ECE 219 Project2

Jiuru Shao
UID:204288539
Haoxiang Zhang
UID:104278461

Feb 12th 2018

1 Introduction

In this experiment, we implemented a simple text classifier with unsupervised method(K-means clustering). Clustering differs from classification in that no *a priori* labeling (grouping) of the data points is available.

We represent of text file as TF-IDF matrix, and perform SVD and NMF on the matrix to reduce dimension. We evaluates our result with five metrics, namely the homogeneity score, the completeness score, the V-measure, the adjusted Rand score and the adjusted mutual info score.

We tried our code on both binary classification and multi-label classification. We find that the binary clustering works fine, but the 20-class clustering result are below our expectation despite all the methods we tried.

2 Problem 1. Building the TF-IDF Matrix

We use the same method as project 1 to perform TF-IDF, and the dimension of the TF-IDF matrix we get is 7882 × 27768.

3 Problem 2. Apply K-means Clustering using the TF-IDF

The results we get are:

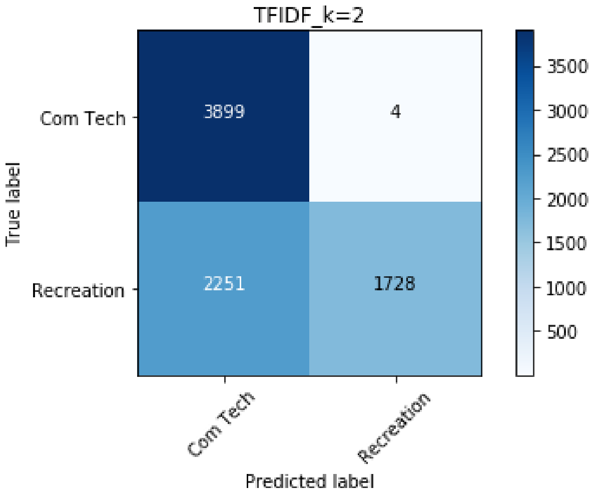


Figure 1: **Contingency Matrix of TFIDF Data**

Homogeneity	0.255
Completeness	0.336
V-measure	0.290
Adjusted rand score	0.183
Adjusted mutual info score	0.255

Table 1: **5 Measures of TFIDF**

As we can see from figure 1 and table 1, the original TFIDF data provides a result that is not that bad. However, there are many cluster 2 data points are predicted as cluster 1 points, which is the reason that 5 measure scores are not very high.

4 Problem 3. Dimensionality Reduction

4.1 3a Report the percent of variance

In this part, we use the explained_variance_ratio in the TruncateSVD class to plot the graph. The result is in Figure 1.

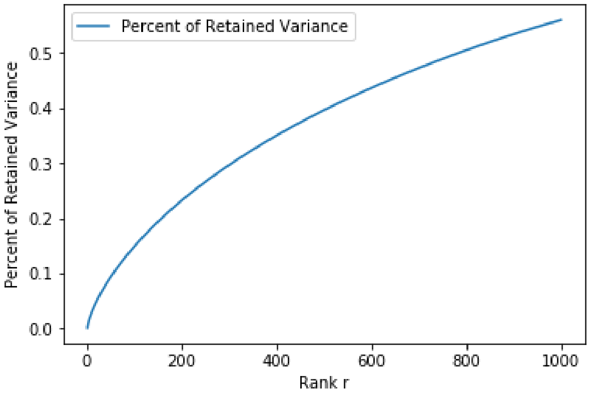
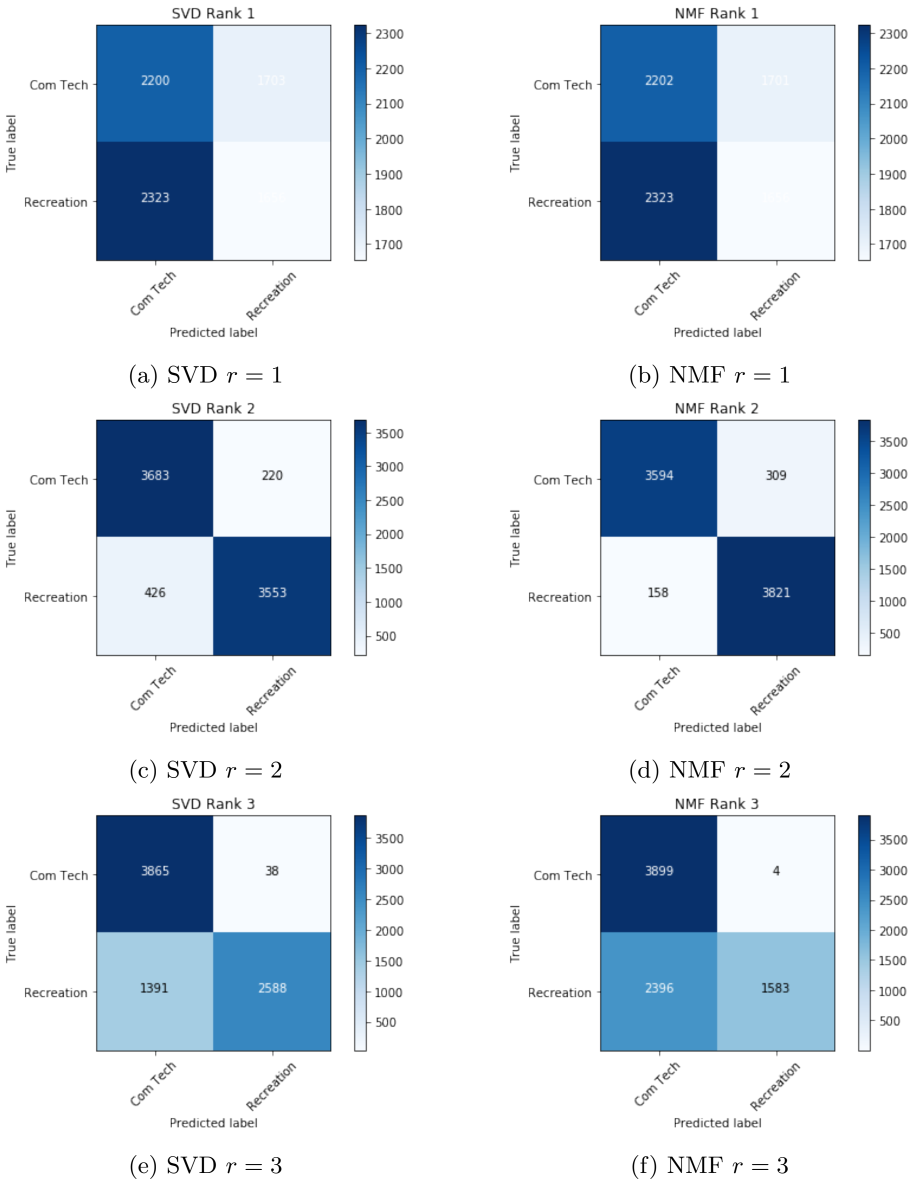


Figure 2: **The plot of the percent of variance the top r principle components can retain v.s. r, for r = 1 to 1000.**

As r keeps increasing, percent of retained variance increases as well.

