Clustering

Akshya Arunachalam UID : 904943191
Ashwin Kumar Kannan UID : 605035204

February 12, 2018

1 Introduction

When you have unlabeled data and no prior information, clustering is used to group them together based on their feature similarity. The number of groups is K (K-means clustering). The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. The objects in a single cluster are similar to each other that the objects in another cluster. The K means clustering algorithm repeats two steps iteratively :

1. Data Assignment Step : In this stem the object is assigned to the cluster that it has the least euclidean distance to.

2. Center updating step : After every time a data object is assigned to a cluster, the center of the cluster is recalculated. This is done by taking the mean of all data points assigned to that centroid's cluster.

These two steps are repeated until a certain criteria is met.

2 Data Set

We work with "20 Newsgroups" dataset. It is a collection of approximately 20,000 documents, partitioned (nearly) evenly across 20 different newsgroups, each corresponding to a different topic. Each topic can be viewed as a "class". The following eight categories are loaded, each belonging to two distinct classes. The documents belonging to these classes are used to test the clustering algorithm and measure performance parameters.

- comp.graphics
- comp.os.ms-windows.misc
- comp.sys.ibm.pc.hardware
- comp.sys.mac.hardware
- rec.autos

- rec.motorcycles
- rec.sport.baseball
- rec.sport.hockey

3 Question 1

Proper document representation and the non inclusion of irrelevant words in the document is important for computation accuracy of the algorithm. Hence, common words, stopping words, punctuation and the like are being excluded from the documents. The goal of the question is to tokenize the document, remove stopping words, punctuations and represent the occurrence of words in the document and the data set.
Words that occur in a less that three documents are removed by setting $mindf = 3$.
By performing vectorization, we are representing the document in terms of numerical values and occurrence of words in the document irrespective of their position in the document. The data can be represented as a matrix wherein each row represents a document and each column represents a token (word). TF-IDF transformation is done on the data set in order to obtain a vector representation of the document based on the frequency of occurrence of words in the documents.

Dimension of TF-IDF Matrix : (7882, 18469)

4 Question 2

In this section we run the K-means algorithm on the TF-IDF Matrix that we got from the last question with k=2 and evaluate the performances. This will cluster the documents into two different categories based on their euclidean distance from the center of the cluster.
Once the clustering is done, the performance of the clustering algorithm can be compared to the ground truth. There are various purity measures to do so.

Homogeneity : Homogeneity is a metric that is satisfied if all the data points in a cluster belong to only one class.
Completeness : Completeness is satisfied if all cluster have data points belonging only to a single class.
V Measure : V Measure is the harmonic mean of Completeness and Homogeneity.
Rand Index : Rand Index is an accuracy measure that gives us the similarity between the clustering labels and the ground truth.
Mutual Index : Mutual Information gives the MI between cluster labels and ground truth.

The following results were observed :

Homogeneity	0.426939956854
Completeness	0.463834468349
V-Measure	0.444623155541
Adjusted Rand Index	0.43718683605
Adjusted Mutual Index	0.42688749143

Table 1: Stats for K-means clustering with k=2

2606	1297
38	3941

38	3941
2606	1297

Table 2: Contingency Matrix

We do not get good results for the algorithm due to the high dimensionality of the TF-IDF Matrix. In the next parts, we aim to reduce the dimensionality and hence better the performance of the clustering algorithm.

5 Question 3

The TF-IDF vector representation of the document may have very high dimensions, and the algorithms may not perform well on them. This is because the K means algorithm assumes that the clusters are in a shape of a circle, which is why reducing the Euclidean distance seems like the best option.

5.1 Part a

In this part, the aim is to reduce the dimensions of the vector by Latent Semantic Indexing (LSI) and Non-negative Matrix Factorization (NMF) . LSI reduces dimension by reducing the SVD of the term document matrix to its best rank k approximation. The aim is to inspect the top singular values of TF-IDF matrix and find out which ones are relevant in reconstructing the matrix with SVD. We can do this by calculating the ratio of variance of original data that is retained after SVD reduction.

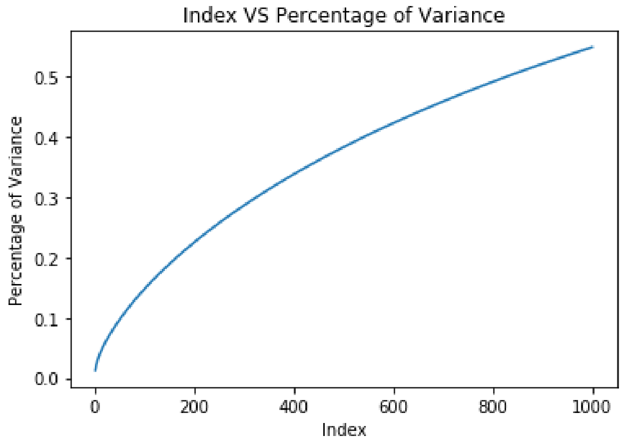

Figure 1: Percentage of Variance VS r

5.2 Part b

In this part, we analyze the performance of the algorithm for various values of r (principle components). We measure the performance metrics for r=1,2,3,5,10,20,50,100,300.

