EE219 Project 2 - Report
Clustering
Winter 2018

Nrithya Theetharappan
nrithya@ucla.edu
004946349

Shraddha Manchekar
smanchekar@ucla.edu
004945217

Introduction

The project aims to explore Clustering algorithms, which provide unsupervised learning methods for grouping of similar data points together, when no *a priori* labelling is available.

K-means clustering tries to divide the datapoints into K clusters with each data point belonging to one and only one cluster. The algorithm works as follows:

1. Assign data points to clusters(initially random) such that the distance between each point and the center is minimized
2. Reassign data points depending on the distance from the center at each turn
3. Calculate centers based on the mean of distances of the data points

In this project, our goal is to find (a) proper representation of the data such that the clustering is efficient, (b) to perform K-means clustering on the dataset and evaluate its performance and (c) to try different preprocessing methods to improve the clustering performance.

Dataset

We use the "20 newsgroups" dataset. It is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups, each corresponding to a different topic. We pretend that the class labels are unknown to us and aim to group the data. We thus work with a well differentiated portion of the data, namely the two major classes: 'Computer Technology' and 'Recreational activity'.

Part 1: Building the TF-IDF Matrix

In this part, we find a good representation of the data to perform K-means clustering on the data in further steps. Following the steps in Project 1, in this step, we compute the TF-IDF representation of the data, in order to make the text data into analyzable and numeric format by setting min_df=3 and excluding stopwords.

The dimensions of our TF-IDF matrix are (7882, 18469)

Part 2: Applying K-means with k=2 to the TF-IDF data

In this part, we applied K-means clustering algorithm with k=2. The following measures are computed and inspected to get a sense of the effectiveness of the clustering algorithm against the known class labels.

The contingency matrix gives the number of data points that are members of a both a given class and the corresponding cluster.

Various purity measures used here for a more complete analysis are: Homogeneity score, Completeness score, V-measure, Adjusted Rand Index and Adjusted Mutual Info Score

The following table shows the values of purity measures that we obtained.

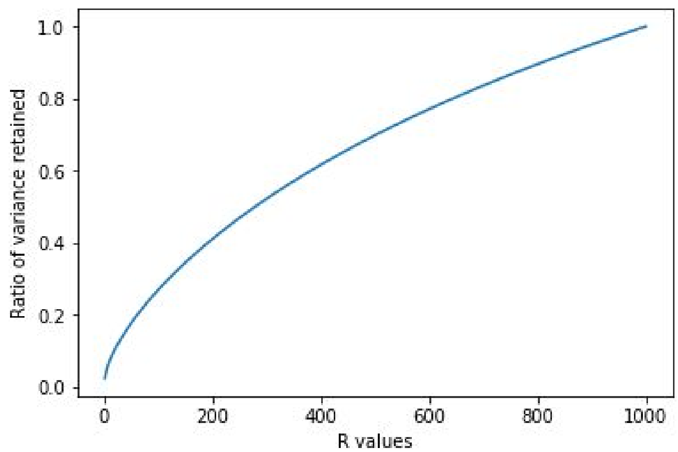
Contingency Matrix	**Homogeneity score**	**Completeness score**	**V_measure**	**Adjusted Rand-Index**	**Adjusted Mutual info score**
[[2588 1315] [46 3933]]	0.417	0.453	0.434	0.429	0.417

From the contingency matrix, it can be seen that the clustering algorithm performed well for class 1(in our case, recreational activity) while it performed poorly for class 0, grouping a third of the documents from the other class into class 0. Our analysis is further affirmed by the lower values of purity scores we have obtained.

Part 3: Preprocessing the data

Part 3a(i): Dimensionality Reduction using LSI and NMF

In this part, first, we use Truncated SVD to reduce the dimensions of the data. Also, we plot the percentage of variance the top r principal components can retain v.s. r, for r=1 to 1000, which is as shown below:



We calculate the ratio of variance of the original data retained after dimensionality reduction by using SciPy's linalg.svds to decompose the matrix into U, S and V' and then calculate the ratio using trace(S'S).

The plot is observed to be monotonically increasing in nature.

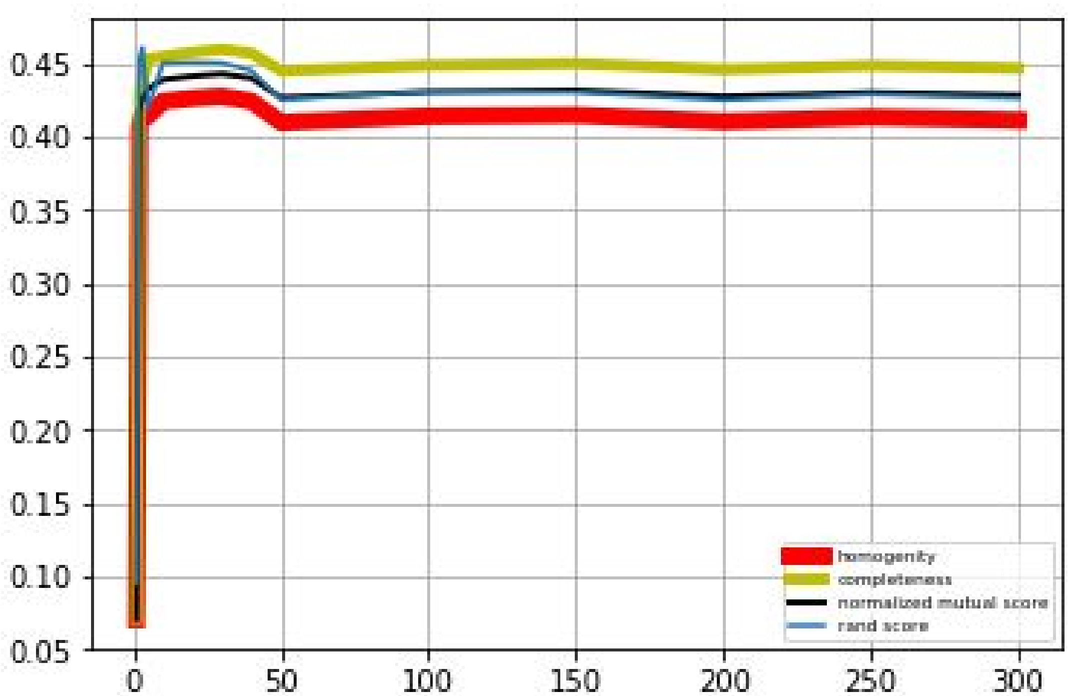
Part 3a(ii): Finding the best dimension parameter r

Here, we use TruncatedSVD and NMF for dimensionality reduction and vary the dimension parameter r, calculating various metrics.