4.2 3b Apply SVD and NMF



(a) SVD $r = 1$

(b) NMF $r = 1$

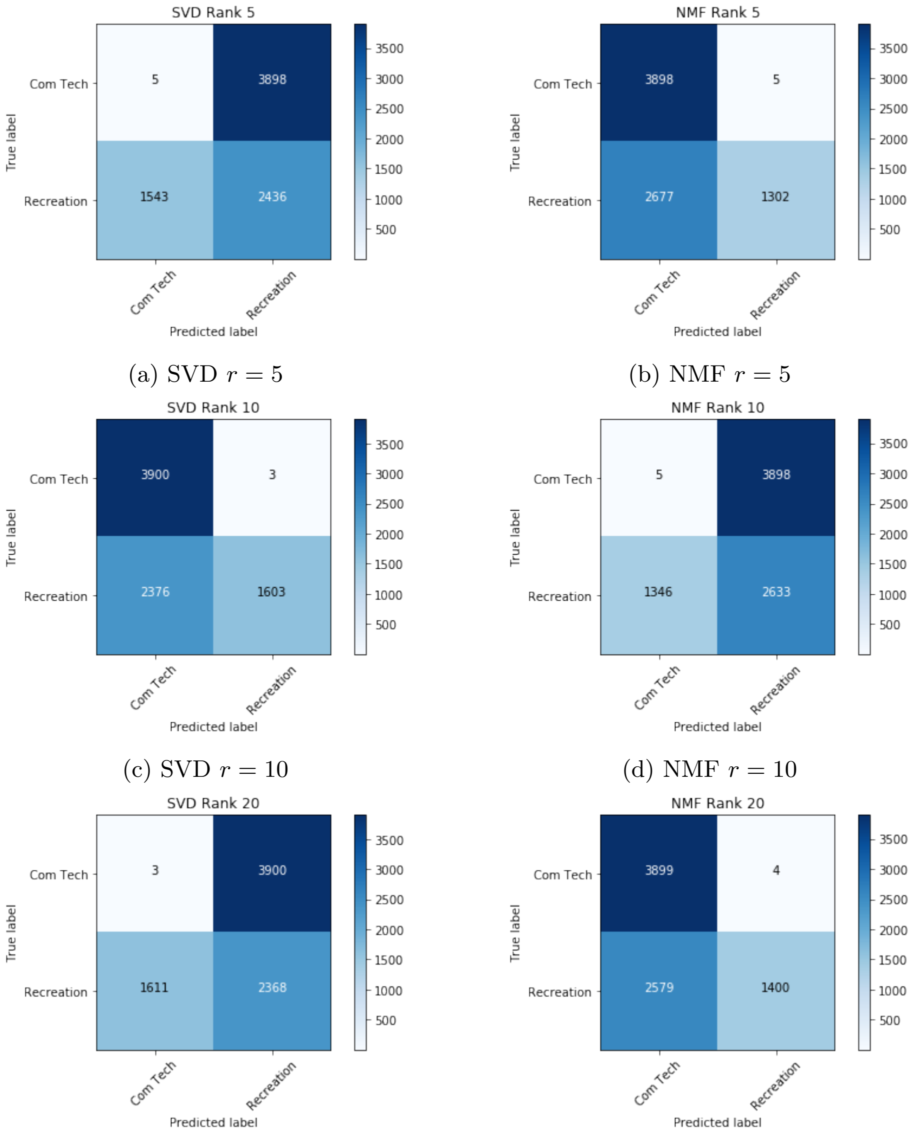
(c) SVD $r = 2$

(d) NMF $r = 2$

(e) SVD $r = 3$

(f) NMF $r = 3$

Figure 3: **Contingency Matrix for** $r = 1, 2, 3$



(a) SVD $r = 5$

(b) NMF $r = 5$

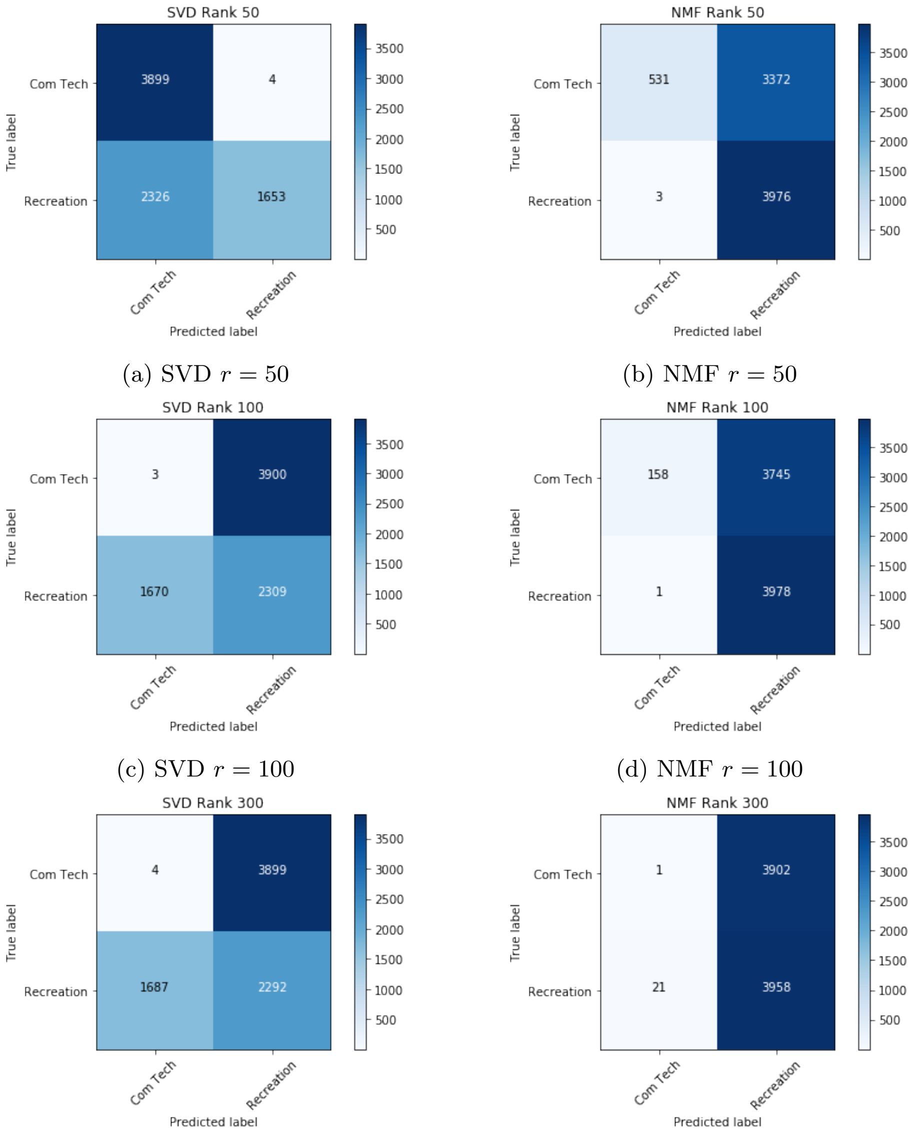
(c) SVD $r = 10$

(d) NMF $r = 10$

(e) SVD $r = 20$

(f) NMF $r = 20$

Figure 4: **Contingency Matrix for** $r = 5, 10, 20$



(a) SVD $r = 50$

(b) NMF $r = 50$

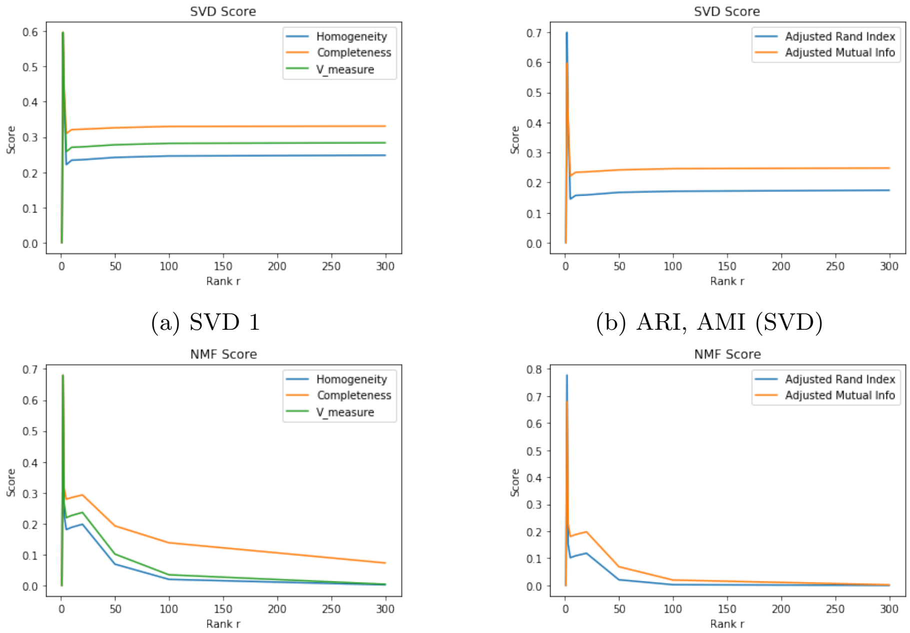
(c) SVD $r = 100$

(d) NMF $r = 100$

(e) SVD $r = 300$

(f) NMF $r = 300$

Figure 5: **Contingency Matrix for** $r = 50, 100, 300$



(a) SVD 1

(b) ARI, AMI (SVD)

(c) NMF 1

(d) ARI, AMI (NMF)

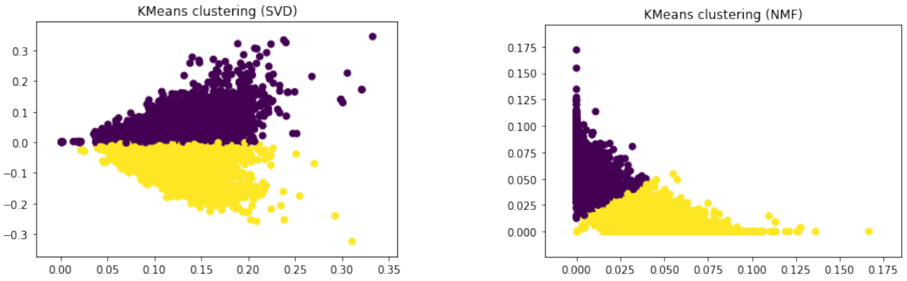
Figure 6: **5 Measures**

From plots of contingency matrix and plots of 5 measures v.s. r, we conclude that the best r choice for both SVD and NMF is $r = 2$.

We observe the non-monotonic behavior of the measure and one possible explanation is that Euclidean distance is not a good distance in high dimensions. In high dimensions, most of the mass of a multivariate Gaussian distribution is not near the mean, but in an increasingly distant "shell" around it; and most of the volume of a high-dimensional orange is in the skin, not the pulp.

5 Problem 4