Index	Homogeneity	Completeness	V Measure	ARI	AMI
1	0.0606	0.0616	0.0611	0.0815	0.0605
2	0.4161	0.4499	0.4324	0.4348	0.4160
3	0.4168	0.4506	0.4331	0.4358	0.4168
5	0.4009	0.4464	0.4224	0.3924	0.4008
10	0.4169	0.4554	0.4353	0.4238	0.4168
20	0.4193	0.4579	0.4378	0.4258	0.4193
50	0.4180	0.4574	0.4368	0.4228	0.4180
100	0.1840	0.2821	0.2227	0.1045	0.1839
300	0.1812	0.2800	0.2200	0.1017	0.1811

Table 3: Stats for SVD against various index values

2210	1693
1122	2857

Table 4: Contingency Matrix for r=1(SVD)

1283	2620
3920	59

Table 5: Contingency Matrix for r=2(SVD)

2623	1280
59	3920

Table 6: Contingency Matrix for r=3(SVD)

2454	1449
23	3956

Table 7: Contingency Matrix for r=5(SVD)

2566	1337
38	3941

Table 8: Contingency Matrix for r=10(SVD)

1333	2570
3943	36

Table 9: Contingency Matrix for r=20(SVD)

1344	2559
3945	34

Table 10: Contingency Matrix for r=50(SVD)

3899	4
2662	1317

Table 11: Contingency Matrix for r=100(SVD)

3899	4
2679	1300

Table 12: Contingency Matrix for r=200 (SVD)

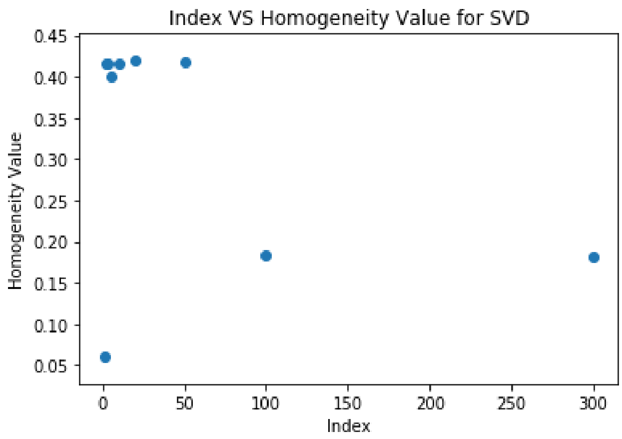

Figure 2: Homogeneity VS r (SVD)

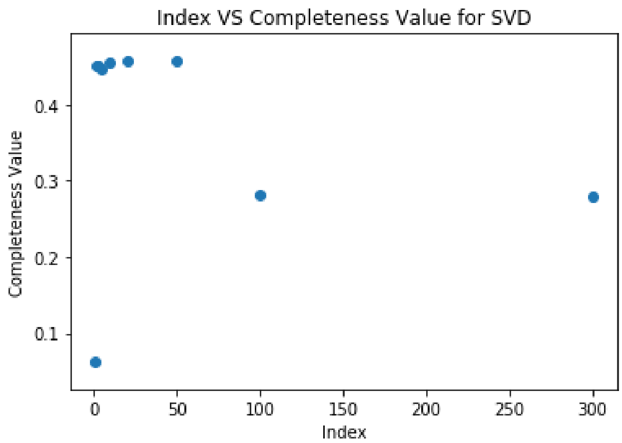

Figure 3: Completeness VS r (SVD)

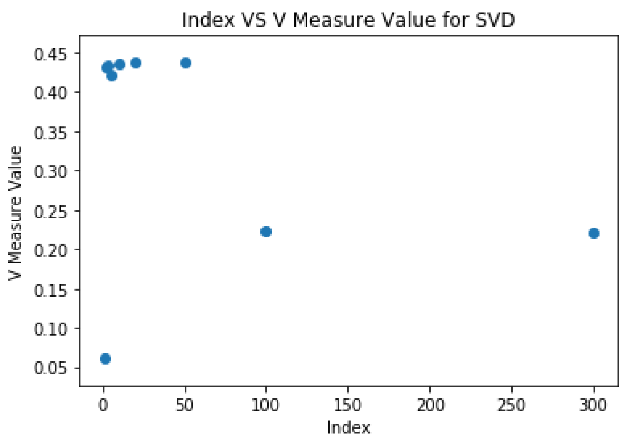

Figure 4: V Measure VS r (SVD)

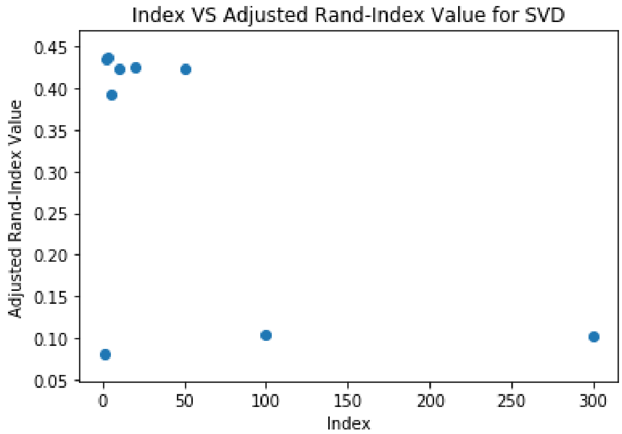

Figure 5: ARI VS r (SVD)

