ECE 219 Project 2 Report
Clustering
Winter 2018

Hengyu Lou (005035476)
email: hylou@ucla.edu

Zhonglin Zhang (005030520)
email: evanzhang@ucla.edu

February 12, 2018

1 Introduction

Clustering algorithms are unsupervised methods for finding groups of data points that have similar representations in a proper space. Clustering differs from classification in that no a priori labeling (grouping) of the data points is available.

2 Dataset and Problem Statement

2.1 Building the TF-IDF matrix

Here, we transformd the documents into TF-IDF vectors. In detail, we used min_df = 3 and excluded the stopwords without stemming. The dimension of the TF-IDF matrix is (7882, 27768).

2.2 Applying K-means Clustering

(a) The contingency table is shown below

Table 1: Contingency Table

	Cluster 1	**Cluster 2**
Class 1	4	3899
Class 2	1717	2262

(b) To make a concrete comparison of different clustering results, the measures are shown below
– Homogeneity: 0.253
– Completeness: 0.335
– V-measure: 0.288
– Adjusted Rand-Index: 0.181
– Adjusted Mutual Info Score: 0.253

2.3 Preprocess the Data

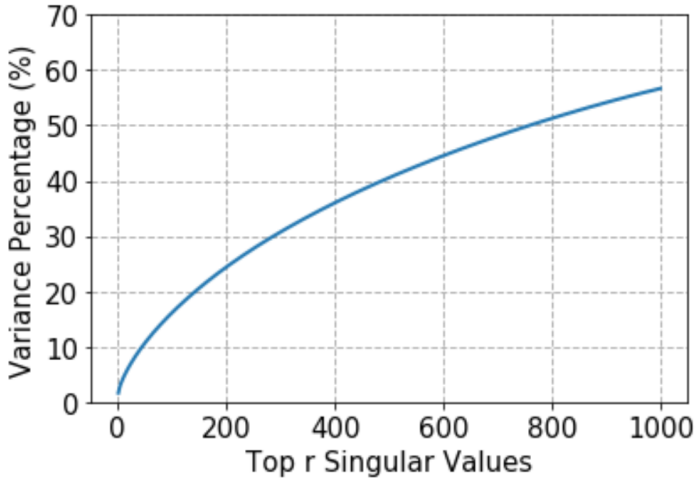

Figure 1: Varience Percent V.S. r

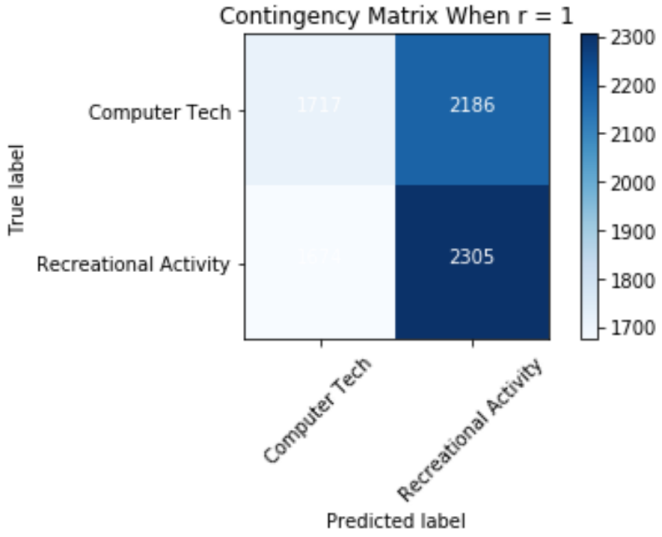

Figure 2: LSI Contingency Matrix with r = 1

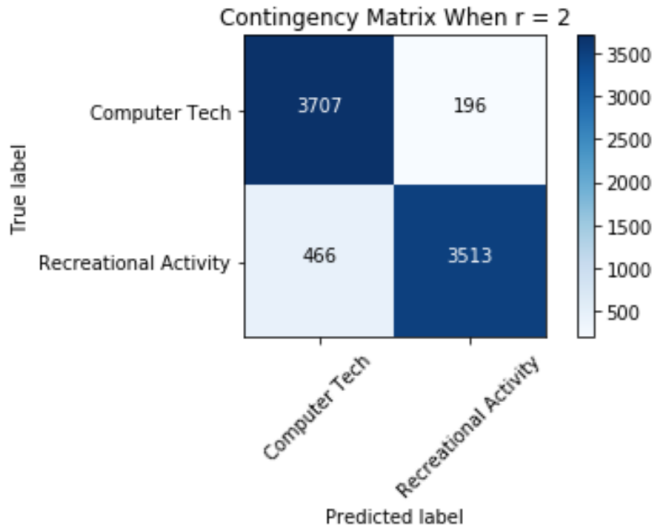

Figure 3: LSI Contingency Matrix with r = 2

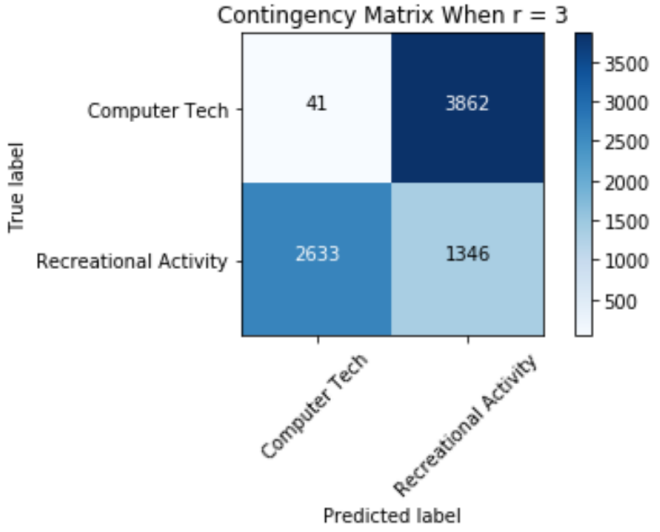

Figure 4: LSI Contingency Matrix with r = 3

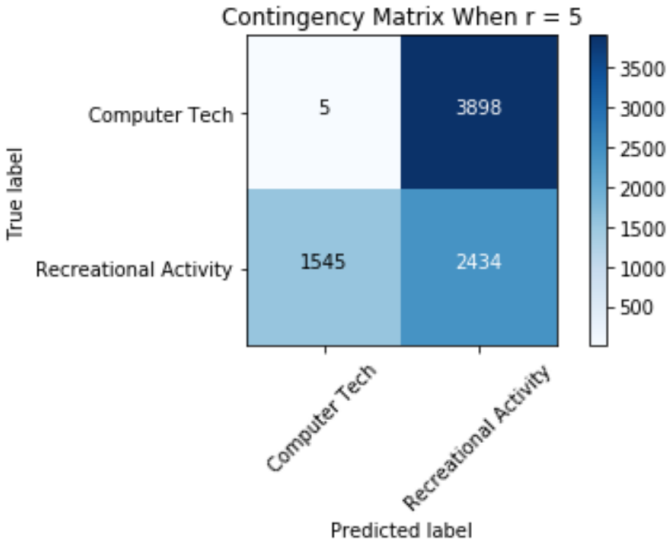

Figure 5: LSI Contingency Matrix with r = 5

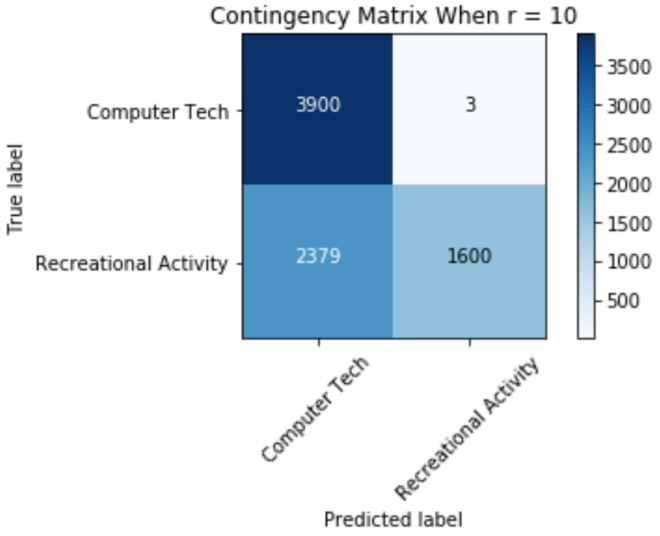

Figure 6: LSI Contingency Matrix with r = 10

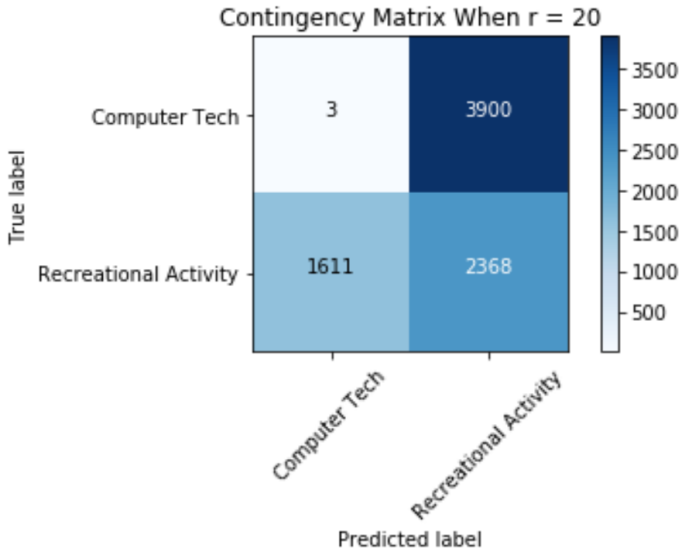

Figure 7: LSI Contingency Matrix with r = 20

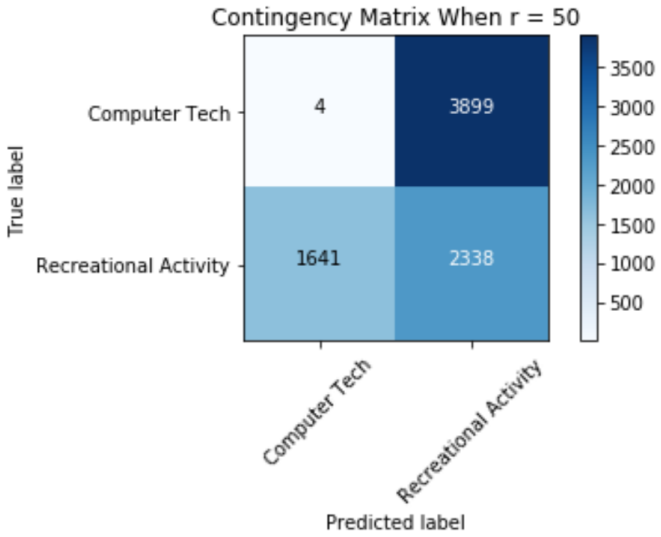

Figure 8: LSI Contingency Matrix with r = 50

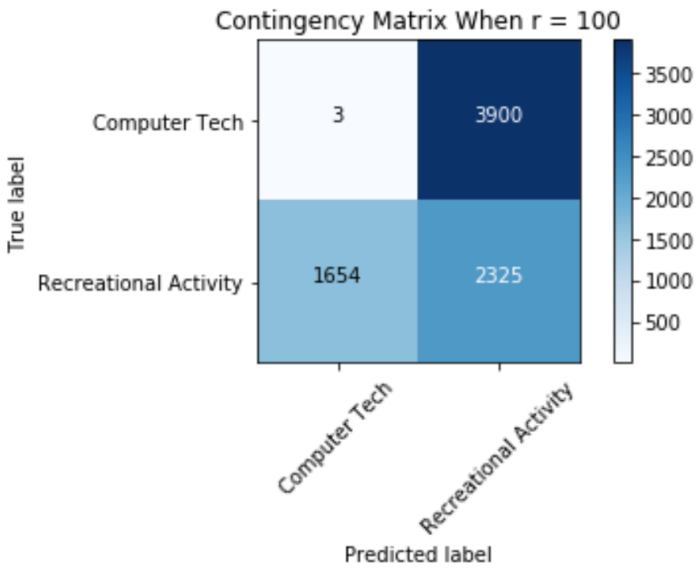

Figure 9: LSI Contingency Matrix with r = 100

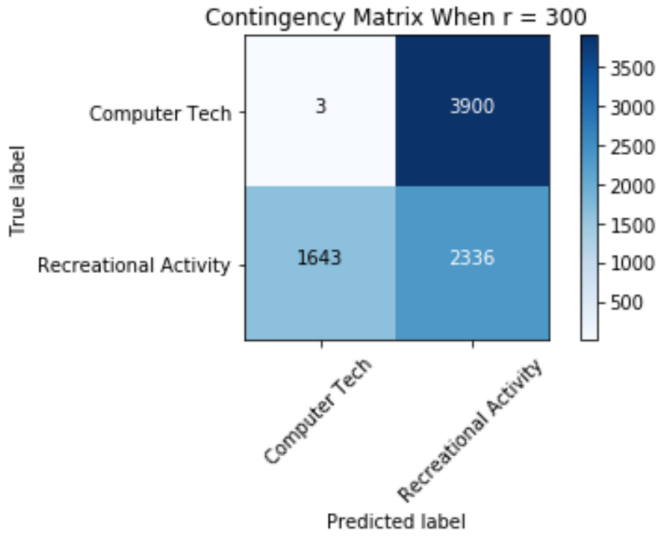

Figure 10: LSI Contingency Matrix with r = 300

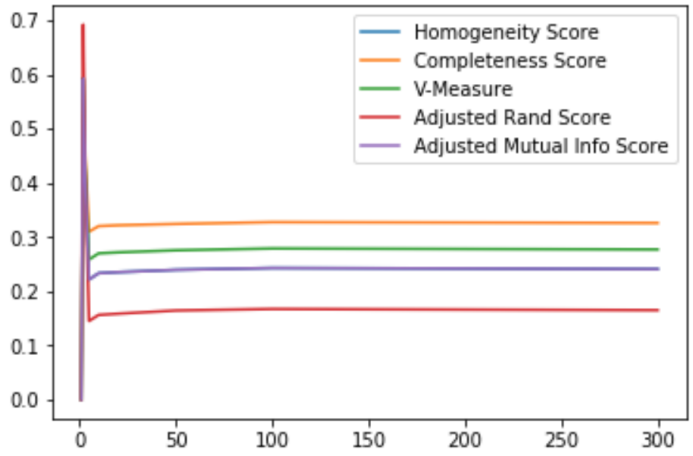

Figure 11: LSI Measures V.S. Dimensions

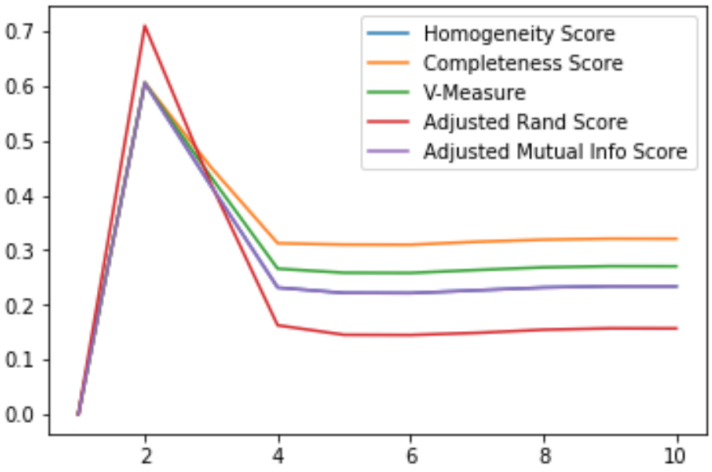

Figure 12: LSI Measures V.S. Dimensions (only from 1 to 10)

```
Contingency matrix
[[2189 1714]
 [2311 1668]]
```