5.1 4a Visualize Best Clustering Results



(a) SVD

(b) NMF

Figure 7: **Best Clustering Results**

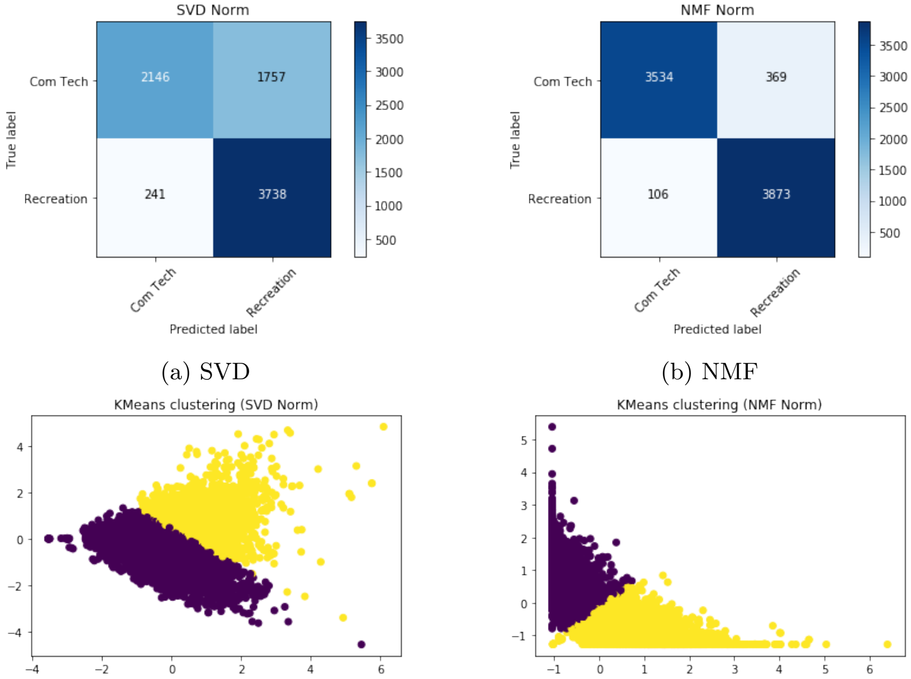
	SVD	NMF
Homogeneity	0.596	0.679
Completeness	0.597	0.680
V-measure	0.596	0.680
Adjusted rand score	0.699	0.777
Adjusted mutual info score	0.596	0.679

Table 2: **5 Measure of Best Clustering Results**

From the distribution of the data points after SVD, we can observe that it is very hard to cluster them into 2 classes, because most of the points are very close to each other and there is no clear boundary. In contrast, NMF is slightly better, showing a triangular distribution. Table 2 confirms this observation.

5.2 4b Different Transformations

1. **Normalization**



(a) SVD

(b) NMF

(c) SVD

(d) NMF

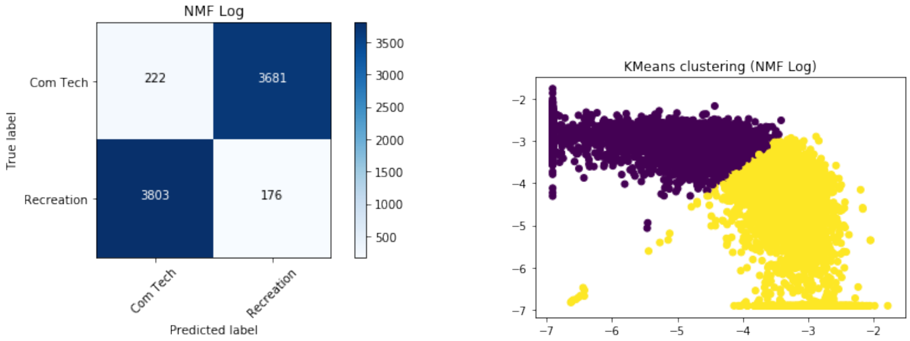
Figure 8: **Plots after Normalization**

	SVD	NMF
Homogeneity	0.235	0.683
Completeness	0.264	0.686
V-measure	0.249	0.684
Adjusted rand score	0.255	0.773
Adjusted mutual info score	0.235	0.683

Table 3: **5 Measures after Normalization**

We can see the result of normalization after SVD decomposition becomes significantly worse, while the result after NMF remains the same, we can confirm this by visualizing the data points: the distribution of the observation points still remain the similar shape.