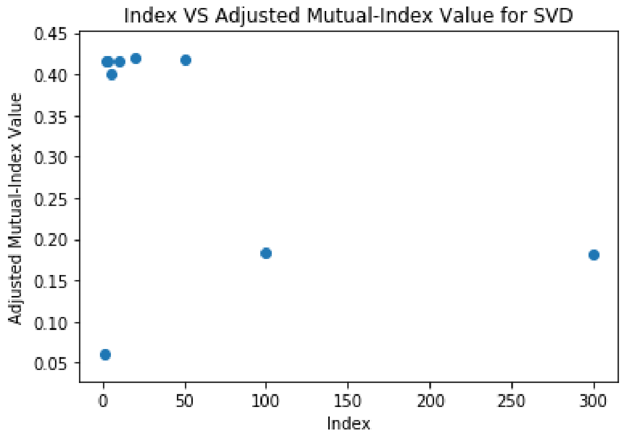

Figure 6: AMI VS r (SVD)

Index	Homogeneity	Completeness	V Measure	ARI	AMI
1	0.0600	0.0610	0.0605	0.0809	0.0599
2	0.3778	0.3930	0.3852	0.4314	03777
3	0.4633	0.4819	0.4724	0.5203	0.4632
5	0.1349	0.2471	0.1746	0.0668	0.1348
10	0.1179	0.2271	0.1552	0.0481	0.1178
20	0.1130	0.2290	0.1513	0.0420	0.1129
50	0.0552	0.1559	0.0815	0.0132	0.0551
100	0.0022	0.0966	0.0044	-6.7533	0.0021
300	0.0194	0.1437	0.0342	0.0021	0.0193

Table 13: Stats for NMF against various index values

1684	2219
2844	1135

Table 14: Contingency Matrix for r=1 (NMF)

3701	202
1150	2829

Table 15: Contingency Matrix for r=2 (NMF)

984	2919
3865	114

Table 16: Contingency Matrix for r=3 (NMF)

2916	987
3974	5

Table 17: Contingency Matrix for r=5 (NMF)

3896	7
3068	911

Table 18: Contingency Matrix for r=10 (NMF)

3901	2
3130	849

Table 19: Contingency Matrix for r=20 (NMF)

3886	17
3469	510

Table 20: Contingency Matrix for r=50 (NMF)

0	3903
18	3961

Table 21: Contingency Matrix for r=100 (NMF)

3754	149
3979	0

Table 22: Contingency Matrix for r=200 (NMF)

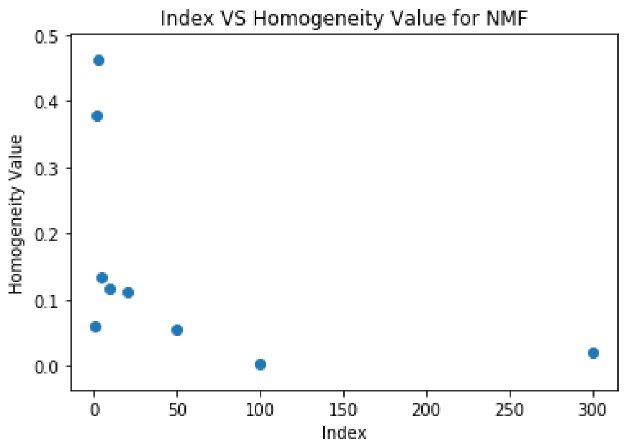

Figure 7: Homogeneity VS r (NMF)

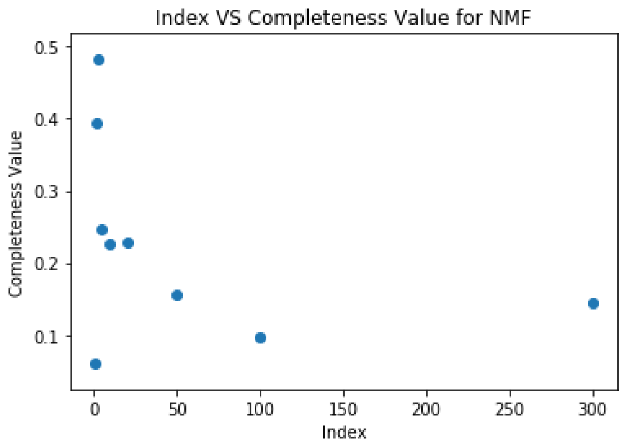

Figure 8: Completeness VS r (NMF)

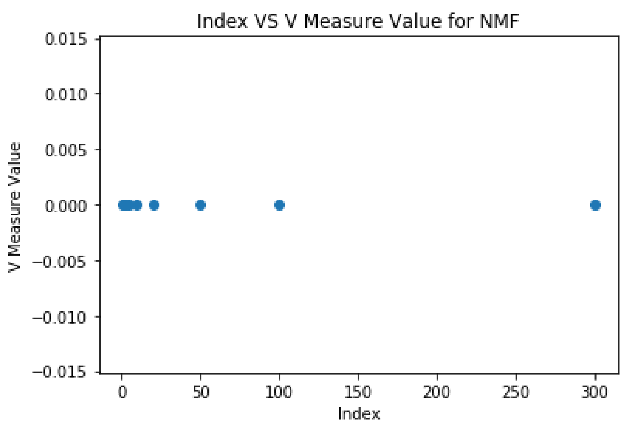

Figure 9: V Measure VS r (NMF)