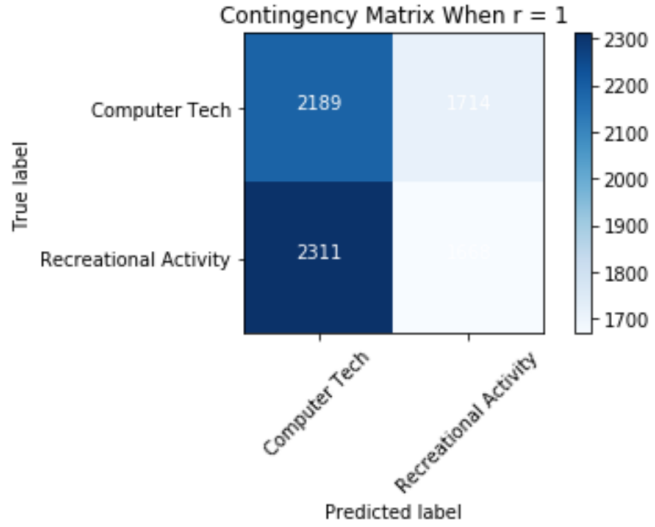

Figure 13: NMF Contingency Matrix with r = 1

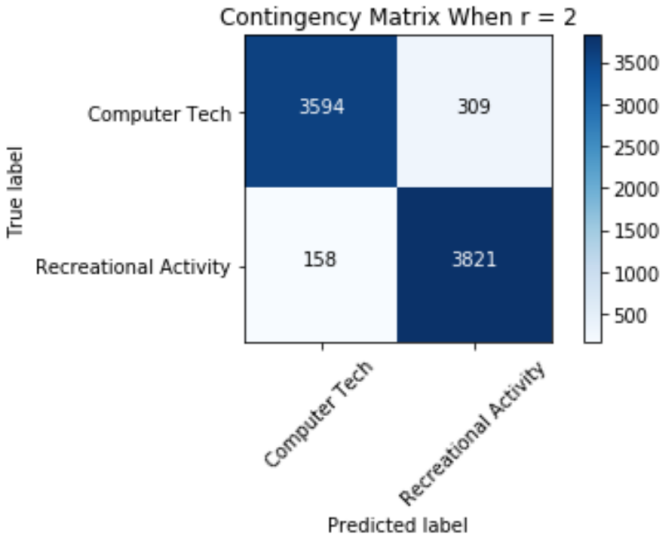

Figure 14: NMF Contingency Matrix with r = 2

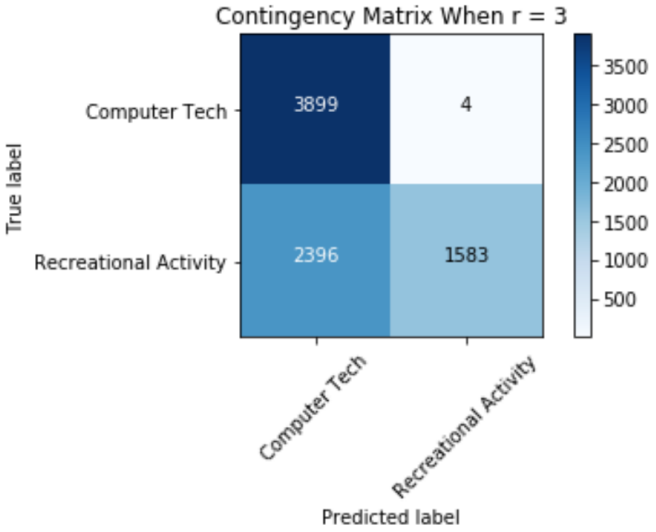

Figure 15: NMF Contingency Matrix with r = 3

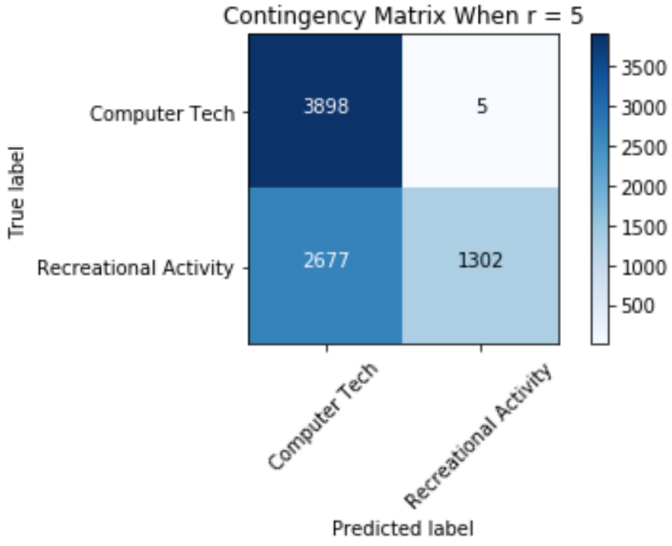

Figure 16: NMF Contingency Matrix with r = 5

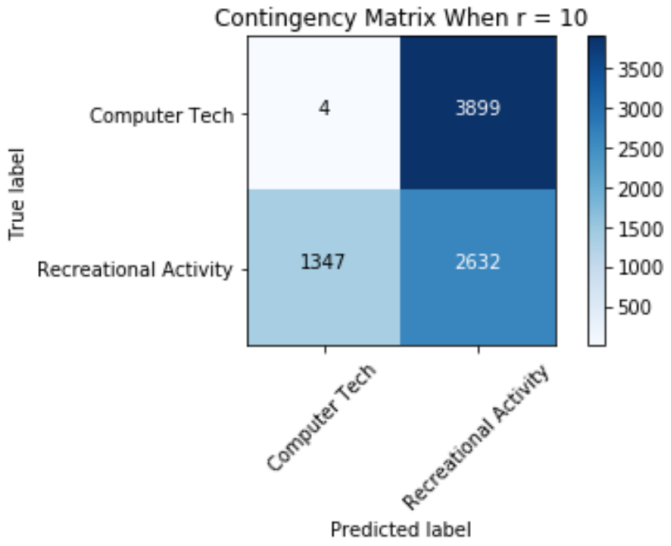

Figure 17: NMF Contingency Matrix with r = 10

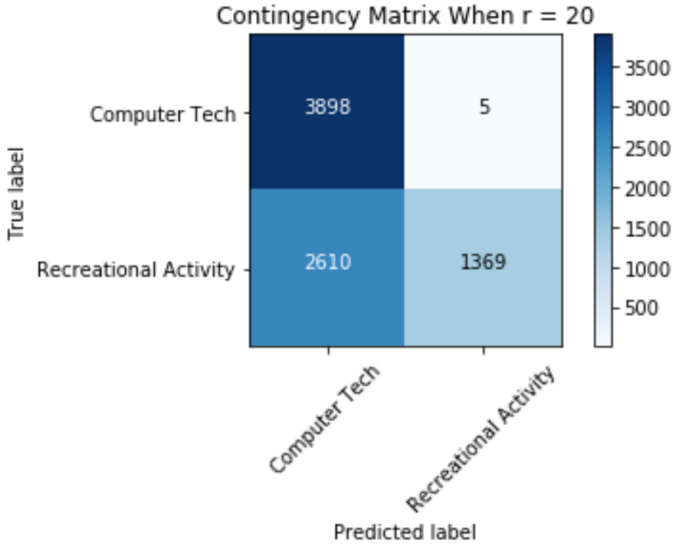

Figure 18: NMF Contingency Matrix with r = 20

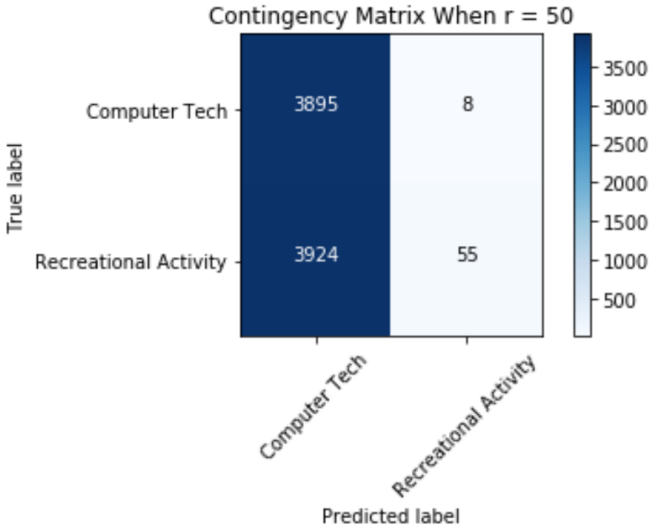

Figure 19: NMF Contingency Matrix with r = 50

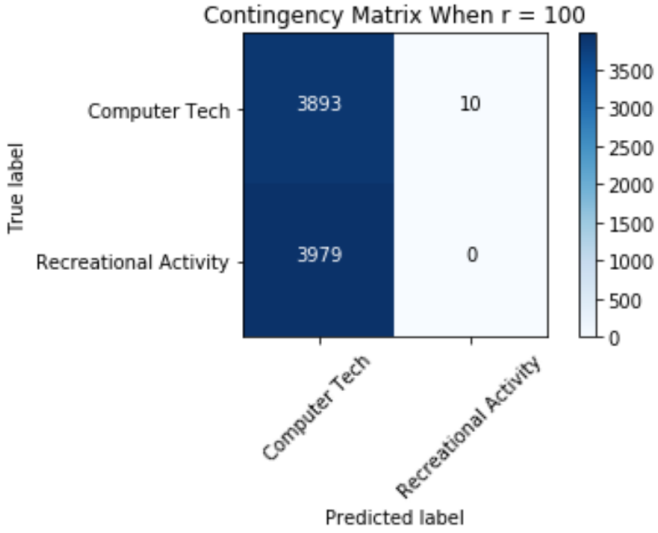

Figure 20: NMF Contingency Matrix with r = 100

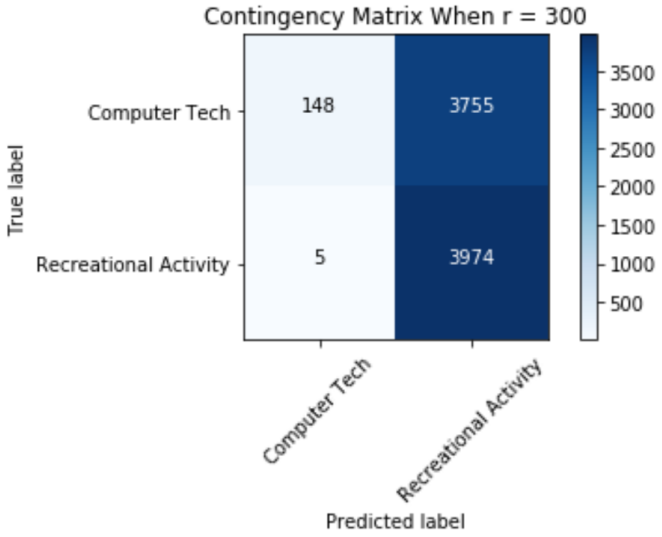

Figure 21: NMF Contingency Matrix with r = 300