Truncated SVD:

We observed that the homogeneity values were around 0.41 on average. The following table and plot show the values of various metrics for dimensions r = [1, 2, 3, 5, 10, 20, 30 , 40, 50, 100, 150, 200, 250, 300]. The homogeneity and completeness scores were found to be highest for **r=30.**

R	Contingency Matrix	Homogeneity Score	Completeness score	V-measure	Adjusted Rand Score	Adjusted mutual info score
1	[[1557 2346] [2813 1166]]	0.070445375015008249	0.071049158265143728	0.07074597841287493	0.095402819288579882	0.070360273771262
2	[[2748 1155] [133 3846]]	0.40662838691795128	0.42928128019082612	0.41764789035596445	0.45310245269969657	0.40657406250570954
3	[[1140 2763] [3853 126]]	0.41418460472985036	0.43688743339020164	0.42523321365337691	0.46065036633303785	0.41413097212235217
5	[[2551 1352] [37 3942]]	0.41391287710227881	0.45321288251607822	0.43267229940120838	0.419251431298241	0.41385921884221583
10	[[2676 1227] [69 3910]]	0.42405782067972853	0.45472019029533872	0.43885406896277857	0.45037357216015644	0.42400509165447903
20	[[1231 2672] [3915 64]]	0.42618164319127272	0.45750611161609461	0.44128869130047033	0.45071430317761102	0.42612910858501563
30	[[1236 2667] [3919 60]]	0.42746593516022319	0.45940009782806085	0.44285807579996539	0.45037362421229793	0.42741351811158901
40	[[2649 1254] [58 3921]]	0.42415149212156966	0.45699306109079668	0.43996024952820356	0.44493991736649258	0.42409877157496906
50	[[2589 1314] [57 3922]]	0.40966701992395554	0.44494798542304842	0.42657925280732817	0.42518804075614225	0.40961297311380784
100	[[2606 1297] [57 3922]]	0.41391668324333331	0.44853629138386492	0.43053165681091454	0.43083328214074418	0.41386302555109289
150	[[2606 1297] [57 3922]]	0.41463058686183296	0.44979178172198742	0.4314960769156288	0.43016722335621016	0.41457699450713359
200	[[2586 1317] [55 3924]]	0.41022523551229495	0.44585740176026928	0.42729977149103898	0.42485713990731089	0.41017123979409847
250	[[2598 1305] [55 3924]]	0.4132204641145456	0.44838280484421961	0.43008414056418204	0.42883660667029683	0.41316674265314107
300	[[2590 1313] [55 3924]]	0.41122194806756335	0.44669746411863759	0.42822623846128466	0.4261815681428498	0.41116804361302606

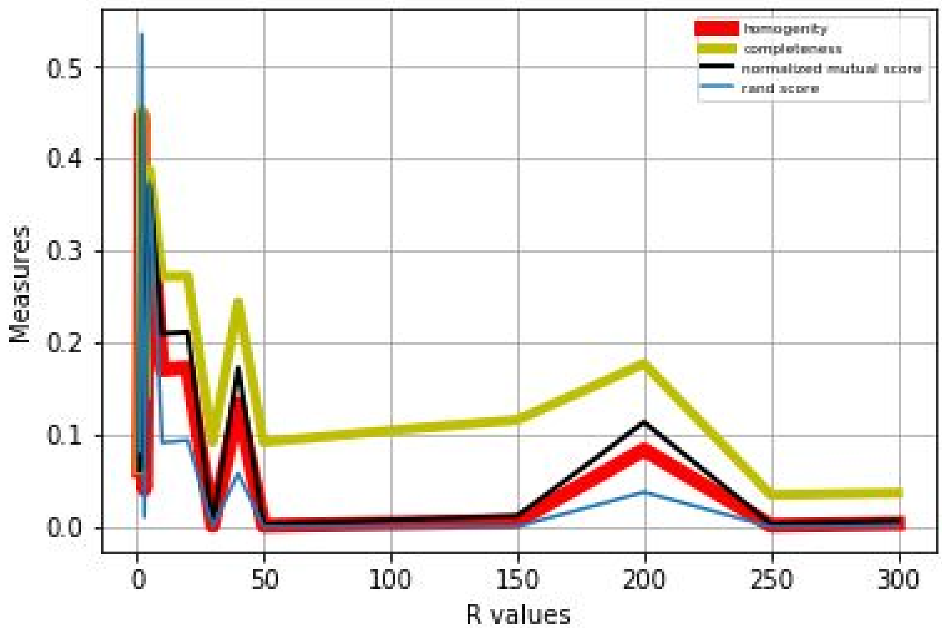


NMF:

For NMF, we observed the following metrics for different r values. The best dimension parameter r was observed to be at r=2.

R	Contingency Matrix	Homogeneity Score	Completeness score	V-measure	Adjusted Rand Score	Adjusted mutual info score
1	[[1690 2213] [2855 1124]]	0.060701243874015265	0.061747532751851165,	0.061219918197158898	0.081661929849436113	0.060615250286200378
2	[[3633 270] [788 3191]]	0.44717190609737173	0.45199058291946975	0.44956833268958962	0.53509233922553578	0.44712129433799758
3	[[3517 386] [3962 17]]	0.040593827540793721	0.13940483431334139	0.062877976360230362	0.010566870787302477	0.040505900735849566

5	[[2486 1417] [110 3869]]	0.35415499164683206	0.38733806889581135	0.37000403066188103	0.37512217678006504	0.35409586237684787
10	[[3899 4] [2746 1233]]	0.17035731909746471	0.27171081663207342	0.20941534822379637	0.091228613549751675	0.1702813479106218
20	[[5 3898] [1251 2728]]	0.17233076995976931	0.27233238774706481	0.21108674848378528	0.093854530423786428	0.17225497993163041
30	[[3893 10] [3979 0]]	0.0012877195572840966	0.091729426573068443	0.0025397850071869981	5.5204716378814561e-05	0.001190770702596841
40	[[3 3900] [991 2988]]	0.13340076593020481	0.24400666587271397	0.17249621166174325	0.058006129706685221	0.13332140294893544
50	[[3893 10] [3979 0]]	0.0012877195572840966	0.091729426573068443	0.0025397850071869981	5.5204716378814561e-05	0.001190770702596841
100	[[0 3903] [30 3949]]	0.0037639895218942393	0.10433864137514494	0.0072658648472445778	-8.9374349206971393e-05	0.00367120686194346
150	[[3855 48] [3979 0]]	0.0062028073957758586	0.11576742163576162	0.011774726091996297	0.0003824367057137385	0.0061108667769254497
200	[[3142 761] [3945 34]]	0.083442146005835605	0.17687562218316386	0.11339127247253541	0.037578607310198311	0.083358197399429235
250	[[7 3896] [32 3947]]	0.0015519188963649621	0.034471703212745383	0.0029701226291853041	-8.2766774183442112e-05	0.001459313258783166
300	[[82 3821] [21 3958]]	0.0036871667555791821	0.036683936572652245	0.0067008221348199715	0.0005363702660300263 4	0.003595522725708821



The best r choice for SVD is at r=30 and for NMF is at r=2 as seen from the above results.

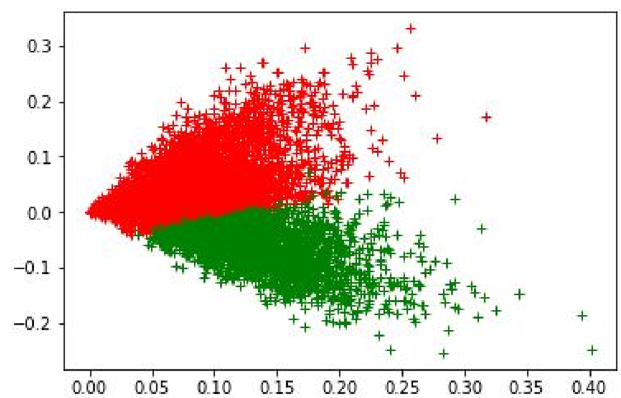
Q: How do you explain the non-monotonic behavior of the measures as r increases?
A: We observed that after a certain dimension, these metrics became almost constant, indicating higher dimensions didn't change the clustering result much. The number of singular values actually significant in the reconstruction of original matrix dips after a certain r. At this point, inclusion of these values in computing the metrics might be attributed to drop in scores.