2. **Logarithm Transformation (NMF Only)**



(a) Contingency Matrix

(b) Visualize

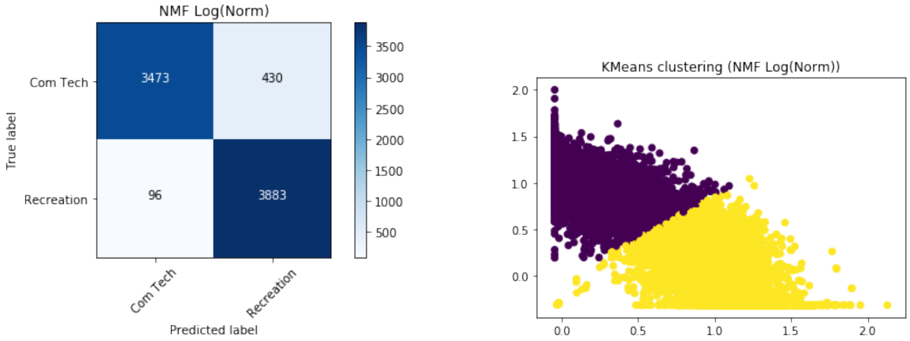
Figure 9: **Plots after Log**

Homogeneity	0.712
Completeness	0.712
V-measure	0.712
Adjusted rand score	0.808
Adjusted mutual info score	0.712

Table 4: **Measures after Log**

We find log transformation greatly improves the result because the original data points are skewed with wide distribution, and taking the log of the features may restore symmetry to it. As we can see from the distribution of the observations after taking log, the points are mapped to a sector that is easier to cluster.

3. **Norm then Log (NMF Only)**



(a) Contingency Matrix

(b) Visualize

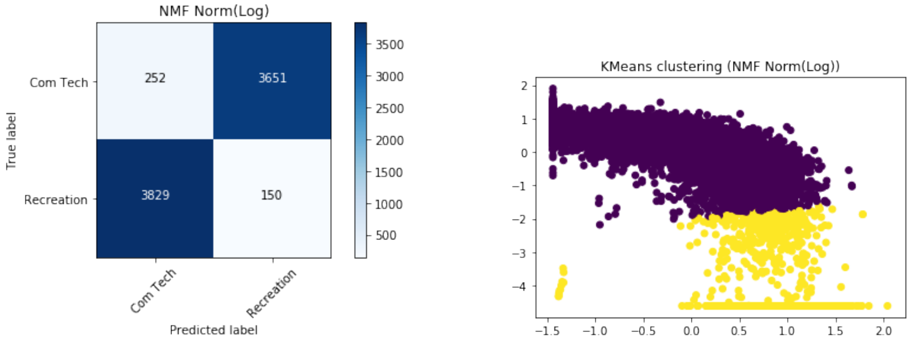
Figure 10: **Plots (Norm then Log)**

Homogeneity	0.663
Completeness	0.667
V-measure	0.665
Adjusted rand score	0.751
Adjusted mutual info score	0.663

Table 5: **5 Measures (Norm then Log)**

The result is roughly the same as part 4a, there is no improvement.

4. **Log then Norm (NMF Only)**



(a) Contingency Matrix

(b) Visualize

Figure 11: **Plots (Log then Norm)**

Homogeneity	0.711
Completeness	0.712
V-measure	0.712
Adjusted rand score	0.806
Adjusted mutual info score	0.711

Table 6: **5 Measures (Log then Norm)**

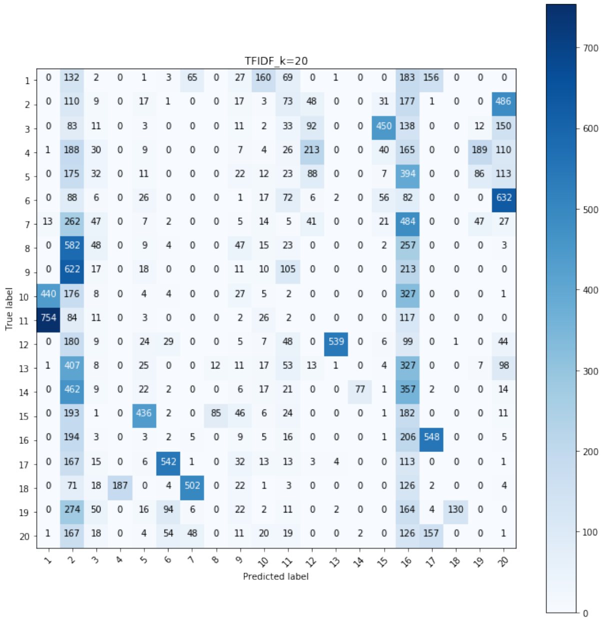
The result is similar to taking log only.

6 Problem 5

6.1 Get TF-IDF matrix

The dimension of the TF-IDF matrix is 18846 × 52295.

6.2 Apply K-means clustering with k = 20



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	132	2	0	1	3	65	0	27	160	69	0	1	0	0	183	156	0	0	0
2	0	110	9	0	17	1	0	0	17	3	73	48	0	0	31	177	1	0	0	486
3	0	83	11	0	3	0	0	0	11	2	33	92	0	0	450	138	0	0	12	150
4	1	188	30	0	9	0	0	0	7	4	26	213	0	0	40	165	0	0	189	110
5	0	175	32	0	11	0	0	0	22	12	23	88	0	0	7	394	0	0	86	113
6	0	88	6	0	26	0	0	0	1	17	72	6	2	0	56	82	0	0	0	632
7	13	262	47	0	7	2	0	0	5	14	5	41	0	0	21	484	0	0	47	27
8	0	582	48	0	9	4	0	0	47	15	23	0	0	0	2	257	0	0	0	3
9	0	622	17	0	18	0	0	0	11	10	105	0	0	0	0	213	0	0	0	0
10	440	176	8	0	4	4	0	0	27	5	2	0	0	0	0	327	0	0	0	1
11	754	84	11	0	3	0	0	0	2	26	2	0	0	0	0	117	0	0	0	0
12	0	180	9	0	24	29	0	0	5	7	48	0	539	0	6	99	0	1	0	44
13	1	407	8	0	25	0	0	12	11	17	53	13	1	0	4	327	0	0	7	98
14	0	462	9	0	22	2	0	0	6	17	21	0	0	77	1	357	2	0	0	14
15	0	193	1	0	436	2	0	85	46	6	24	0	0	0	1	182	0	0	0	11
16	0	194	3	0	3	2	5	0	9	5	16	0	0	0	1	206	548	0	0	5
17	0	167	15	0	6	542	1	0	32	13	13	3	4	0	0	113	0	0	0	1
18	0	71	18	187	0	4	502	0	22	1	3	0	0	0	0	126	2	0	0	4
19	0	274	50	0	16	94	6	0	22	2	11	0	2	0	0	164	4	130	0	0
20	1	167	18	0	4	54	48	0	11	20	19	0	0	2	0	126	157	0	0	1