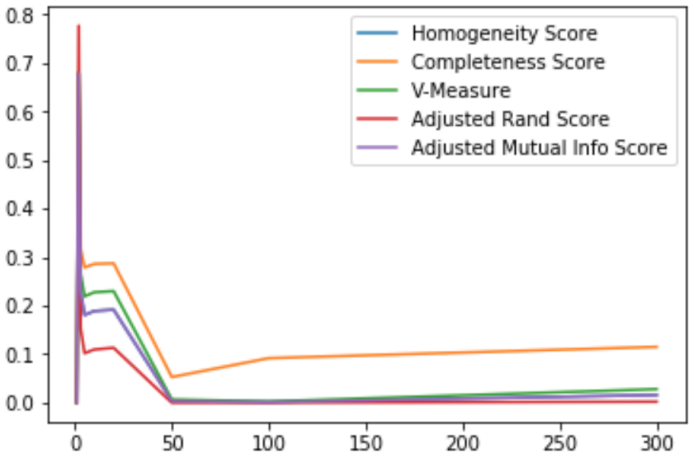

Figure 22: NMF Measures V.S. Dimensions

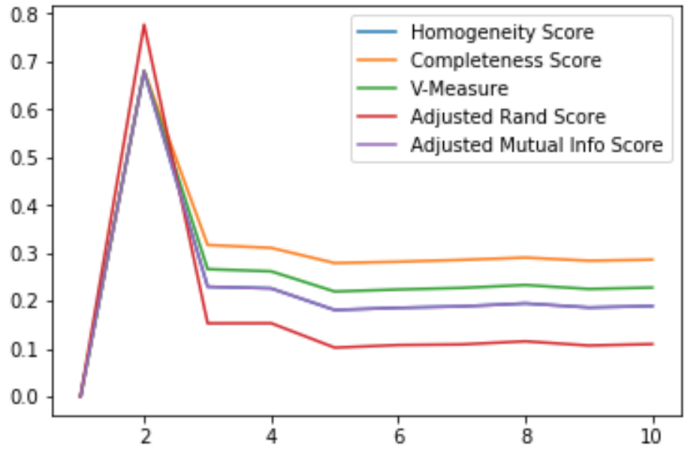

Figure 23: NMF Measures V.S. Dimensions (only from 1 to 10)

The best dimension for LSI is 2 and the best dimension for NMF is also 2.

Table 2: Contingency Table

	Homogeneity Score	**Completeness Score**	**V-measure**	**Adjusted Rand**	**Adjusted Mutual Info**
LSI	0.605	0.606	0.605	0.710	0.605
NMF	0.679	0.680	0.680	0.777	0.679

We find there is non-monotonic behavior of the measures as r increases. The reason is that when the feature dimension is high enough, the Euclidean distance difference between different instances is not notable so that KMeans algorithm cannot classify the points clearly. Hence, it means that the high dimensions would not lead to a better clustering performance necessarily.