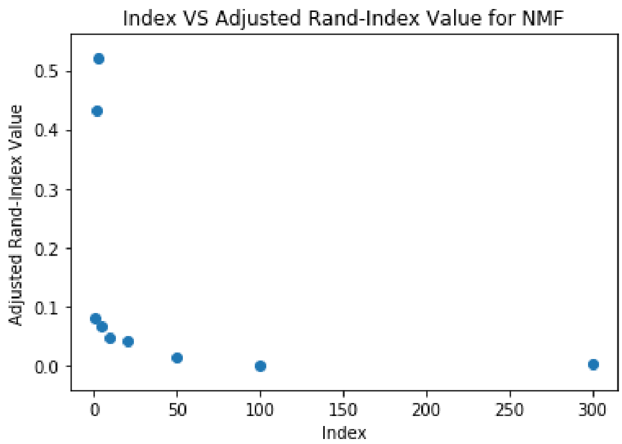

Figure 10: ARI VS r (NMF)

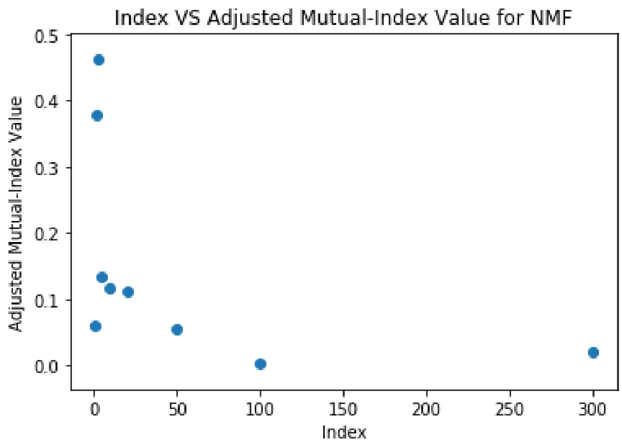

Figure 11: AMI VS r (NMF)

We conclude that the algorithm performs best for r=2 in SVD and r=3 in NMF. We also notice that the performance decays as we increase the value of r from its optimal point.

6 Question 4

6.1 Part a

In this section we represent the clustering in color coded form. The clustering was done from the best r that was obtained from the previous part.

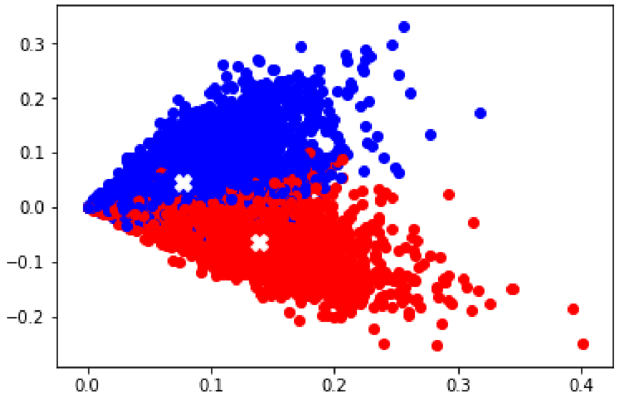

Figure 12: Color coded clustering result for SVD for r=2

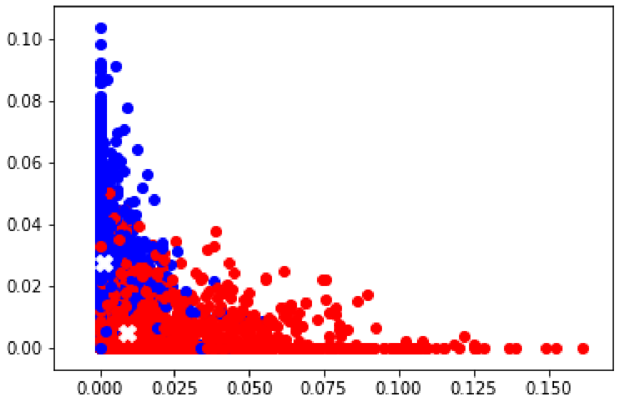

Figure 13: Color coded clustering result for NMF for r=3

6.2 Part b

K-means clustering is isotropic in all directions of space and therefore tends to produce more or less round (rather than elongated) clusters. Due to this, leaving variance that is unequal will invariably put more weight on variables with smaller values. Though normalization does not always improve result, it mostly does not degrade it either.

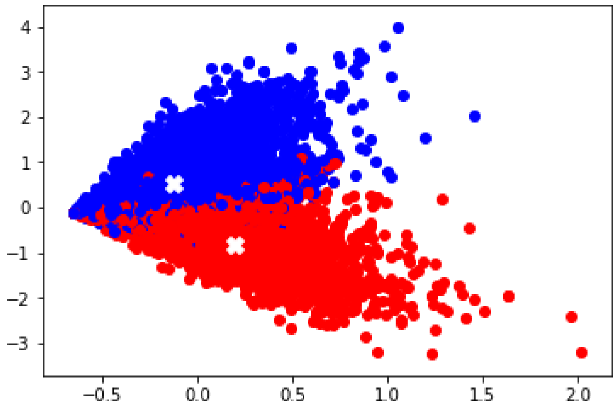

Figure 14: Clustering of Normalized Data (SVD)

- Homogeneity: 0.475905197552
- Completeness: 0.494126141691
- V-measure: 0.484844539684
- Adjusted Rand-Index: 0.533608766189
- Adjusted Mutual-Index: 0.475857215875
- Contingency Matrix: [[956 2947] [3873 106]]