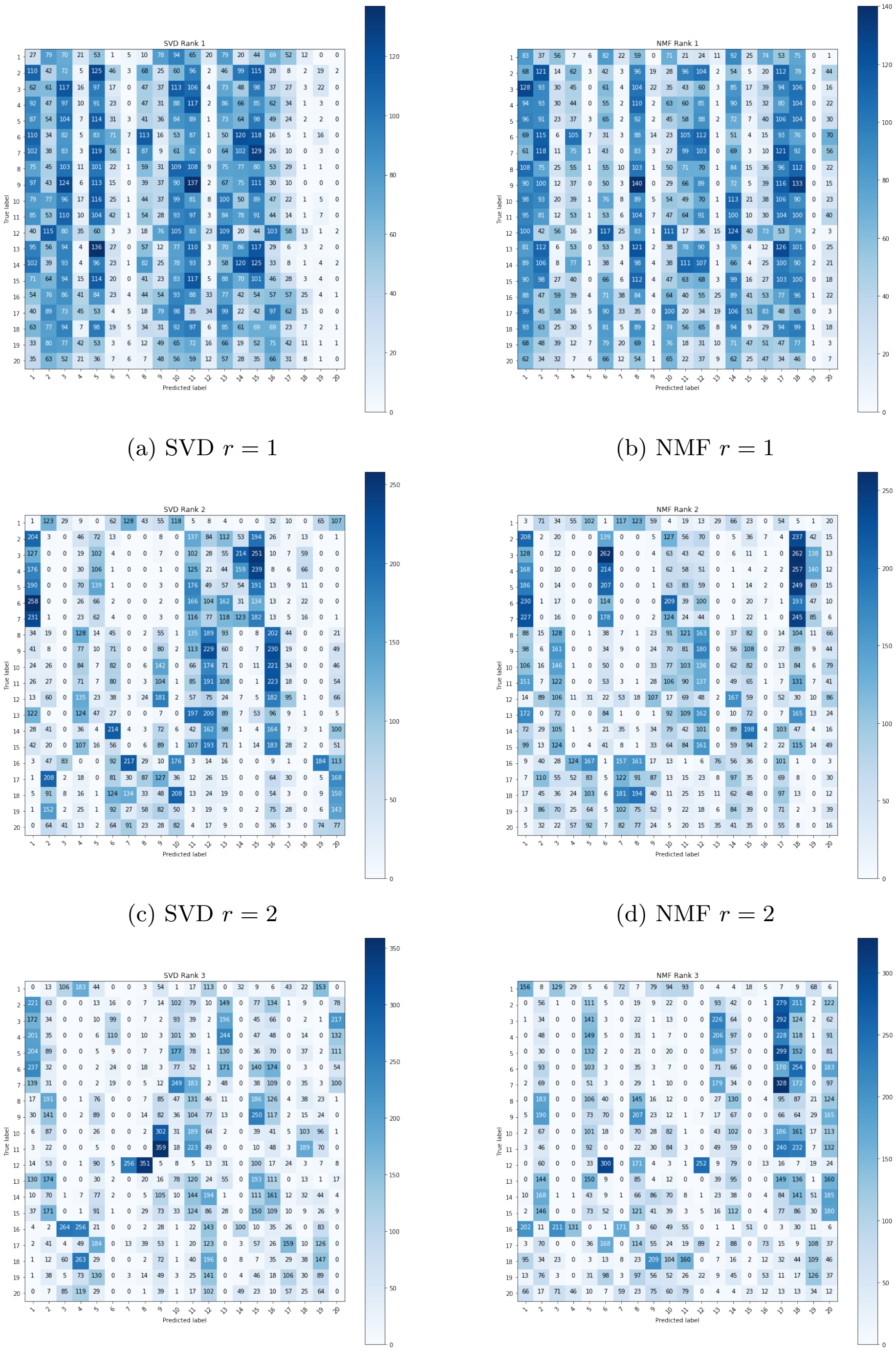
Figure 12: **Contingency Matrix of TFIDF Data**

Homogeneity	0.322
Completeness	0.398
V-measure	0.356
Adjusted rand score	0.119
Adjusted mutual info score	0.320

Table 7: **5 Measures of TFIDF**

As we can see from figure 12 and table 7, the original TFIDF data provides a result that is not bad because this is a clustering for 20 different clusters.