Part 4: Visualization of Clusters

Part 4a: Visualizing the performance of case with best clustering results.

SVD:

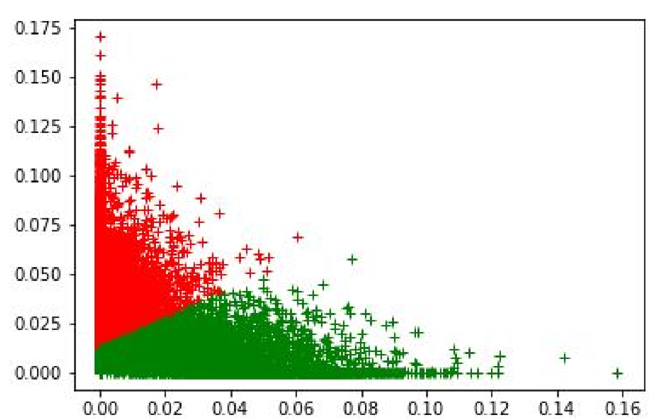
Following figure shows the TF-IDF data reduced in dimensions at r=30 using Truncated SVD. As it can be seen from the plot below, the data is very close to 0 and is not easily separable.



Clustering results with TruncatedSVD

NMF:

Following figure shows the TF-IDF data reduced in dimensions at r=2 using NMF.

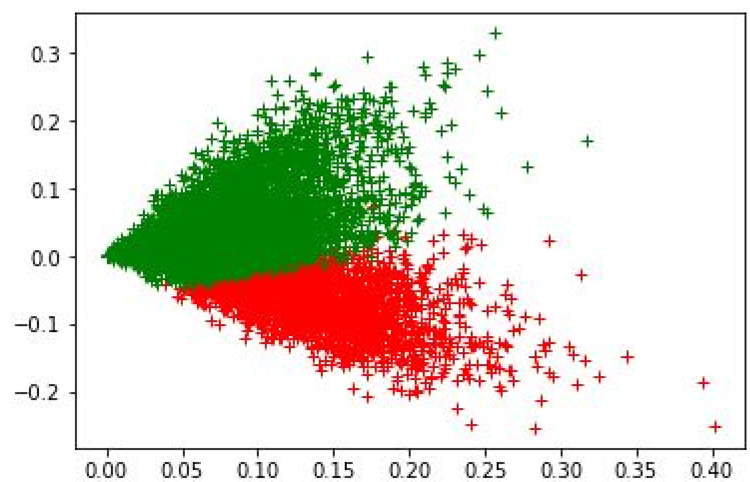


Clustering results with NMF

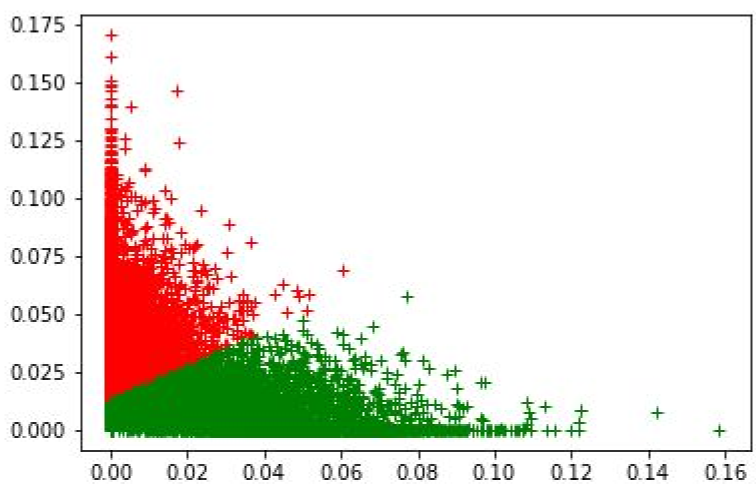
Part 4b(i): Visualizing the transformed data after normalizing

In this part, we apply various methods to improve the clustering performance. We visualized the clusters and calculated various purity scores for SVD and NMF after normalizing the data. We used StandardScaler for normalizing the data. It was observed that the result, after normalizing, is almost the same as without normalizing (though there is a slight improvement).

The following plots show visualization of the clusters.



Clustering results with TruncatedSVD (normalized)



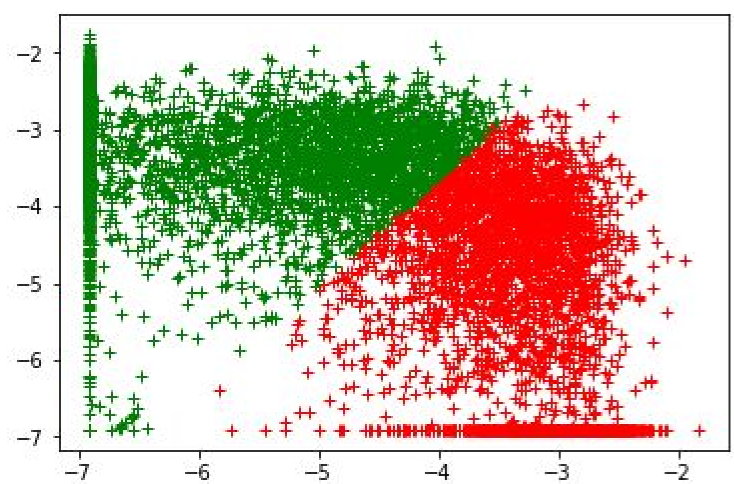
Clustering results with NMF (normalized)

The following table shows the values of purity metrics that we obtained after normalizing the data.

	Contingency Matrix	Homogeneity score	Completeness score	V_measure	Adjusted Rand-Index	Adjusted Mutual info score
SVD	[[2561 1342] [34 3945]]	0.41858669382846714	0.45787200406708439	0.4373489106543057	0.42353491242879165	0.41853346349325249
NMF	[[300 3603] [3269 710]]	0.45837858761389916	0.46131721334117681	0.45984320570718235	0.5530626716804482	0.45832900190967441

Part 4b(ii): Visualizing the transformed data after a non-linear (logarithm) transformation (NMF)

It was observed that applying non-linear logarithmic transformation on the NMF-reduced data improves the clustering performance significantly. The following plots show visualization of the clusters for log transformation.



Clustering results with log transformation on NMF-reduced data

The following table shows the values of purity metrics that we obtained after applying logarithm transformation to the data.