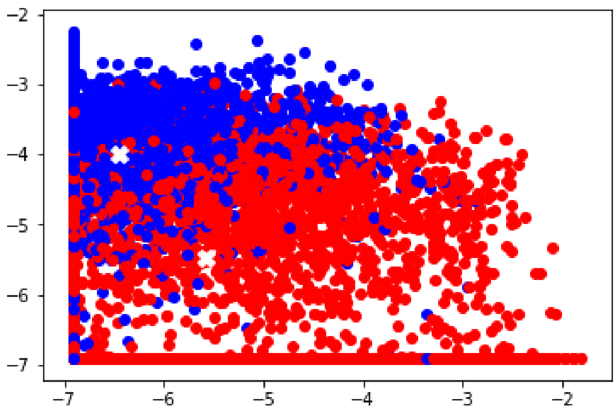

Figure 15: Clustering after logarithmic transformation (NMF)

- Homogeneity: 0.505891837486
- Completeness: 0.506595279058
- V-measure: 0.506243313908
- Adjusted Rand-Index: 0.61311822321

- Adjusted Mutual-Index: 0.505846601723
- Contingency Matrix: [[497 3406] [3621 358]]

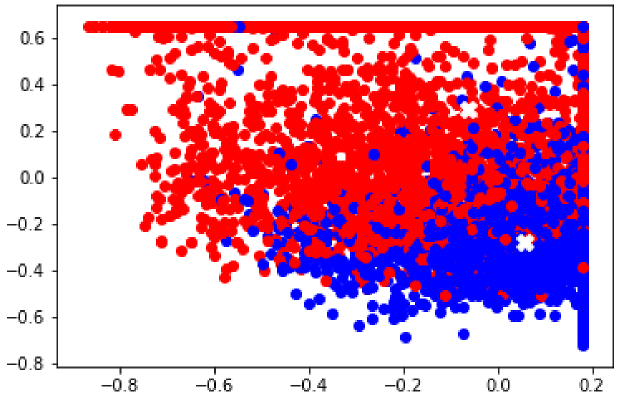

Figure 16: Clustering after performing logarithmic transformation then normalizing (NMF)

- Homogeneity: 0.466101019428
- Completeness: 0.466366307243
- V-measure: 0.466233625598
- Adjusted Rand-Index: 0.572481255597
- Adjusted Mutual-Index: 0.466052140817
- Contingency Matrix: [[3384 519] [440 3539]]

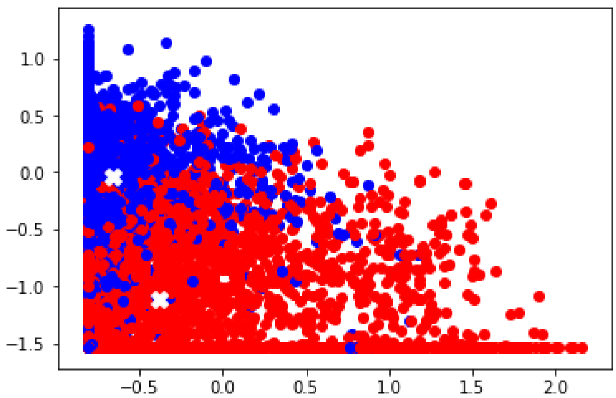

Figure 17: Clustering after normalizing then performing logarithmic transformation then normalizing (NMF)

- Homogeneity: 0.46486810155
- Completeness: 0.465035993491
- V-measure: 0.464952032365
- Adjusted Rand-Index: 0.571329701944
- Adjusted Mutual-Index: 0.464819110069
- Contingency Matrix: [[3393 510] [452 3527]]

We notice that normalizing the data does not have a huge effect on the data in SVD. However, there is a noticeable difference after performing a logarithmic transformation. When the distribution of the continuous data is non-normal, transformations of data are applied to make the data as normal as possible and, thus, increase the validity of the associated statistical analyses. The log transformation is, the most popular among the different types of transformations used to transform skewed data to approximately conform to normality. Here, the log transformation is to spread out the data and that in turn gives better clustering performance. Since the log transformation can only be used for positive outcomes, it is common to add a small positive constant, M, to all observations before applying this transformation. We have added a value of "0.718657278385" to all observations.

7 Question 5

In this part, we work with 20 sub classes instead of the eight as we did with the previous parts.

Dimensions of TF-IDF Matrix : 18846, 35684

Index	Homogeneity	Completeness	V Measure	ARI	AMI
1	0.0148	0.0161	0.0154	0.0029	0.0116
2	0.1730	0.1847	0.1786	0.0506	0.1703
3	0.2116	0.2308	0.2208	0.0679	0.2091
5	0.2666	0.2989	0.2818	0.0881	0.2642
10	0.2753	0.3089	0.2911	0.0887	0.2730
20	0.2713	0.3382	0.3011	0.0681	0.2689
50	0.2687	0.3422	0.3010	0.0625	0.2663
100	0.2610	0.3541	0.3005	0.0591	0.2586
300	0.2489	0.3732	0.2986	0.0603	0.2464