6.3 Dimension reduction with SVD and NMF



(a) SVD $r = 1$

(b) NMF $r = 1$

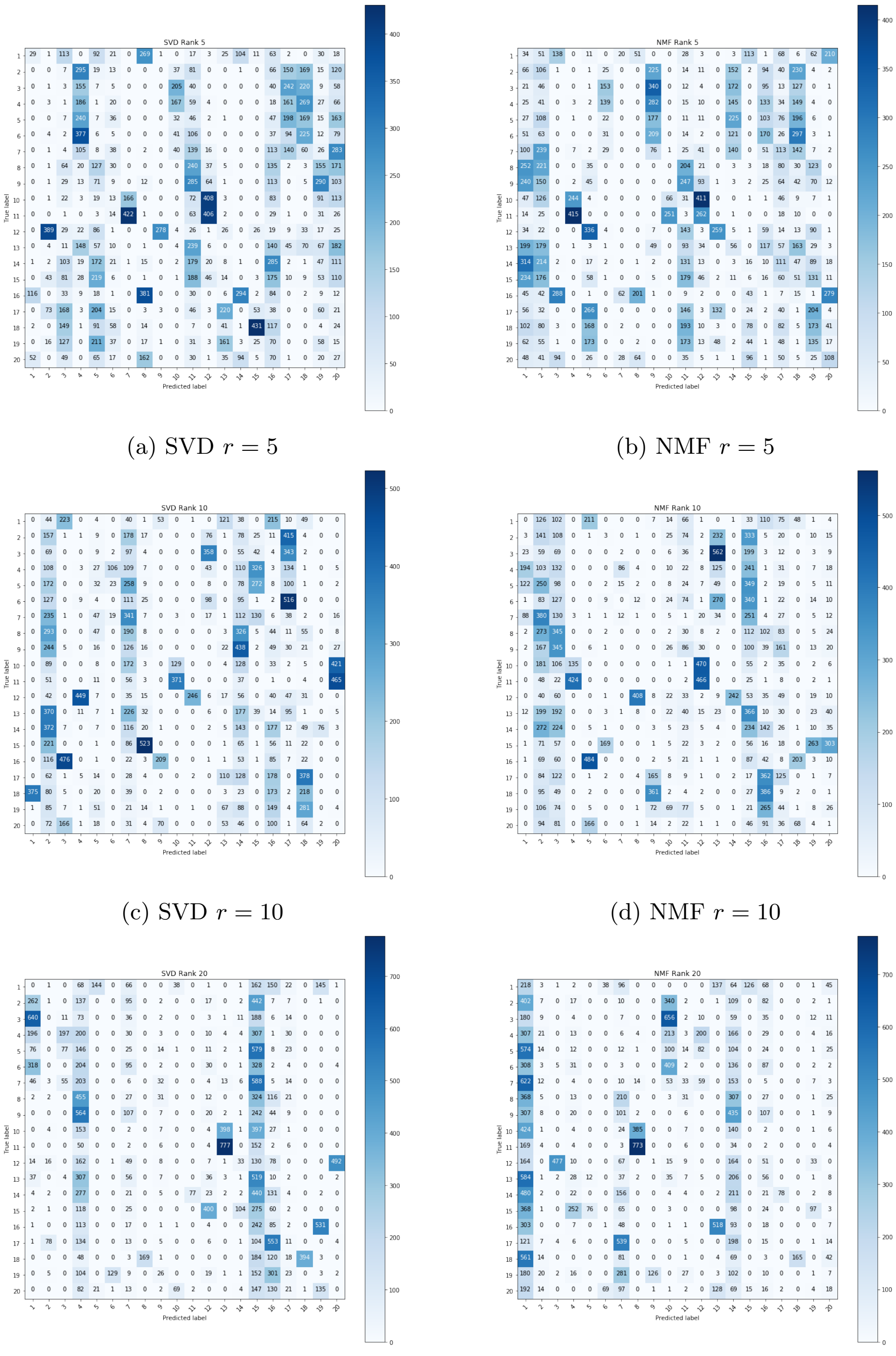
(c) SVD $r = 2$

(d) NMF $r = 2$

(e) SVD $r = 3$

(f) NMF $r = 3$

Figure 13: **Contingency Matrix for** $r = 1, 2, 3$



(a) SVD $r = 5$

(b) NMF $r = 5$

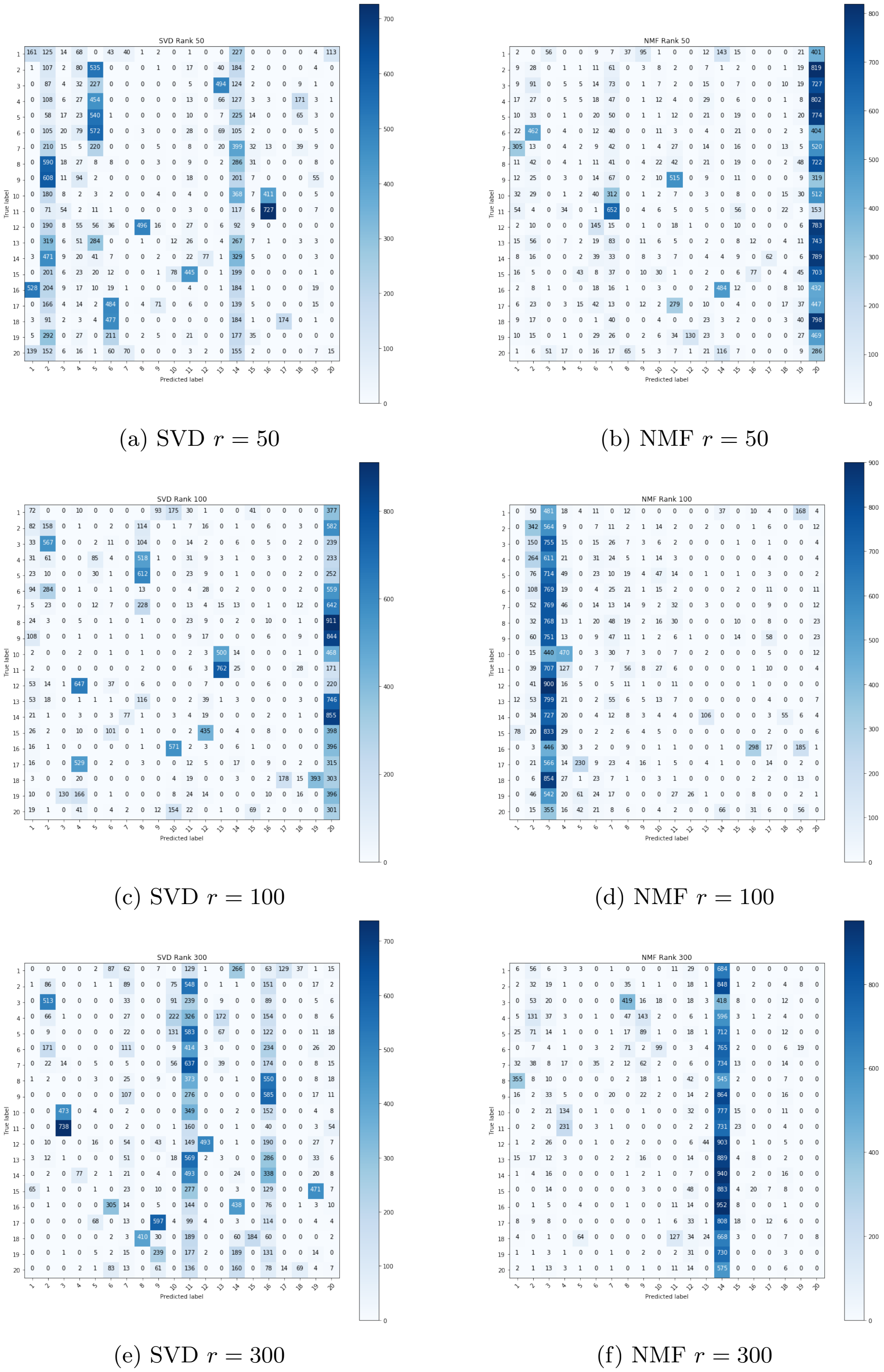
(c) SVD $r = 10$

(d) NMF $r = 10$

(e) SVD $r = 20$

(f) NMF $r = 20$

Figure 14: **Contingency Matrix for** $r = 5, 10, 20$



(a) SVD $r = 50$

(b) NMF $r = 50$

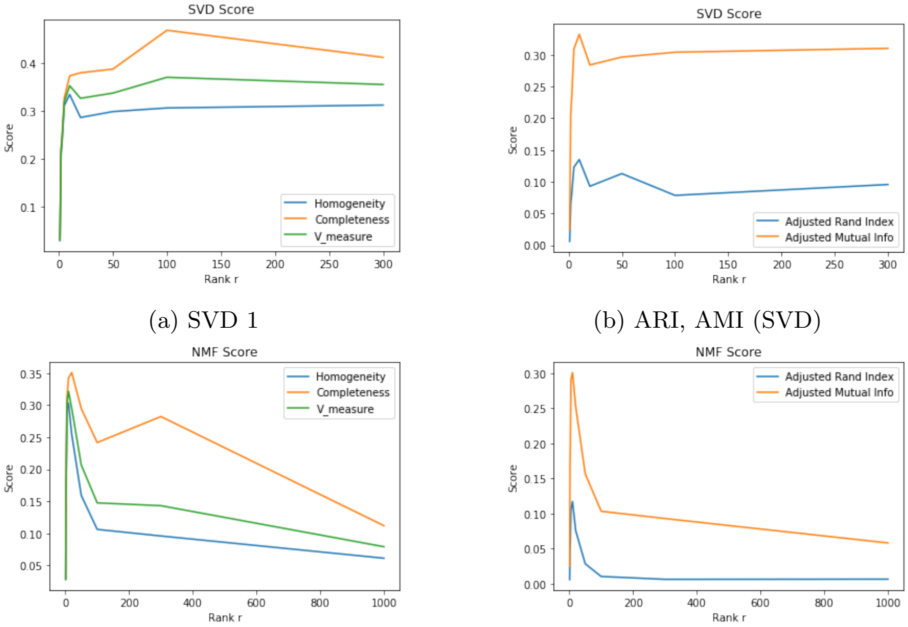
(c) SVD $r = 100$

(d) NMF $r = 100$

(e) SVD $r = 300$

(f) NMF $r = 300$

Figure 15: **Contingency Matrix for** $r = 50, 100, 300$



(a) SVD 1

(b) ARI, AMI (SVD)