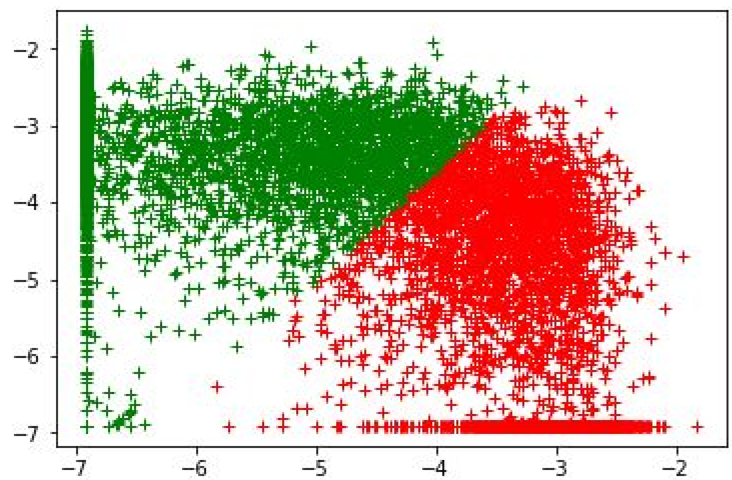
	Contingency Matrix	Homogeneity score	Completeness score	V_measure	Adjusted Rand-Index	Adjusted Mutual info score
NMF	[[3422 481] [302 3677]]	0.53549672000289172	0.53663502705530663	0.53606526924564746	0.64206750415703473	0.53545419456551313

In this task, we observed that the clustering results improve on a logarithm transformed data. However, while performing log transformation, one must be very careful of the 0s in the original TF-IDF matrix. The logarithm of 0 is undefined. Hence, to avoid this problem, we add a very small constant - 0.001 to all the values while performing this transformation.

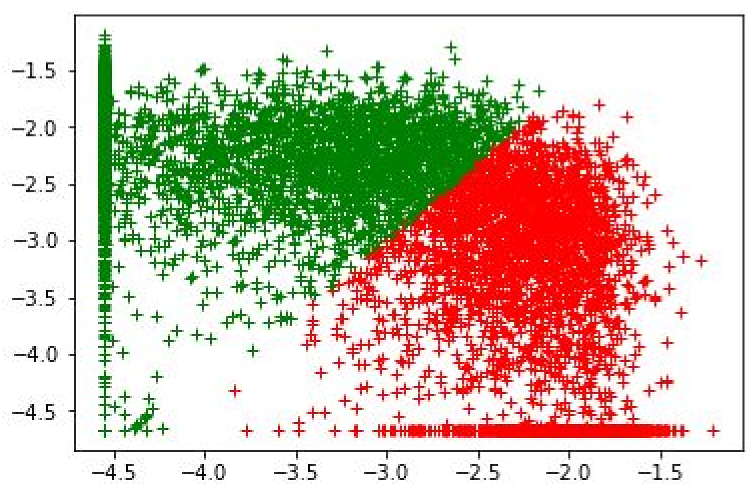
Q: Can you justify why logarithm transformation may increase the clustering results?
A: The scoring functions can be considered additive by means of applying log transformation based on the following :
For independant terms, P(a.b)= P(a) * P(b)
Applying log, log(P(a.b))= log(P(a))+log(P(b))
Thus the influence of outliers is amortized

Part 4b(iii): Combination of normalization and log transformation on NMF-reduced data

It was observed that applying logarithm transformation after normalization on the NMF-reduced data and vice-versa improves the clustering result. The visualization plots for these transformation are as shown below:



Clustering results with log transformation on normalized NMF-reduced data



Clustering results normalization on log transformed NMF-reduced data

	Contingency Matrix	Homogeneity score	Completeness score	V_measure	Adjusted Rand-Index	Adjusted Mutual info score
norm ->log	[[3438 465] [322 3657]]	0.53294375394727811	0.53372038411572298	0.53333178630219658	0.64044168522514522	0.53290099479503594
log-> norm	[[3406 497] [288 3691]]	0.53562830842554687	0.53711530816422526	0.5363707776813319	0.64125435364257122	0.53558579502538894

Part 5: Expand Dataset into 20 categories

All the analysis thus far was conducted by setting k=2 for two classes. The analysis is extended to the whole dataset of 20 subclasses with 20 clusters. The TF-IDF representation is found and dimensions are reduced using both truncated SVD and NMF and then different transformations are applied. The contingency matrix, purity scores and visualization is obtained in each case and tabulated below.

Part 5.1: Building the TFIDF matrix

We compute the TF-IDF representation of the data, in order to make the text data into analyzable and numeric format by setting min_df=3 and excluding stopwords

Shape of TF-IDF matrix: (18846, 52295)

Part 5.2: Applying K means with k=20 to the TFIDF matrix

We applied K-means clustering algorithm with k=2. The following measures are computed and inspected to get a sense of the effectiveness of the clustering algorithm against the known class labels.

Contingency matrix:

We obtained the following contingency matrix for k=20.

```
[[  0  68   0 174  92   1   0   0  28   0   1   0 229 168   0   0  38   0  0   0]
 [ 22  73   1   1  93   0   2 379   2  36  16   2 318   0   0   0   0   4 22   2]
 [ 13  36   5   0  74   0   1 133   0  71   5   0 207   0  11   0   0   2 417  10]
 [144  27   1   0 145   0   3  36   3 104  13   0 256   0 197   0   0   5 43   5]
```

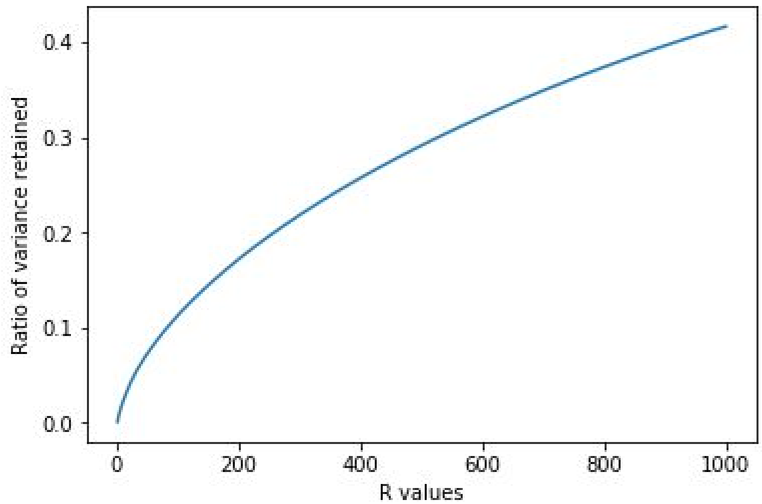
```
[542  21   1   0  47   0  13   8   2   7  11   1 231   0  69   0   0   3 6   1]
[  4  73   2   0  97   2   2 543   0   4  31   0 170   0   0   0   0   4 55   1]
[ 92   5  14   0 164   0  30   2   6   8   4   0 561   0  46   0   0  12 23   8]
[  0  25   2   0 527   0  32   0   9   0  11   3 373   0   0   0   0   5 2   1]
[  3 108   0   0 562   0   6   0   2   0  24   0 278   0   0   0   0  12 0   1]
[  3   2  14   0 149   0   7   1 425   1   7   0 381   0   0   0   0   3 0   1]
[  8   3  19   0  42   0   6   0 660   0  10   0 198   0   0   0   0  53 0   0]
[  4  49   0   0 148 552  11  20  32   0  11   0 123   0   0   0   0   0 4  37]
[ 45  53  17   0 257   1   6  26   1  18  35   2 505   0   6   0   0   7 4   1]
[  2  22  15   3 300  75   5   2  14   0  24   0 523   0   0   0   0   1 1   3]
[  1  22   4   0 113   0   0   8   9  84 114 316 213   0   0   0   0   0 0 103]
[  0  15  29 576 105   0   0   0  18   0   5   0 247   0   0   0   1   0 1   0]
[  0  16   7   1 274   5   5   1 429   3   7   0 154   0   0   0   0   5 0   3]
[  0   5   3   4  49   0   2   0 292   0   0   0 157   0   0 410   0  18 0   0]
[  0  12   0   4 170   2 155   0 183   0  23   0 206   0   0   0   0  19 0   1]
[  0  10   0 147 101   1   2   1  53   0   0   1 185  39   0   0  70  14 0   4]]
```

We observed the following purity metrics for the entire dataset with 20 subclasses. It can be observed that the quality of clustering results decrease as the number of clusters increase.