2.4 Visualization

(a) Visualize the performance of the case with best clustering results in the previous part your clustering by projecting final data vectors onto 2 dimensional plane and color-coding the classes.
The best dimension for both LSI and NMF is 2.

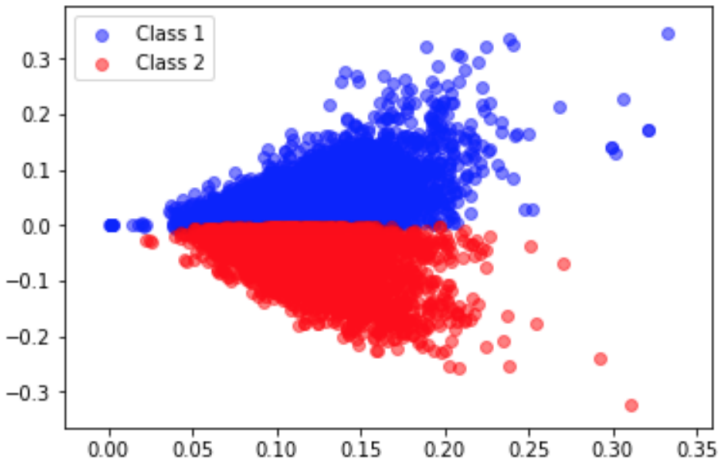

Figure 24: LSI Best Clustering Results Visualization

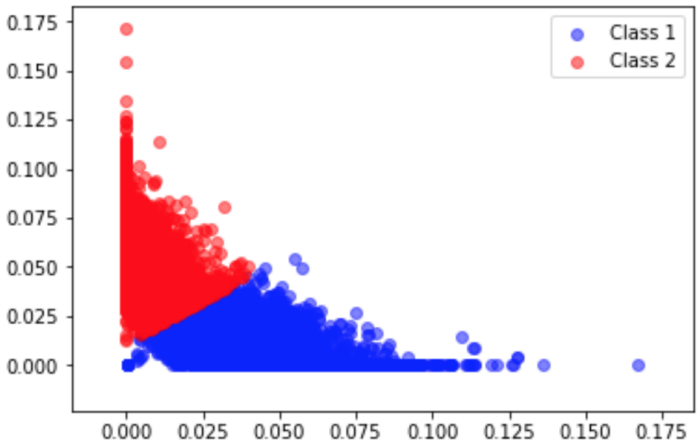

Figure 25: NMF Best Clustering Results Visualization

(b) Now try the three methods below to see whether they increase the clustering performance. Still use the best r we had in previous parts. Visualize the transformed data as in part (a). Report the new clustering measures including the contingency matrix after transformation.
(i)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).

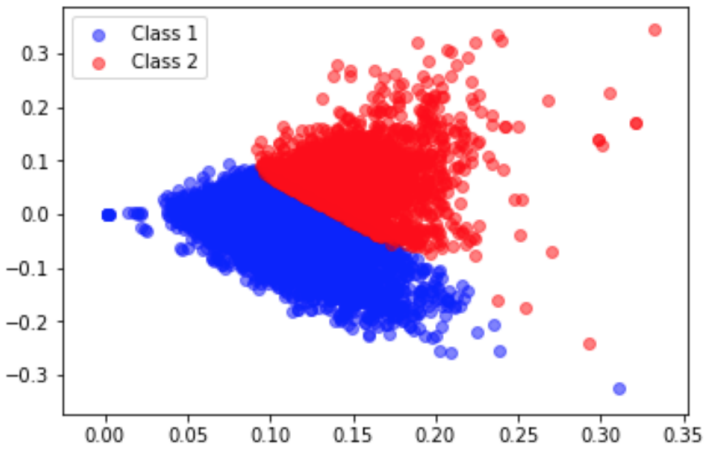

Figure 26: LSI Clustering Results Visualization with Normalizing Features

– Homogeneity: 0.236
– Completeness: 0.264
– V-measure: 0.249
– Adjusted Rand-Index: 0.255
– Adjusted Mutual Info Score: 0.236

```
Contingency matrix
[[1708 2195]
 [3737  242]]
```

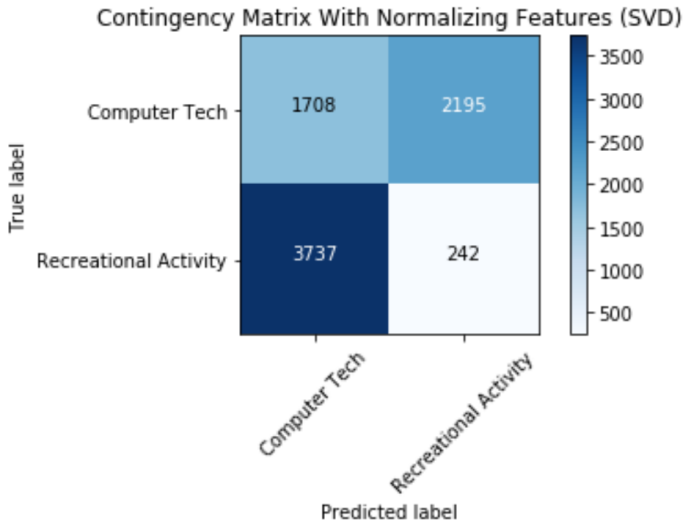

Figure 27: LSI Contingency Matrix with Normalizing Features

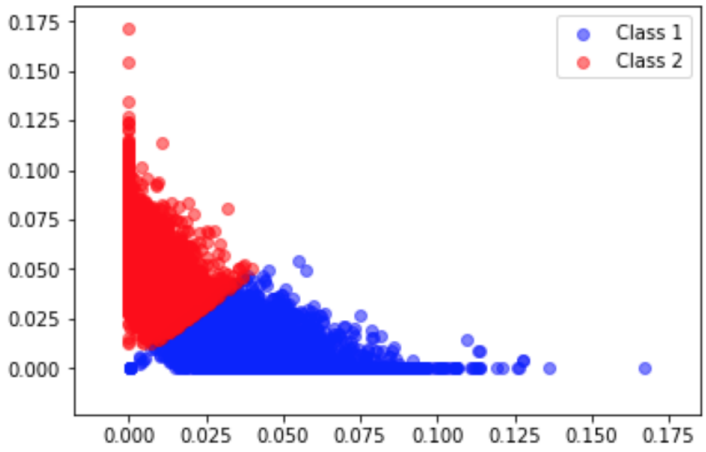

Figure 28: NMF Clustering Results Visualization with Normalizing Features

– Homogeneity: 0.683
– Completeness: 0.686
– V-measure: 0.684
– Adjusted Rand-Index: 0.773
– Adjusted Mutual Info Score: 0.683

```
Contingency matrix
[[3534  369]
 [ 106 3873]]
```

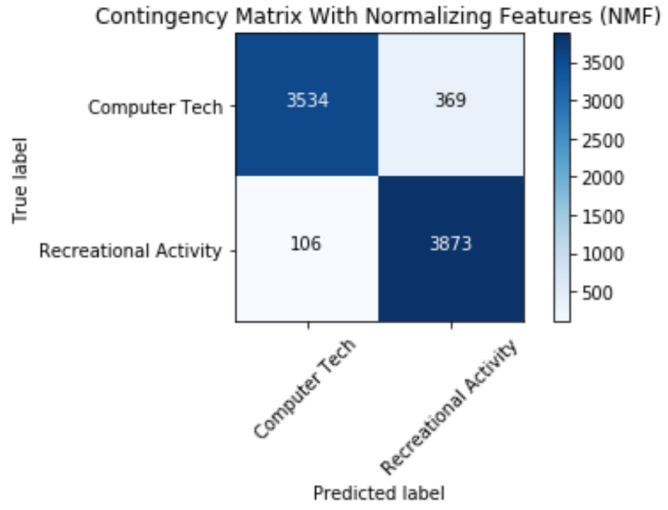

Figure 29: NMF Contingency Matrix with Normalizing Features

(ii) Applying a non-linear transformation to the data vectors only after NMF. Here we use logarithm transformation as an example.