(c) NMF 1

(d) ARI, AMI (NMF)

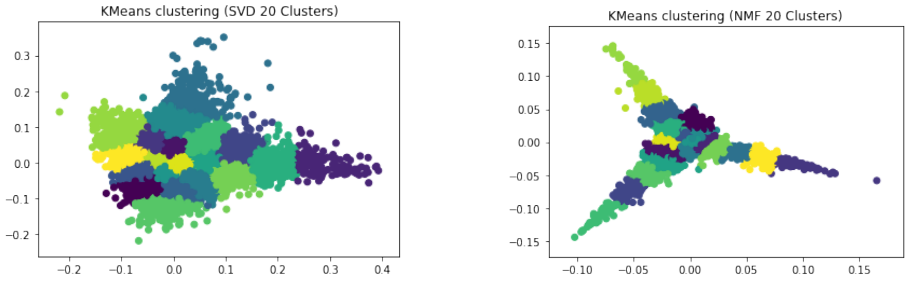
Figure 16: **5 Measures of 20 Clusters**

As in problem 3, we explored the performance measures for $r = 1, 2, 3, 5, 10, 20, 50, 100, 300$. For NMF, we expaned the performance measures to $r = 1000$ as well to observe the trend. However we find that unlike the previous part, the optimal r for each metric are not the same at each time. We pick adjusted mutual info score as the metric, and the result shows $r = 10$ is the best value for both SVD and NMF. The details of the measures at $r = 10$ are:

	SVD	NMF
Homogeneity	0.335	0.317
Completeness	0.374	0.357
V-measure	0.353	0.336
Adjusted rand score	0.135	0.124
Adjusted mutual info score	0.332	0.315

Table 8: **Measures of Best Clustering Results**

We visualize the best clustering results of 20 clusters:



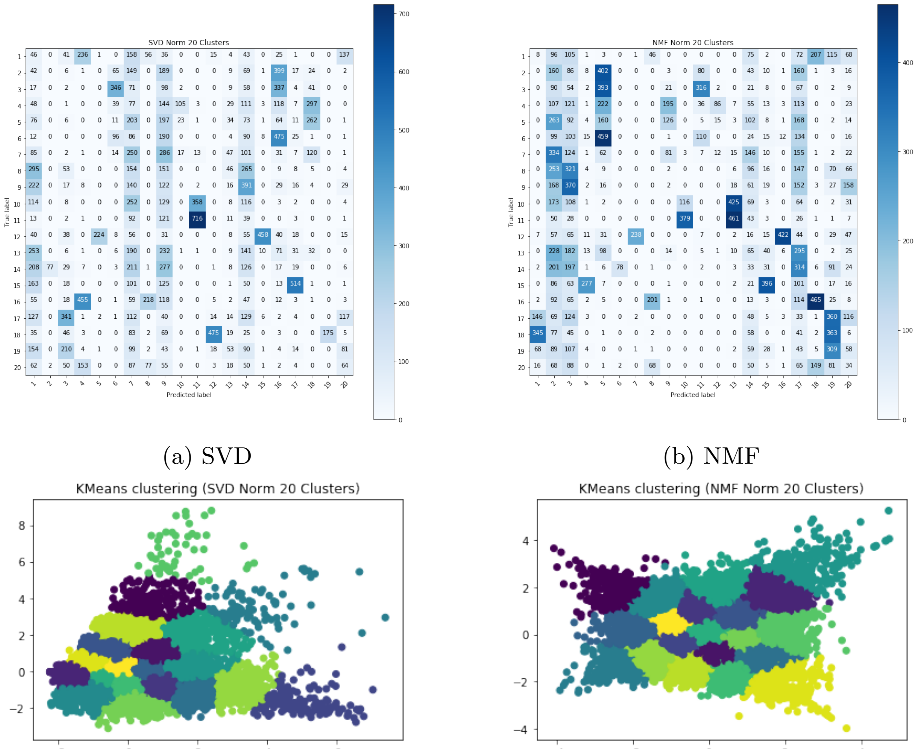
(a) SVD

(b) NMF

Figure 17: **Best Clustering Results**

6.4 Apply Normalization and Log

1. Normalization



(a) SVD

(b) NMF

(c) SVD

(d) NMF

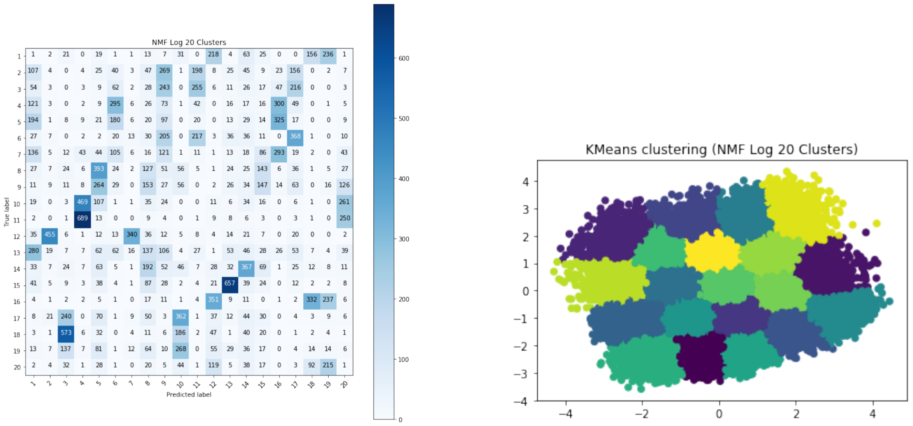
Figure 18: **Plots after Normalization**

	SVD	NMF
Homogeneity	0.310	0.305
Completeness	0.349	0.341
V-measure	0.328	0.322
Adjusted rand score	0.127	0.117
Adjusted mutual info score	0.308	0.303

Table 9: **5 Measures after Normalization**

Despite the changes in the pattern(the data points become more "spread-out") of distribution after normalization, the result is still not better than before.

2. **Log Transformation with NMF**



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	2	21	0	19	1	1	13	7	31	0	218	4	63	25	0	0	156	236	1
2	107	4	0	4	25	40	3	47	269	1	198	8	25	45	9	23	156	0	2	7
3	54	3	0	3	9	62	2	28	243	0	255	6	11	26	17	47	216	0	0	3
4	121	3	0	2	9	295	6	26	73	1	42	0	16	17	16	300	49	0	1	5
5	194	1	8	9	21	180	6	20	97	0	20	0	13	29	14	325	17	0	0	9
6	27	7	0	2	2	20	13	30	205	0	217	3	36	36	11	0	368	1	0	10
7	136	5	12	43	44	105	6	16	121	1	11	1	13	18	86	293	19	2	0	43
8	27	7	24	6	393	24	2	127	51	56	5	1	24	25	143	6	36	1	5	27
9	11	9	11	8	264	29	0	153	27	56	0	2	26	34	147	14	63	0	16	126
10	19	0	3	469	107	1	1	35	24	0	0	11	6	34	16	0	6	1	0	261
11	2	0	1	689	13	0	0	9	4	0	1	9	8	6	3	0	3	1	0	250
12	35	455	6	1	12	13	340	36	12	5	8	4	14	21	7	0	20	0	0	2
13	280	19	7	7	62	62	16	137	106	4	27	1	53	46	28	26	53	7	4	39
14	33	7	24	7	63	5	1	192	52	46	7	28	32	367	69	1	25	12	8	11
15	41	5	9	3	38	4	1	87	28	2	4	21	657	39	24	0	12	2	2	8
16	4	1	2	2	5	1	0	17	11	1	4	351	9	11	0	1	2	332	237	6
17	8	21	240	0	70	1	9	50	3	362	1	37	12	44	30	0	4	3	9	6
18	3	1	573	6	32	0	4	11	6	186	2	47	1	40	20	0	1	2	4	1
19	13	7	137	7	81	1	12	64	10	268	0	55	29	36	17	0	4	14	14	6
20	2	4	32	1	28	1	0	20	5	44	1	119	5	38	17	0	3	92	215	1