Homogeneity score	Completeness score	V_measure	Adjusted Rand-Index	Adjusted Mutual info score
0.303	0.385	0.339	0.100	0.301

Part 5.3a(i): Dimensionality Reduction

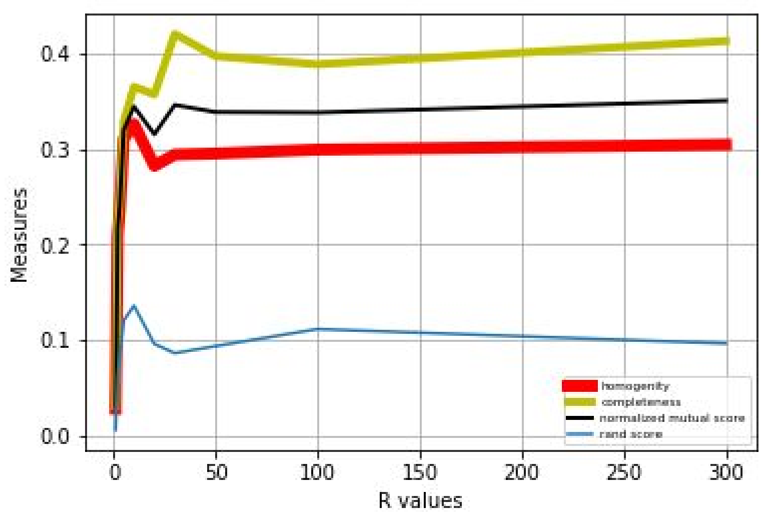
Visualization of variance retained for different r values. We performed TruncatedSVD on the TF-IDF data and used the explained_variance_ratio_ attribute to calculate the variance ratio. The graph is seen to be monotonic in nature as shown below.



Part 5.3a(ii): Preprocessing data to find the best dimension parameter

Truncated SVD:

We observed that the homogeneity values were around 0.29 on average. The following table and plot show the values of various metrics for dimensions r = [1, 2, 3, 5, 10, 20,30, 50, 100, 300]. The homogeneity score was highest for **r=10**.

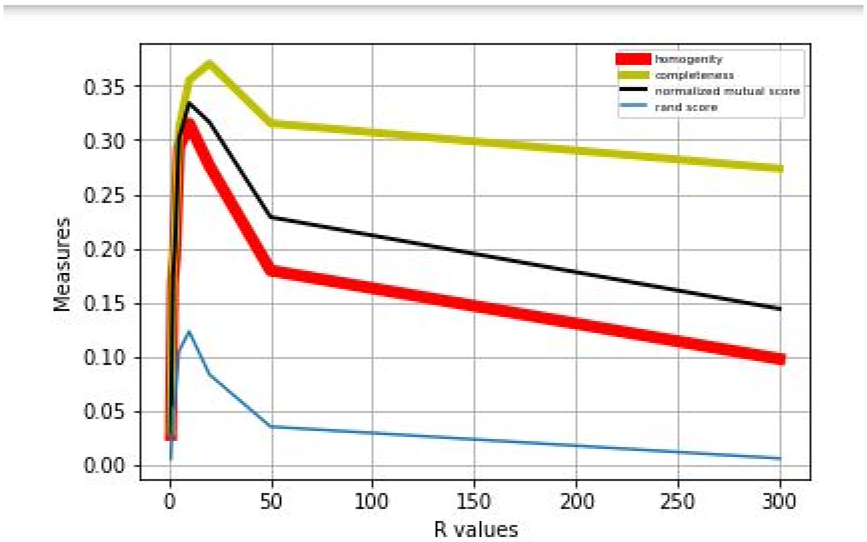


R	Homogeneity Score	Completeness score	V-measure	Adjusted Rand Score	Adjusted mutual info

					score
1	0.028107019773196011	0.03091595331920349	0.029444647252510375	0.0060211751706676581	0.024936303456201764
2	0.20951914133391025	0.22306941421762116	0.21608205545403908	0.06424905302647238	0.20695783280876393
3	0.23558058716226349	0.24549010650209704	0.24043328431326175	0.081917412999576403	0.23310933214839799
5	0.31101910700863244	0.32918596083978752	0.31984477700014213	0.12059337914262762	0.30879086583944593
10	0.32596093769562934	0.36501435814036226	0.34438401247862177	0.13591322714360388	0.323769613002518
20	0.28224939991399189	0.35711827796697615	0.31530032918949835	0.09582046706647783	0.27990415487132947
30	0.29405050619360712	0.42031228436815843	0.34602317368925778	0.086230069501824477	0.29173333027897275
50	0.29512333514028388	0.39711440184060065	0.3386054253978738	0.09355392916604989	0.29281381242239518
100	0.29924506498024189	0.38858658054117684	0.33811359887736797	0.11142773816226147	0.29694881134254736
300	0.30449722455296929	0.41297838819996124	0.3505367172683273	0.096385719561590269	0.30220353968889985

NMF:

For NMF, we observed the following metrics for different r values. The best dimension parameter r was observed to be at **r=10.**



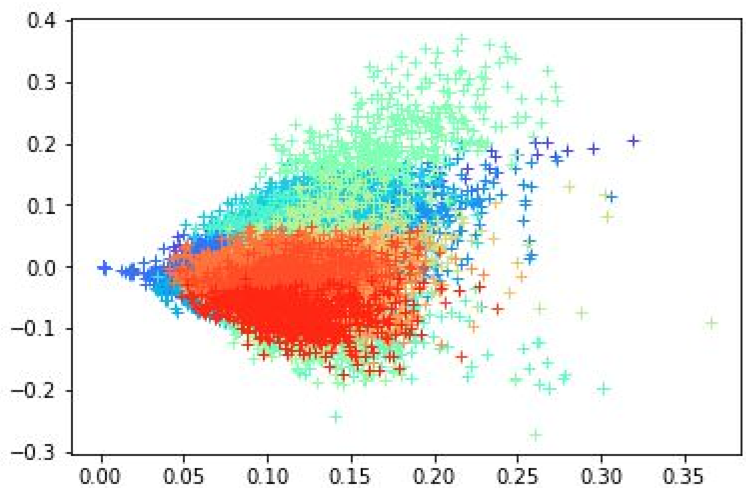
R	Homogeneity Score	Completeness score	V-measure	Adjusted Rand Score	Adjusted mutual info score
1	0.0278531623 77013976	0.0309756590 98749334	[0.029331543 313950131	0.005782541 4233596993	0.02468068829 7136693
2	0.1688224344 9500522	0.1796214674 0078624	0.1740546082 1199817	0.050696060 258931044	0.16611255406 74267
3	0.1911178761 4765977	0.2078624127 613814	0.1991387743 3148153	0.056577107 014621417	0.18849267824 615845
5	0.2927499762 0352683	0.3101205781 5998179	0.3011850262 6955956	0.105156580 47825127	0.29046111307 105998
10	0.3149473319 994191	0.3550860580 5453686	0.3338144285 7784027	0.123181728 56292847	0.31272075720 611181