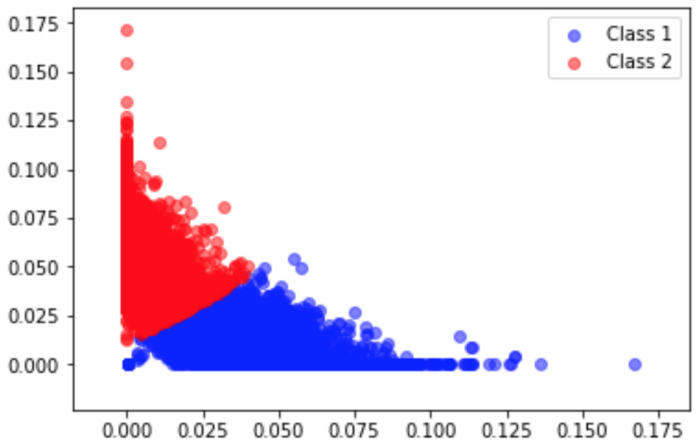

Figure 30: Clustering Results Visualization with Log Transformation

– Homogeneity: 0.675
– Completeness: 0.676
– V-measure: 0.676
– Adjusted Rand-Index: 0.773
– Adjusted Mutual Info Score: 0.675

```
Contingency matrix
[[3578  325]
 [ 152 3827]]
```

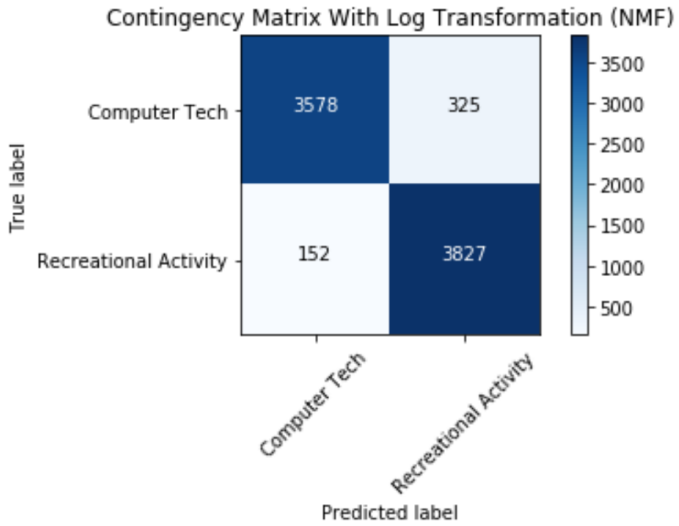

Figure 31: Contingency Matrix with Log Transformation

Here, we find that logarithm transformation indeed can increase the clustering results. According to our analysis, we think the logarithm function, as a non-linear transformation, can make the points from the same class located closer to each other in the feature space. In this way, the clustering of different classes is more obvious and therefore, the performance is better.

(iii) Now try combining both transformations (in different orders) on NMF reduced data.

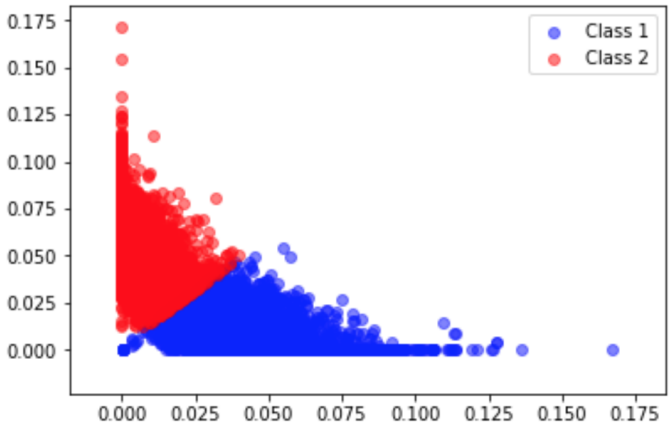

Figure 32: Clustering Results Visualization with Normalization and Log Transformation

– Homogeneity: 0.683
– Completeness: 0.686
– V-measure: 0.684
– Adjusted Rand-Index: 0.773
– Adjusted Mutual Info Score: 0.683

```
Contingency matrix
[[3534  369]
 [ 106 3873]]
```

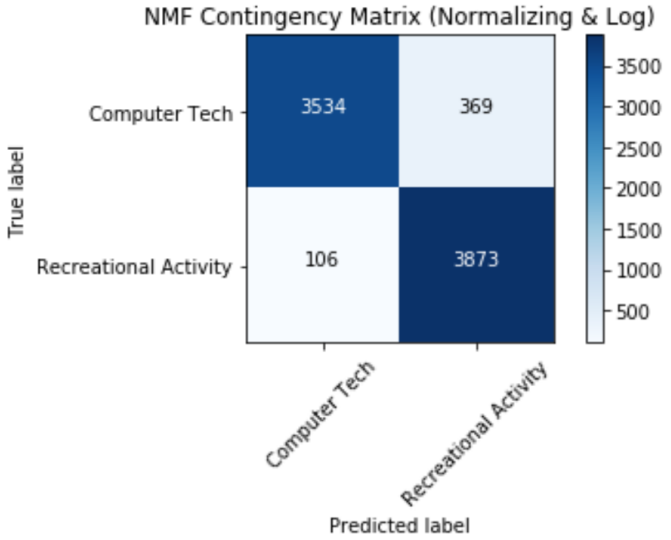

Figure 33: Contingency Matrix with Normalization and Log Transformation

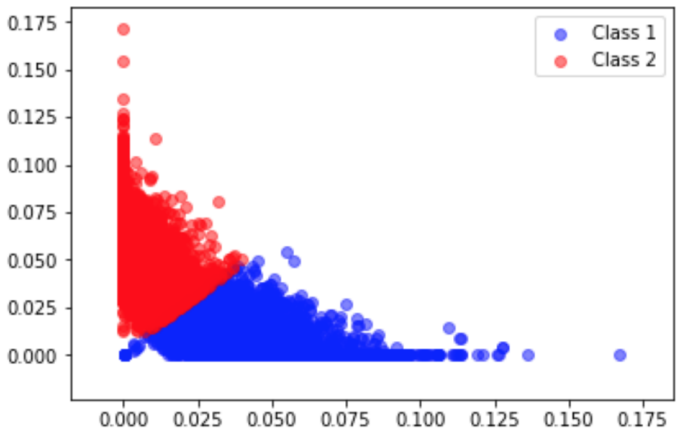

Figure 34: Clustering Results Visualization with Log Transformation and Normalization

– Homogeneity: 0.683
– Completeness: 0.686
– V-measure: 0.684
– Adjusted Rand-Index: 0.773
– Adjusted Mutual Info Score: 0.683

```
Contingency matrix
[[3534  369]
 [ 106 3873]]
```