(a) Contingency Matrix

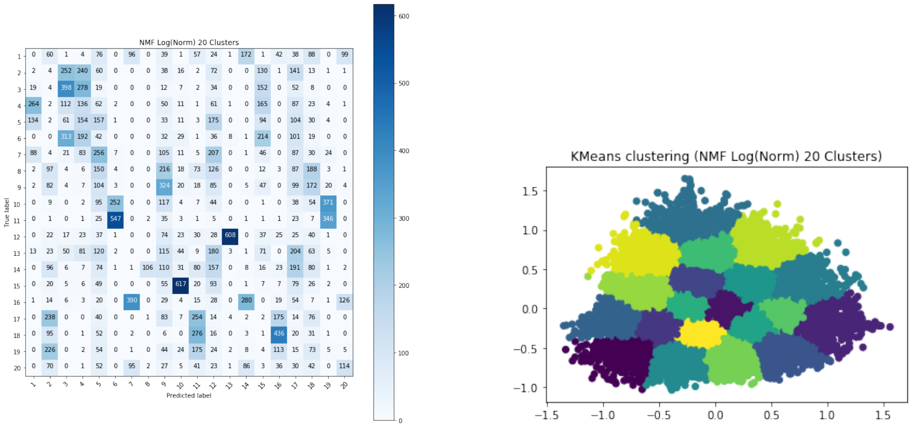
(b) Visualize

Figure 19: **Plots after Log**

Homogeneity	0.376
Completeness	0.380
V-measure	0.378
Adjusted rand score	0.205
Adjusted mutual info score	0.374

Table 10: **Measures after Log**

3. Norm then Log with NMF



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0	60	1	4	76	0	96	0	39	1	57	24	1	172	1	42	38	88	0	99
2	2	4	252	240	60	0	0	0	38	16	2	72	0	0	130	1	141	13	1	1
3	19	4	398	278	19	0	0	0	12	7	2	34	0	0	152	0	52	8	0	0
4	264	2	112	136	62	2	0	0	50	11	1	61	1	0	165	0	87	23	4	1
5	134	2	61	154	157	1	0	0	33	11	3	175	0	0	94	0	104	30	4	0
6	0	0	313	192	42	0	0	0	32	29	1	36	8	1	214	0	101	19	0	0
7	88	4	21	83	256	7	0	0	105	11	5	207	0	1	46	0	87	30	24	0
8	2	97	4	6	150	4	0	0	216	18	73	126	0	0	12	3	87	188	3	1
9	2	82	4	7	104	3	0	0	324	20	18	85	0	5	47	0	99	172	20	4
10	0	9	0	2	95	252	0	0	117	4	7	44	0	0	1	0	38	54	371	0
11	0	1	0	1	25	547	0	2	35	3	1	5	0	1	1	1	23	7	346	0
12	0	22	17	23	37	1	0	0	74	23	30	28	608	0	37	25	25	40	1	0
13	13	23	50	81	120	2	0	0	115	44	9	180	3	1	71	0	204	63	5	0
14	0	96	6	7	74	1	1	106	110	31	80	157	0	8	16	23	191	80	1	2
15	0	20	5	6	49	0	0	0	55	617	20	93	0	1	7	7	79	26	2	0
16	1	14	6	3	20	0	390	0	29	4	15	28	0	280	0	19	54	7	1	126
17	0	238	0	0	40	0	0	1	83	7	254	14	4	2	2	175	14	76	0	0
18	0	95	0	1	52	0	2	0	6	0	276	16	0	3	1	436	20	31	1	0
19	0	226	0	2	54	0	1	0	44	24	175	24	2	8	4	113	15	73	5	5
20	0	70	0	1	52	0	95	2	27	5	41	23	1	86	3	36	30	42	0	114

(a) Contingency Matrix

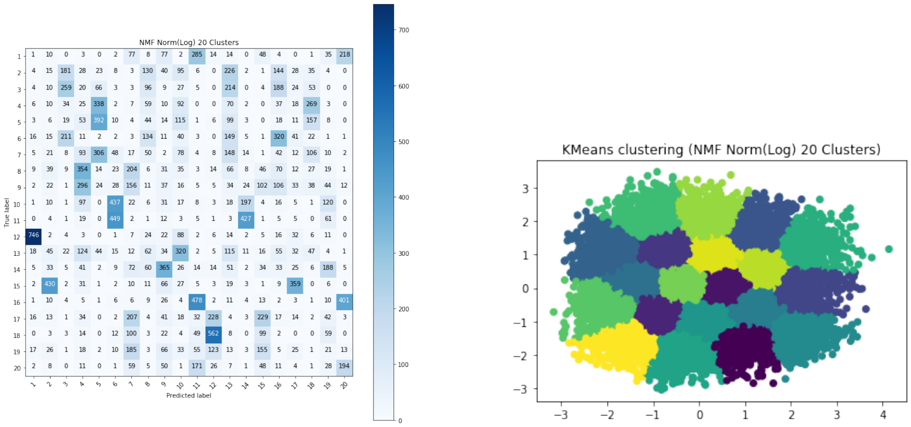
(b) Visualize

Figure 20: **Plots (Norm then Log)**

Homogeneity	0.337
Completeness	0.348
V-measure	0.342
Adjusted rand score	0.157
Adjusted mutual info score	0.335

Table 11: **5 Measures (Norm then Log)**

4. **Log then Norm with NMF**



(a) Contingency Matrix

(b) Visualize

Figure 21: **Plots (Log then Norm)**

Homogeneity	0.368
Completeness	0.370
V-measure	0.369
Adjusted rand score	0.202
Adjusted mutual info score	0.366

Table 12: **5 Measures (Log then Norm)**

From our experiment, we can see that the best result is generated with the log transformation only. We also notice that even the best result only have a minor improvement on the metrics despite the transformation in the distribution of data points.

7 Conclusion

In summary, we achieve the goal of this project. We used TF-IDF and SVD/NMF to effectively represent the features and reduce the dimensions. In the case of binary classification, our result shows that NMF is more effective than SVD if we perform a log transformation after dimension reduction. Also, a log transformation followed by normalization will yield similar result.

On the other hand, we also tried to expand our dataset by clustering the 20 original data sets. However, the result is much worse than the binary case. Similarly, the best result is achieved with log transformation with NMF. This indicates that our method is not able to cluster the texts into more refined classes effectively.