20	0.27588442670080543	0.370388442162289	0.31622680742364301	0.083526929197327515	0.27351282948276689
50	0.17942356833031942	0.31520341067060154	0.22867705601760596	0.035448645286705514	0.17672613002141202,
300	0.097885513830824361	0.2735190331817568	0.14417460055813272	0.0061614448866317457	0.094810501472847683

Part 5.4a: Visualizing the performance of case with best clustering results

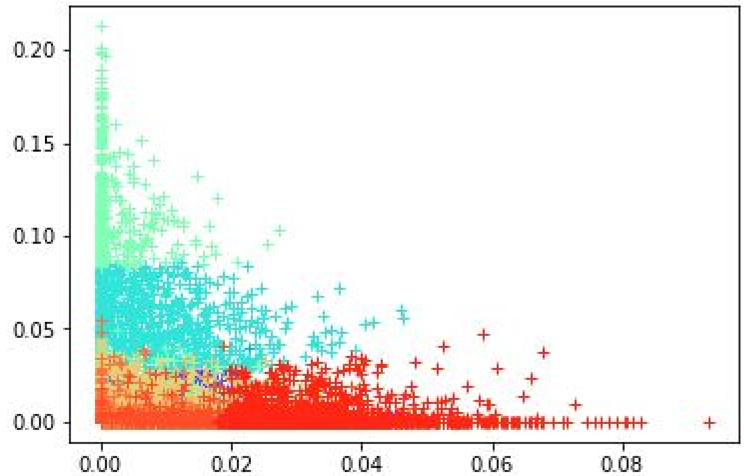
SVD:

Following figure shows the TF-IDF data reduced in dimensions at r=10 using Truncated SVD.



NMF:

Following figure shows the TF-IDF data reduced to dimensions at r=10 using NMF.

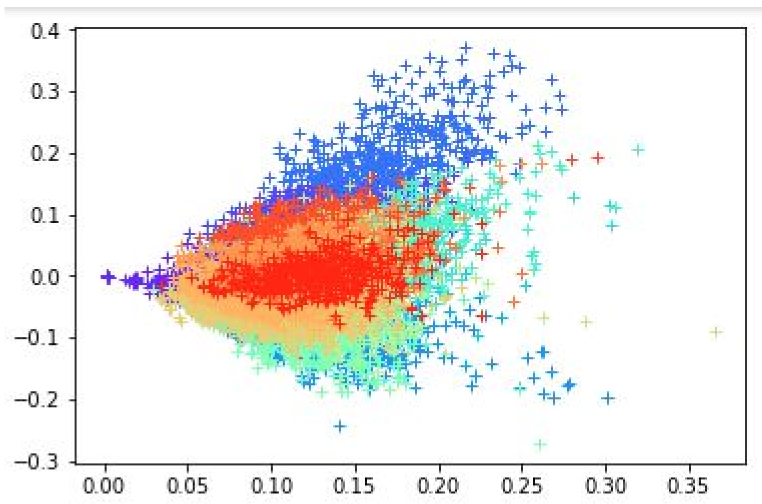


Part 5.4b(i): Visualizing the transformed data after normalizing

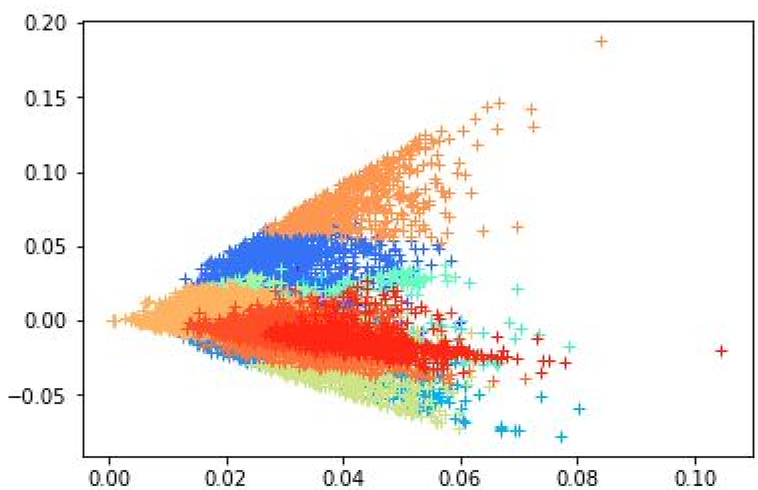
We visualized the clusters and calculated various purity scores for SVD and NMF after normalizing the data. We used StandardScaler for normalizing the data. It was observed that the result improve slightly after normalizing.

The following plots show visualization of the clusters.

SVD:



Clustering results with TruncatedSVD(normalized)



Clustering results with NMF (normalized)

The obtained Contingency matrices for normalized TruncatedSVD and NMF are listed below:

SVD:

```
[[  0  59 282  58   0   1 121   1  38   0   0  68   7 116   0  38  10   0 0   0]
 [  0 150   1   0  69  16 164   0   4  19   0  38   9   1 420  82   0   0 0   0]
 [  0  64   0   0 363   3  73   0   2  39   2   4   9   0 369  57   0   0 0   0]
 [  4 104   0   0  43   4  88   0   1 305 105  13  29   0 171 113   0   0 0   2]
 [  2 173   0   0   9   9 239   0   2 258  22  13  34   0 119  83   0   0 0   0]
 [  0 119   0   0  95  23 108   0   0   1   0  40   4   0 491 101   0   0 0   6]
 [ 22 242   1   0  16   5 312   0   3 124  16   6  45   1  55 125   0   2 0   0]
 [  8 292   0   0   0   8 227   0  79   6   0   8  40   4   8 310   0   0 0   0]
 [ 22 260   5   0   0  16 161   0  29   2   0  52  15  30   1 403   0   0 0   0]
 [416  95   0   0   0   2 205   0  10   0   0   1   8   3   1 126   0 127 0   0]
 [471  42   0   0   0   3  64   0   3   0   0   1  10   1   0  36   0 368 0   0]
 [  0  49   0   0   6   5  55 222  43   0   0  35   6  14  32  64   0   0 0 460]
 [  7 372   1   0   5  33 222   0   2  33   1  27   6   0  94 176   0   0 0   5]
 [  3 425  13   1   2  18 203   0  50   0   0 101   7   5   8 154   0   0 0   0]
 [  0 239   0   0   0 526 114   0  21   1   0   9   1   1   6  69   0   0 0   0]
 [  1 135 494 228   1   3  53   0  16   1   0   9   1   2   3  48   2   0 0   0]
 [  0  86   1   0   1   4  93   2 455   0   0   8  19 100   0 132   6   0 0   3]
 [  0  82   5   2   0   0  90   0  54   0   0   3  19   5   1  26 482   0 171   0]
 [  7 110  14   2   0  13  87   1 285   0   0  71  45  47   0  87   5   0 0   1]
 [  0  89 172  81   0   2  76   0  68   0   0  21  18  49   0  48   3   0 0   1]]
```