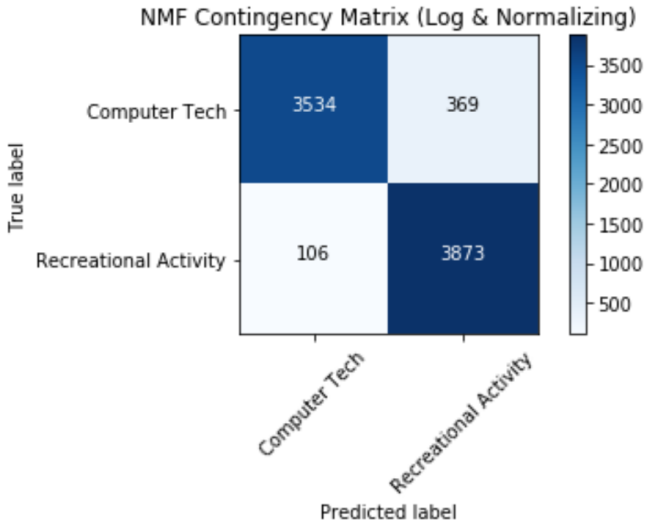

Figure 35: Contingency Matrix with Log Transformation and Normalization

2.5 Expand Dataset into 20 categories

The dimension of the TF-IDF matrix is (18846, 52295).

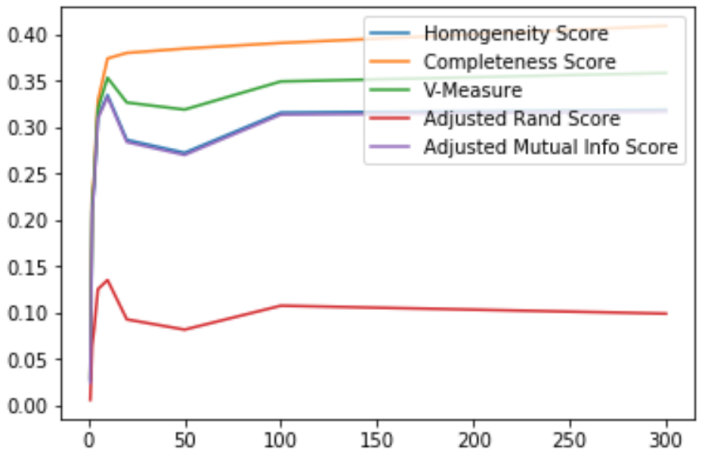

Figure 36: LSI Measures V.S. Dimensions

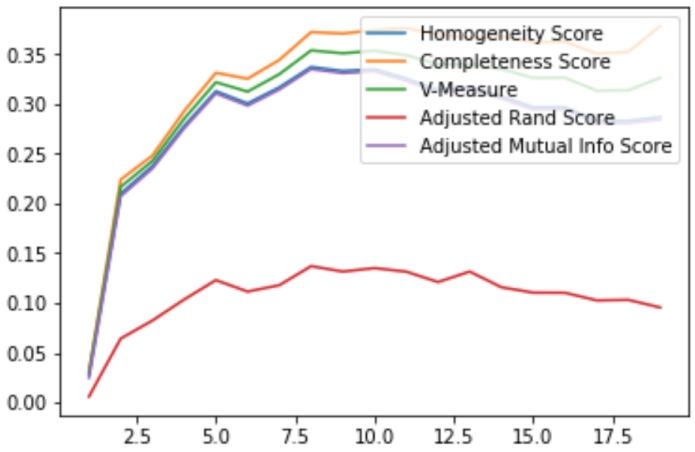

Figure 37: LSI Measures V.S. Dimensions (only from 1 to 20)

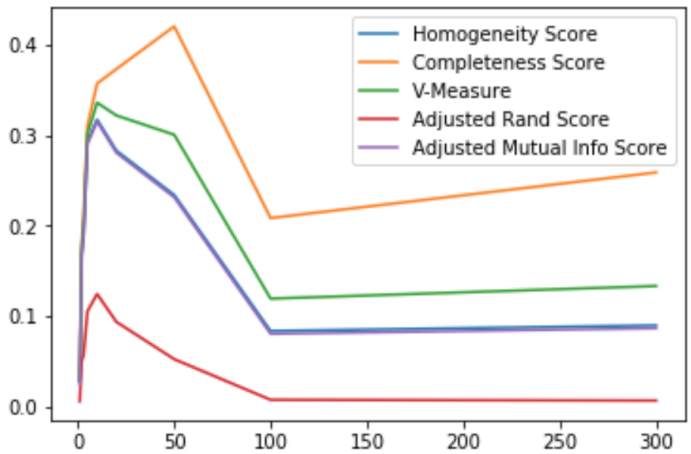

Figure 38: NMF Measures V.S. Dimensions

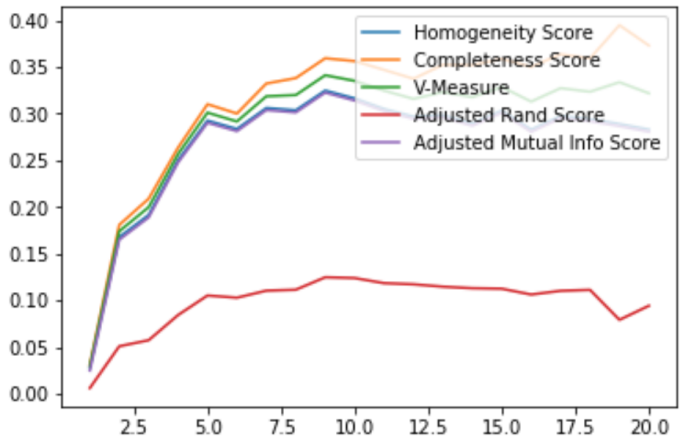

Figure 39: NMF Measures V.S. Dimensions (only from 1 to 20)

The best dimension for LSI is 10 and the best dimension for NMF is 9.

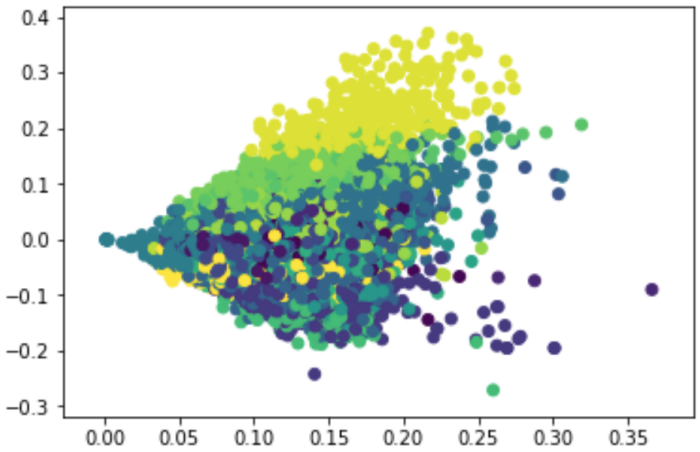

Figure 40: LSI Best Clustering Results Visualization

– Homogeneity: 0.334
– Completeness: 0.374
– V-measure: 0.353
– Adjusted Rand-Index: 0.134
– Adjusted Mutual Info Score: 0.332