Table 23: Stats for SVD against various index values

```
Contingency Matrix:  [[ 61  61  11  21  35  50  45  36  57  48  25   0   9  23  81  52  56  22
   72  34]
 [ 72  70   3  12  28  84  46  47 109  66  23   0   5  21  91  53 106  12
   97  28]
 [ 93  57   1  24  53  68  74  37 100  83  39   0   5  20  73  34  97  13
   60  54]
 [ 85  60   0  24  21  83  80  24  89  88  27   0   4  16  93  31 108  11
   80  58]
 [ 87  74   3  10  40  85  59  42  89  60  38   0   2  23  87  60  92   6
   68  38]
 [ 85  81   5  10  22 106  34  51 106  47  15   0   5  27 112  62  83   6
  109  22]
 [ 56 106   1   5  21 108  29  62  90  31  11   0   0  46 101  83  81   4
  122  18]
 [ 78  96   1   8  68  94  31  54  87  53  13   0   2  46 101  73  71   5
   87  22]
 [ 51  97   0   8  46 101  31  78  79  39  17   0   0  59  92 102  68   3
  110  15]
 [ 70  93   0  12  66  97  30  77  70  44  15   0   1  49  91  88  56  10
   97  28]
 [ 62 106   0   9  43 111  30  80  79  42  22   0   2  57  77  85  62   7
  102  23]
 [ 75  52   5  35  33  69  72  33  83  81  53   3  14  32  68  47  82  25
   65  64]
 [ 64  90   0   6  35 113  34  48  90  45  17   0   4  34 111  65  89   6
  111  22]
 [ 53  92   1  11  40  97  36  69  93  50  12   0   2  45 102  90  70   6
  102  19]
 [ 57 102   1  11  42  89  39  70  83  49  20   0   1  41 100  89  86   9
   77  21]
 [ 79  45  11  43  28  64  67  24  94  70  58   1  31  23  80  39  92  37
   52  59]
 [ 69  72   3  31  26  75  43  63  76  51  35   0  16  30  65  57  69  16
   66  47]
 [ 61  83   6  14  33  83  44  56  80  50  17   3   6  34 106  61  85  10
   78  30]
 [ 49  60   5  23  30  78  32  40  77  68  17   2   9  17  51  46  52  18
   69  32]
 [ 41  53   5  19  30  48  31  35  39  39  24   1  10  33  49  42  42  18
   44  25]]
```

Figure 18: Contingency matrix for r=1

```
Contingency Matrix:  [[ 79   8   0  58   1 102  54   3  17   2  53 152  26   3   0  20   0   5
  151  65]
 [ 12 108 158   4 127  31   0  22  89  63  44   7   0 172  74   0  19  43
    0   0]
 [  8  75 176   1  98  23   0   9  43  68  69   5   0 113 204   0  66  25
    1   1]
 [  4  71 198   2 134  10   0  10  52  95  36   2   0 151 151   0  47  19
    0   0]
 [  3 120 134   2 145  22   0  20  74 101  60   2   0 171  53   0   8  48
    0   0]
 [  5 168 148   2 145  12   0   6  85  52  40   4   0 226  42   0  19  33
    0   1]
 [  5 179 110   4 109  44   0   5 114  29  61   6   0 190  76   0   9  32
    0   2]
 [ 93  86   4  37  32 145   2  31 209   5 116  48   0  23   0   1   0 140
    9   9]
 [124  72   1  49  15 207   0  29 163   2 107 110   0  11   0   0   0  91
    6   9]
 [164  33   6  95   6 210   0  17  69   2 114 158   0  17   2   8   0  39
   27  27]
 [182  32   2 112   5 225   0   6  57   1 104 188   0   8   0   6   0  13
   16  42]
 [107  43   6  88  42  84   1 100 112  20  65  75   0  18   6  24   3 136
   19  42]
 [ 33 113  35   5 112  83   0  22 246  44  69  18   0  94   7   0   0 102
    0   1]
 [147  40   2  84  16 216   4  15 155   2  86 104   0  21   0   3   2  59
   10  24]
 [123  73   7  61  28 197   1  28 157  11  83  57   0  22   0   5   1 101
   11  21]
 [ 69   3   1  45   5  61 166   4  23   2  47 138  74   4   0  32   1  12
  235  75]
 [123  18   0  97   2 143   6   7  40   1  52 155   5   1   0  33   0  16
  104 107]
 [ 92  10   0  49   0 126  18   0  18   1  63 236   7   0   0  20   0   7
  219  74]
 [113   3   3  89   0 106   4   5  25   0  46 178   3   2   1  32   0  10
   83  72]
 [ 49   9   0  46   0  83  56   1  11   1  57 113  32   1   0  14   0   7
  104  44]]
```

Figure 19: Contingency matrix for r=2

```
Contingency Matrix:  [[ 18   5   0  82  32 169   0  94  17 113   0   3  13  14  72   0  53   3
    3 108]
 [  0 221 198   1   0  14  83   0  31  54  56  23  76   0  74   0   0 139
    2   1]
 [  0 152 249   2   0   8  92   0  14  30 212   7  44   0  88   0   0  82
    1   4]
 [  0 179 202   1   0   4 187   0  15  24  65  13  68   0  45   0   0 178
    1   0]
 [  1 221 100   2   0   3 123   0  31  49  10  26 102   0  83   0   0 208
    3   1]
 [  0 321 227   3   0   5  56   0  12  33  43   5  68   0  74   0   0 138
    1   2]
 [  0 177  73   1   0   9  77   0  30 108  19  14 252   0  80   3   0 112
   19   1]
 [  7  37   1  18   0  66   1   1 188 212   0  39 161   0 126  14   0  29
   86   4]
 [  2  34   2  30   0 125   0   2 148 272   0  28 121   0 116   9   0  15
   85   7]
 [ 15  24   7  29   0  75   0   1  69 222   0  11  31   0 107 140   1   7
  252   3]
 [ 13   8   2  20   0  38   1   0  31 174   0   7  23   0  86 217   0   5
  370   4]
 [ 45  26  11  81   0 108   6   2 187 112   8 114 112   0  81  23   0  43
   29   3]
 [  0 134  32   6   0  34  27   0  92 136   1  30 265   0  88   0   0 123
   13   3]
 [  8  48   4  73   1 197   3   5 105 233   1  20 111   1 103   3   2  16
   49   7]
 [ 13  39   8  48   0 125   5   0 151 251   0  36 137   0  98   4   0  28
   34  10]
 [  9  11   3  27  76  73   0 137   8  58   2   3  11  71  61   0 177   9
    0 261]
 [ 87   6   0 207   0 207   0   7  57 170   0  16  23   1  69  18   2   1
   25  14]
 [ 49   4   0 253   1 305   1  14  18 167   0   1  10   2  67   4   2   0
   23  19]
 [ 63   2   3 168   0 212   1  14  47 132   0   8  18   0  51   6   1   0
   28  21]
 [ 12   5   1  69  22 100   1  59  12  95   0   1   5  30  71   0  49   0
    2  94]]
```