NMF:

```
[[ 14   0 105   1 230   1   0   2   8   0 108   0   5  49   1  88   0   0 115  72]
 [ 19   3  98   1   1   0   0  13   0   0 127 102 419   0   8 159   0   1 4  18]
 [ 12  24  70   1   0   0   2  12   0   0  51 340 382   0   2  77   0   0 2  10]
 [ 16 199 137   6   0   0  91  16   0   0  95  38 228   0   2 126   0   4 1  23]
 [ 17 130  97   7   0   0  15  10   0   0 236   6 186   0   4 237   0   2 1  15]
 [ 12   1 124   1   0   0   0  18   0   0  73 133 472   0  15 108   0  12 0  19]
 [ 18  89 129  19   1   0  12  11   0   0 343   8  68   0   1 246   3   1 3  23]
 [  5   2 333   7   0   0   0  16   0   0 244   0   8   0   4 220   0   0 82  69]
 [ 10   2 373  27   8   0   0  20   0   0 137   0  16   0   2 207   0   0 35 159]
 [ 25   0 104 459   0   0   0   3   0   0 152   0   2   0   1  89 127   0 2  30]
 [  9   0  24 470   1   0   0   3   0   0  42   0   0   0   0  35 407   0 1   7]
 [  8   0  71   2   0 243   0  15   7   0  36   7  33   0  11  50   0 431 31  46]
 [ 22  14 190  11   0   0   1  45   0   0 174   5 113   0  14 356   0   8 5  26]
 [120   0 196   4   7   0   0  29   2  75 114   2  11   1   2 317   0   0 87  23]
 [  5   1  65   2   0   0   0 401   0   0  57   0   8   0 285 130   0   0 17  16]
```

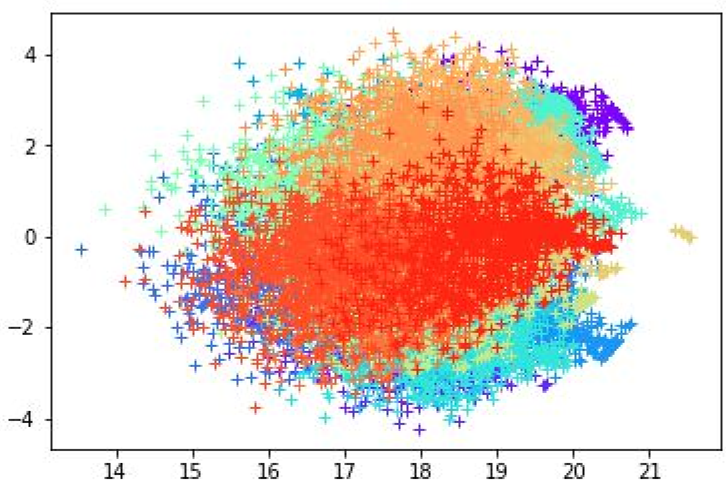
```
[ 2   1  56   0 495   0   0   3   2   0  44   1   6 209   2 141   0   0 27   8]
[ 11   0 123   1   1   2   0   5 153   0  72   0   0   0   3  43   0   4 375 117]
[ 9   0  50   0   2   0   0   0 352   0  89   0   1   2   0  58   0   0 370   7]
[ 7   0 108   4   6   1   0  29  74   0  98   0   1   1   4  59   0   1 322  60]
[ 14   0  78   0 170   0   0   5  14   2  67   0   1  69   0  89   0   1 83  35]]
```

The following table shows the values of purity metrics that we obtained after normalizing the data

	Homogeneity score	**Completeness score**	**V_measure**	**Adjusted Rand-Index**	**Adjusted Mutual info score**
SVD	0.33423956714648284	0.37698192846077722	0.3543264014628591	0.13146145137821147	0.33208181010857502
NMF	0.31677119852076052	0.3567313158799324	0.33556580432881578	0.12424793493359659	0.31455061654218192

Part 5.4b(ii): Visualizing the transformed data after a non-linear (logarithm) transformation (NMF)

The following plot shows visualization of the clusters for log transformation.



Clustering results with log transformation on NMF-reduced data

The following table shows the values of purity metrics that we obtained after applying logarithm transformation to the data.

Contingency matrix:

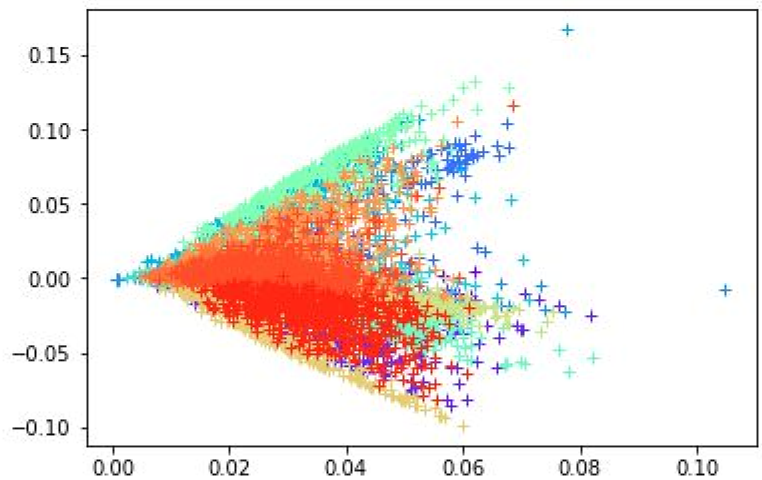
```
[[  0 188   1  84  22   4 251  34   7   4   0   2   1 166   2   0   1  19  13   0]
 [208   8   7  44   0  73   1   1 239  24  86   5   1   0 151  13  36  20  53   3]
 [276   6   6  11   0  47   0   0 223  11  57   4   2   0 199  42  59   9  31   2]
 [ 44   0   4   9   0  37   1   1  73  17 147   3   5   0  44 273 281  13  27   3]
 [ 21   0   8  24   8  25   0   0 109  13 242   1   6   0  21 275 163  22  19   6]
 [238   3   9  21   0  87   0   0 159  34  14   9  13   1 338   0  21   5  35   1]
 [ 13   1  39   3  10  47   0   2 138  20 157   6   4   2  36 262 112  64  19  40]
 [  7   1  38  22  22  12   7  65  67  30  33   7   2   1  77   7  29 409 147   7]
 [  1   2 146  28   8  17  20  69  34  28  17   9   0   0  89  15  37 302 167   7]
 [  0  11 260  39   3  22   0   0  27   6  13   0   1   1   7   0   1  86  44 473]
 [  1   8 230   9   1  10   0   0   5   8   3   0   0   1   5   0   0   8  10 700]
 [  7   3   2  28   6  13   0   6  12  15  26 459 342   0  17   0   9  11  34   1]
 [ 24   1  37  47   6  38   5   4 110  53 261  23  16   6  52  24  57  68 147   5]
 [  8  10  13 317  26 127  12  59  52  31  17   7   1  14  40   1   6  76 168   5]
 [  5  12   5  69   7  16   2   6  29 660  26   6   1   2  18   0   4  34  81   4]
 [  5 344   6  20   2   4 225   1  11   9   5   1   0 342   0   0   1   5  14   2]
 [  1  36   8  54 229   1   9 370   4  12   7  22   9   3  16   0   2  71  56   0]
 [  2  44   1  45 575  18   3 186  10   0   2   1   3   2   2   0   0  31  12   3]
 [  0  49   8  54 138   3  14 273  14  29   3   8  12  13   5   0   1  77  67   7]
 [  1 111   1  48  31   7 215  43   6   5   0   4   0  99   8   0   1  32  15   1]]
```