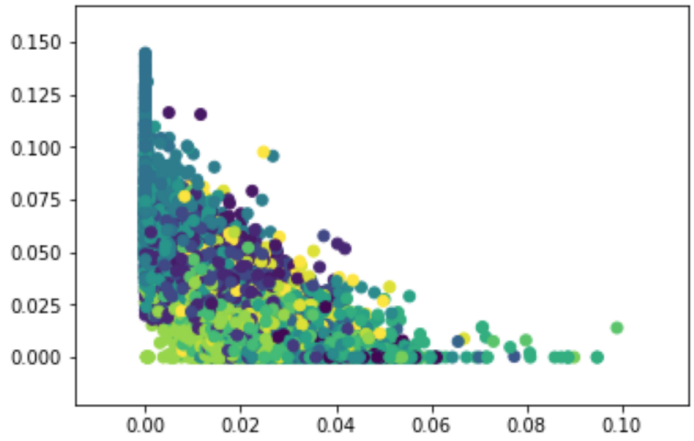

Figure 41: NMF Best Clustering Results Visualization

– Homogeneity: 0.321
– Completeness: 0.357
– V-measure: 0.338
– Adjusted Rand-Index: 0.125
– Adjusted Mutual Info Score: 0.318

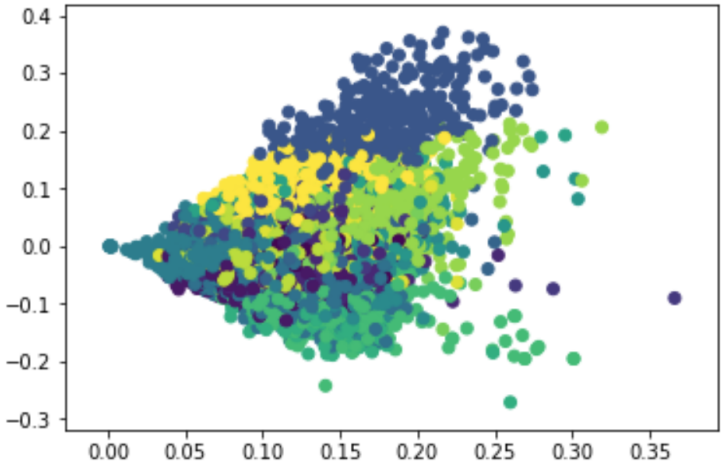

Figure 42: LSI Clustering Results Visualization with Normalizing

– Homogeneity: 0.315
– Completeness: 0.362
– V-measure: 0.337
– Adjusted Rand-Index: 0.123
– Adjusted Mutual Info Score: 0.312

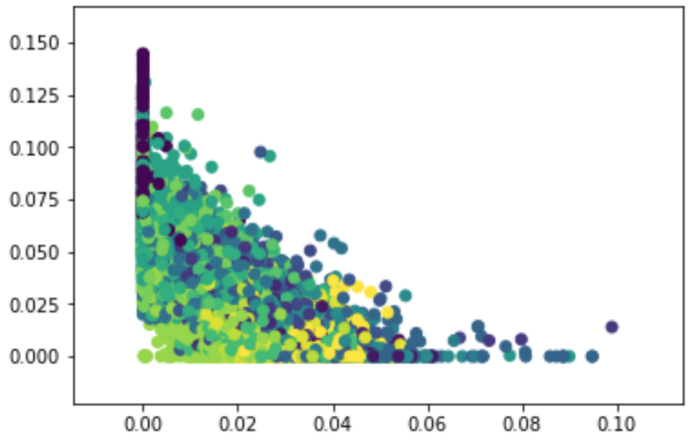

Figure 43: NMF Clustering Results Visualization with Normalizing

– Homogeneity: 0.308
– Completeness: 0.337
– V-measure: 0.322
– Adjusted Rand-Index: 0.119
– Adjusted Mutual Info Score: 0.306

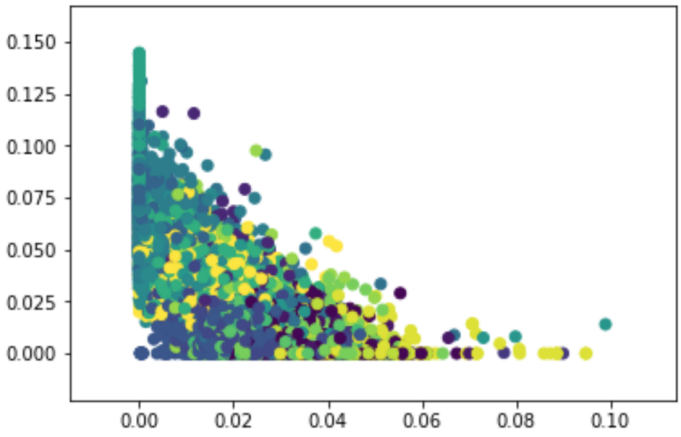

Figure 44: NMF Clustering Results Visualization with Log Transformation

– Homogeneity: 0.327
– Completeness: 0.360
– V-measure: 0.343
– Adjusted Rand-Index: 0.128
– Adjusted Mutual Info Score: 0.325

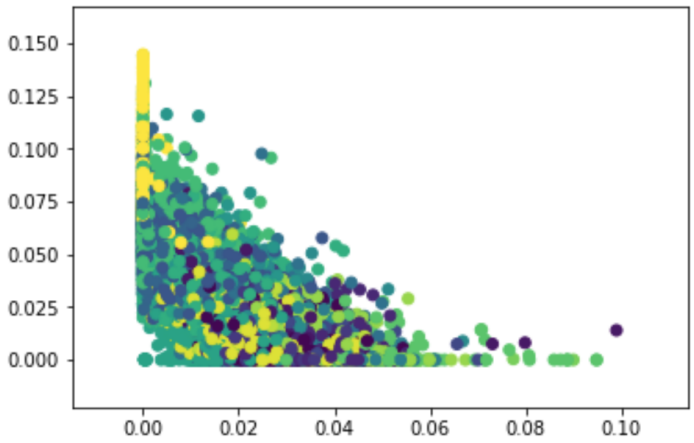

Figure 45: NMF Clustering Results Visualization with Normalizing and Log Transformation

– Homogeneity: 0.309
– Completeness: 0.337
– V-measure: 0.323
– Adjusted Rand-Index: 0.120
– Adjusted Mutual Info Score: 0.307

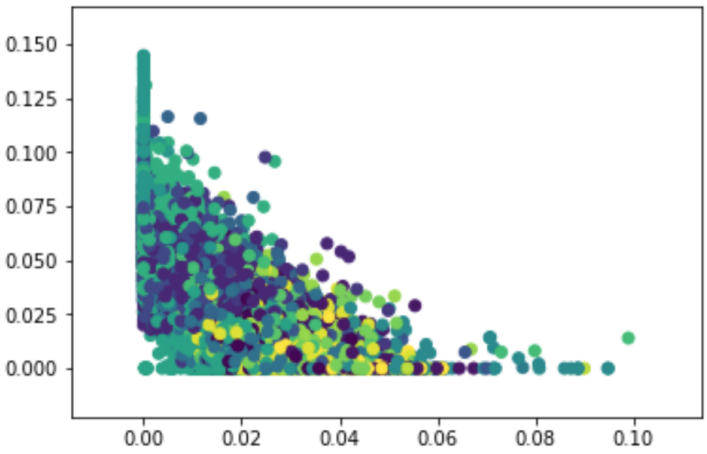

Figure 46: NMF Clustering Results Visualization with Log Transformation and Normalizing

– Homogeneity: 0.307
– Completeness: 0.340
– V-measure: 0.323
– Adjusted Rand-Index: 0.120
– Adjusted Mutual Info Score: 0.305