Figure 20: Contingency matrix for r=3

```
Contingency Matrix:  [[110  37  52   2   0   2 189  48 194  65   0   1  12  19   2   0   0   1
   65   0]
 [ 91   0  26 246   0   2   0   6  72   1   1  27 263   0 111  68   0   1
    0  58]
 [109   0  13 268   4   0   2   2  24   6  15  33 153   0  61 211   0   0
    1  83]
 [ 66   0  17  53  89   1   1   1  30   4 162 193  93   0  65  15   2   0
    0 190]
 [117   0  21  28  28   1   1   3  49   3 114 266 100   0  92   4   0   0
    0 136]
 [ 90   0  11 315   0   0   0   1  27   3   0   9 375   0  79  67   0   1
    0  10]
 [139   0  31  14  38  30   2   2  67   1  83 394  62   0  32   5   8   0
    0  67]
 [203   0 265   8   0  16   6  13 290  19   4  86  47   0  31   0   1   0
    0   1]
 [217   0 238   2   0  12   7   9 372  16  12  53  34   0  22   0   0   0
    2   0]
 [190   0  95   9   0 327   5   6 150   3   0   3  30   0   8   1 165   1
    1   0]
 [133   0  44   4   0 424   4   6  89   1   0   3  15   0   1   1 274   0
    0   0]
 [113   0  16  15   0   1   3  46 126 251   0  12  36   0  64   9   0 299
    0   0]
 [153   0  68  30   1   4   3   0 250  17  17 125 126   0 160   1   0   3
    1  25]
 [188   1 154  10   0   1  14  27 400  54   0   2  89   1  44   2   0   0
    2   1]
 [189   0 144  20   0   4  11  37 354  76   0  18  63   0  70   1   0   0
    0   0]
 [ 90 104  21   7   0   0 340  21  96  13   0   1  15  84   9   2   0   0
  194   0]
 [127   0  71   0   0   2  23 152 229 280   0   3  12   1   6   0   0   2
    2   0]
 [137   5  25   1   0   0  37 106 238 375   0   0   6   2   4   0   0   2
    2   0]
 [104   0  64   4   0   3  30 113 233 210   0   2   5   0   6   0   0   0
    1   0]
 [113  33  23   2   0   2 137  35 130  52   0   0   6  32   4   0   0   0
   59   0]]
```

Figure 21: Contingency matrix for r=5

```
Contingency Matrix:  [[  3   0   5 126   0   2   2  56   0 173   0   0 191   4   1 137  85   1
   13   0]
 [  7 211 142   0   0   2   1   0   0   9  27   0 129 117  23  70   8   0
  185  42]
 [  5 292  61   1  10   0   0   0   0   7  38   0 140  57   4  45   4   0
   57 264]
 [ 11  96 147   0 142   2   3   0   1   1 265   0  77  57  14  42   1   0
   95  28]
 [ 17  38 246   0  53   1   1   0   0   0 223   0 130  65   6  64   4   0
  111   4]
 [  2 361 128   0   0   0   8   0   0   3   5   0 153  60  11  60   4   0
  143  50]
 [ 37  19 244   1  48  13   0   0   6   5  92   0 108  19 269  14   4   0
   89   7]
 [326   3 113   3   2   0   1   0   0  24   3   0 235  20  12 187   9   2
   50   0]
 [190   1  99   6   1   2   0   0   1  22  18   0 281   9   5 307  12   0
   42   0]
 [  4   1   9   3   0 368   1   0 114  15   0   0 283  26  13 112  12   0
   33   0]
 [  2   0   2   0   0 426   0   0 241   8   0   0 184  12  36  52  14   0
   22   0]
 [  1  17 118   1   0   1 357   0   0 136   3   0 156   8   3  76  65   4
   41   4]
 [ 49  21 360   1   2   5   6   0   0  11  36   0 163  49  12 132   4   0
  131   2]
 [  8   1  66   4   0   0   0   3   0  76   0  71 295  23  12 296  35   1
   98   1]
 [ 14   6 178   1   0   4   0   0   0  75   0   0 300  11  52 237  46   1
   62   0]
 [  0   0  11 394   0   0   0 168   0 127   1   0 145   8   1  60  40   4
   37   1]
 [  4   0  19   8   0   2   0   1   0 332   0   0 201   1   4 109 172  40
   17   0]
 [  0   0   2   9   0   1   0   2   0 326   0   0 167   1   4  41  59 315
   13   0]
 [  3   1  15   4   0   5   0   0   0 303   1   0 165   6   5 104 135  20
    8   0]
 [  1   1   8 118   0   3   0  57   0 116   0   2 170   2   0  77  59   5
    9   0]]
```