Homogeneity score	**Completeness score**	**V_measure**	**Adjusted Rand-Index**	**Adjusted Mutual info score**
0.37091220348468257	0.37499089590867285	0.37294039829383263	0.20052309895145071	0.36888179705382163

Part 5.4b(iii): Combination of normalization and log transformation on NMF-reduced data

Clustering results with log transformation on normalized NMF-reduced data

It was observed that applying logarithm transformation after normalization on the NMF-reduced data and vice-versa improves the clustering result. The visualization plots for these transformation are as shown below:



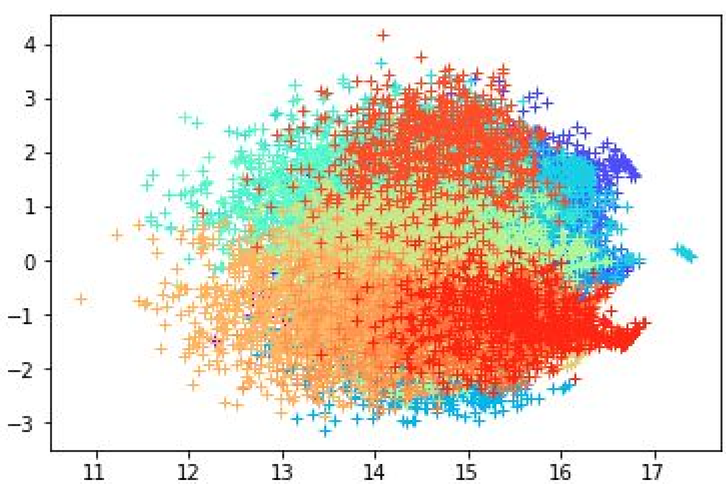
Clustering results with log transformation on normalized NMF-reduced data

Contingency matrix:

```
[[  3 232   7  65  25   3  16   0   0   1   0  19   1  34 133   1   1   0  17 241]
 [ 24   6 270  51   9   6  47 195 155   2   1   0 104   1   0  40   9  24  27   2]
 [ 12   5 243  27  17   2  28 258 214   2   0   0  54   0   0  61   4  47  11   0]
 [ 16   0  78  18  15   2  26  41  47   6   3   0 127   1   0 292   5 296   8   1]
 [ 13   0  99  30  14   9  19  19  18   1   0   8 196   0   0 175   9 328  25   0]
 [ 36   1 210  36  10   3  33 220 366  10   0   0  27   0   1  20  11   0   4   0]
 [ 13   1 115  21  81  59  15  11  18   3   7  13 130   1   2 104  44 288  49   0]
 [ 22   1  50  27 136  15 108   5  34   7   0  24  21  56   1  22  43   7 406   5]
 [ 28   1  23  38 147  15 134   0  55   1   0  11   9  62   0  29 149  15 261  18]
 [  5   3  20  42  13 430  29   0   4   0  86   3  15   0   0   0 239   0 105   0]
 [  3   2   2   7   3 467   9   1   3   0 311   1   1   0   1   0 176   0  12   0]
 [ 18   4  15  44   9   2  44   9  23 721   1  10  54   8   0  15   1   0  13   0]
 [ 52   2 112  51  25   8 131  22  52  16  10   8 280   4   6  64  38  27  73   3]
 [ 31  20  52 378  71  10 195   7  26   1   1  24  29  47  11   5  11   1  63   7]
 [649  15  28  42  26   3  88   4  12   3   3   8  37   1   2   4   6   0  54   2]
 [  9 365  10  16   0   4  29   4   2   1   0   2   5   1 307   1   6   1   4 230]
 [ 12  36   6  48  32   0  46   1   4  13   0 241   8 372   2   1   8   0  69  11]
 [  1  52   7  43  20   6   9   2   1   0   0 570   4 185   0   0   1   0  35   4]
 [ 29  52  10  48  19   7  68   0   4   6   0 142   7 271  10   1   6   0  76  19]
 [  5 115   5  46  19   1  25   0   4   1   1  31   1  43  86   1   1   0  23 220]]
```

Homogeneity score	Completeness score	V_measure	Adjusted Rand-Index	Adjusted Mutual info score
0.37680804025112846	0.38018959683635245	0.37849126572958436	0.20441527455803085	0.37479677850533338

Clustering results with normalization on log transformed on NMF-reduced data



Clustering results with normalization on log transformed on NMF-reduced data

Contingency matrix:

```
[[  1   1   0   9   0 273   7   0   0  10  22  80   5   3 226  85  24  34 1  18]
 [  6   8 164 195   3   0 152  23  52 128 127   3  26  23   0  37   3   1 22   0]
 [  5   3 247 214   0   1 187  61  26  94  60   3  16  17   0   6   0   0 45   0]
 [  6   1  33  56   2   0  52 270 128  61  18   9  17  20   0  10   3   1 295   0]
 [  2  11  17  73   3   0  24 369 169  42  39   5  10  33   0  16   9   8 133   0]
 [ 19   2 197 144   6   1 327   5  14 137  58   3  37   5   1  12   1   2 17   0]
 [  5  52   8 106  15   1  55 329  92  44  60  18  16  65   2   1  12  11 83   0]
 [  9  24   4  51   9   6  78  21  22   6  67 212  28 335   1  32  18  33 21  13]
 [  3  41   0  24  28   7 119  24   6  11  56 175  34 305  12  42  38  12 31  28]
 [  1 435   0  18 185   6  16   0   4   6  34  22   9  90   0  28 135   4 1   0]
 [  0 439   1   3 422   1   6   0   2   1   4   1   9  18   0  11  80   1 0   0]
 [765   1   4  11   2   3  20   0  51  25  29   6  20   4   0  25  10   8 2   5]
 [ 21  15  23  78  11   1  67  43 243  57 135  13  50 135   2  35   7   8 37   3]
 [  5   8   5  30   3   8  39   2   7  58  73  83  33  52   7 364 171  26 5  11]
 [  4   3   3  19   2   3  14   1  16  13  57  17 691  50   1  75   6  11 1   0]
 [  1   2   4   7   4 463   2   1   0   8  36   6  10   8 402  26   8   8 1   0]
 [ 21   4   1   2   2  25  23   0   4   6  28 212  17  47   3  47  25 215 2 226]
```

```
[ 0  9  2  6  0 40  2  0  1  4 16  99  0  16  0  25  39 364 0 317]
[ 20  7  0  7  4  54  6  0  9  3  39 184  37  28  13  68  15 139 0 142]
[ 2  3  0  4  1 182  15  0  0  6  15  56  5  7 191  59  22  39 0  21]]
```

Homogeneity score	Completeness score	V_measure	Adjusted Rand-Index	Adjusted Mutual info score
0.36463410981327699	0.36647250640774032	0.36555099674991942	0.20206093217031865	0.36258388657892437