for NMF, r = 8 with Log Transform

NMF with Normalization -> Log transformation

Table 26: Measurement for NMF, r = 8 with Normalization -> Log Transformation

Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Mutual info score
0.407	0.445	0.425	0.183	1.216

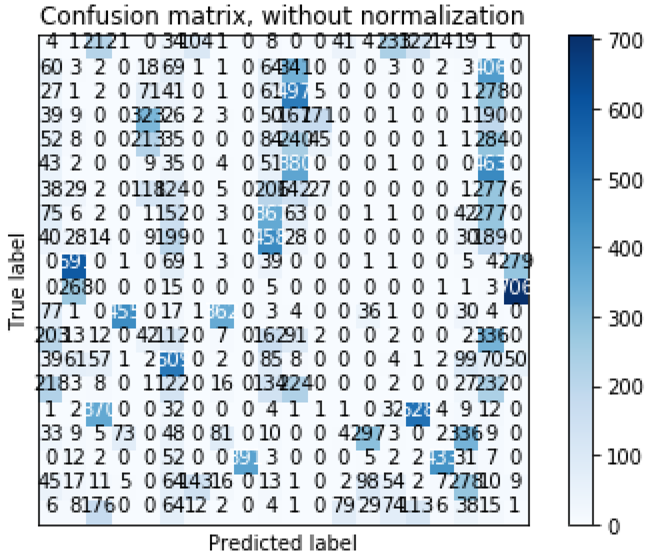


Fig 37: 2-D Visualization of Data Clustering for NMF, r = 8 with Normalization -> Log Transformation

LSI without Normalization

Table 27: Measurement for LSI, r = 5 without Normalization

Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Mutual info score
0.378	0.460	0.415	0.142	1.130

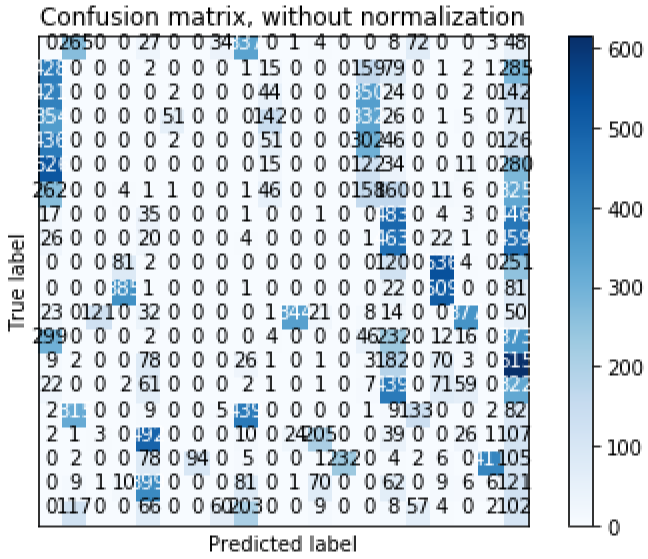


Fig 38: 2-D Visualization of Data Cluster for LSI, r = 5 without Normalization