Figure 22: Contingency matrix for r=10

```
Contingency Matrix:  [[  3 285   0   2 164   6   1  86 172  61   0   2   0   0   0   2   0   0
   14   1]
 [  0 199   0 201   0 144  56   1  88   0   0   2   0  57   0   0   4  29
  192   0]
 [  0 194   2 131   1  60  66   2  70   0  15   1   0  77   0   0   1 293
   72   0]
 [  0 171  82  17   0 103  19   0  62   0 127   5   0 197   0   4   1  43
  149   2]
 [  0 235  22   7   0 111  31   0 101   0 101   1   0 121   0   0   1   5
  226   1]
 [  0 202   0  64   0  94 415   1  49   0   1   7   0   9   0   0   3  44
   99   0]
 [  0 181   4   2   1  45   1   1  26   0  74   1   0  51   0  44  71  16
  439  18]
 [  0 371   0   1   3  35   3  20 119   0   3   1   0   0   0 397   2   0
   35   0]
 [  0 466   0   0   5  30   2   8 216   0  14   0   0   0   0 214   0   0
   39   2]
 [  0 420   0   1   2  39   0   3 127   0   0   1   0   0   0   1   5   0
   21 374]
 [  0 297   0   0   0  19   0   1  82   0   0   0   0   0   0   1  18   0
   18 563]
 [  0 274   0  15   1  19   3 134 121   0   1 329   0   3   0   1   2   5
   83   0]
 [  0 406   2  13   1 104   8   5 146   0  14   6   0  13   0  29   1   3
  231   2]
 [  0 457   0   1   5  62   0  11 296   3   0   0  71   0   1   3   0   1
   79   0]
 [  0 401   0  11   0  22   1  24 268   0   0   0   0   0   0   2   1   0
  256   1]
 [  2 204   0   1 454  24   0  17  97 177   0   0   0   0   0   1   0   1
   19   0]
 [  2 307   0   3   7   6   1 452 109   1   0   0   0   1   0   3   3   0
   14   1]
 [279 295   0   0  12   4   0  67 105   3   0   0   0   0 165   0   2   0
    8   0]
 [  1 311   0   0  10   6   0 255 165   0   1   0   0   0   0   2   0   0
   23   1]
 [  2 235   0   1 139   3   0  68 107  57   0   0   2   0   0   2   0   0
   10   2]]
```

Figure 23: Contingency matrix for r=20

```
Contingency Matrix:  [[ 22   1  41   1 294   2   2   0   0   0   8   0   3   0   0 104   0   6
  313   2]
 [ 55   0  44   0 281   2 256   0   0  54 142   3   0  29  86   0   0   0
   21   0]
 [ 32   0  30   0 227   1 147   0   0  77  64   1   0 304  65   1  16   0
   20   0]
 [ 41   2  27   0 183   3  27   0   0 160  81   1   0  39 228   0 183   0
    5   2]
 [ 60   1  34   0 221   1  16   0   0  62  78   1   0   5 381   0  98   0
    5   0]
 [ 68   0  31   1 278   8 418   0   0   7 100   4   0  48  18   0   1   0
    6   0]
 [  9  15  21   0 228   0   0   0   0  31  52  63   0  12 455   1  51   0
   10  27]
 [ 35   0 112   1 431   1   3   0   0   0  30   2   0   0  14   0   3   1
   35 322]
 [ 33   2 406   0 442   0   0   0   0   0  30   0   0   0  11   3  12   0
   32  25]
 [ 29  88  58 372 374   1   2   0   0   0  35   3   0   0   0   1   0   1
   29   1]
 [ 15 221  35 422 256   0   0   0   0   0  16   8   0   0   2   0   0   0
   23   1]
 [ 29   0  39   0 350 343  21   0   0   1  18   1   0   5  31   0   0   8
  144   1]
 [ 62   2  74   0 504   7  21   0   0   8  75   1   0   3 180   1  13   0
   17  16]
 [ 58   0  64   1 578   0   3   0  71   0  50   0   0   1   6   4   0   1
  153   0]
 [ 28   0  75   4 699   0  15   0   0   0  29   1   0   1  19   0   0   1
  114   1]
 [ 27   0  21   1 286   0   0   0   0   0  21   0   2   1   0 335   0   4
  299   0]
 [ 12   1  37   1 335   1   1   0   0   1   6   3   2   0   0   3   0 206
  297   4]
 [ 19   0  32   0 288   0   0 162   0   0   3   2 273   0   0   6   0   6
  149   0]
 [ 10   1  35   2 328   0   0   0   0   0   6   0   1   0   0   1   1  62
  325   3]
 [ 17   2  27   0 258   0   1   0   2   0   5   0   3   0   0  99   0  48
  164   2]]
```

Figure 24: Contingency matrix for r=50

```
Contingency Matrix:  [[  0   0 342   8  95 209   0  23   0   0   7  24  82   0   1   4   2   0
    2   0]
 [ 59  29 366 168   0  20   0   0   0   1  54  49   0   0   0  11 189  24
    3   0]
 [ 71 296 270  71   0  14   0   0   3   1  37  39   1  16   0   8 128  29
    1   0]
 [183  44 222 123   0   3   0   0  73   1 150  31   0 127   3   1  12   5
    4   0]
 [111   4 295 122   0   6   0   0  32   1 254  33   0  89   1   5   8   1
    1   0]
 [ 10  51 313 115   0   6   0   0   0   2  17  33   0   1   0   6  52 373
    9   0]
 [ 55  18 252  61   0   8   0   0  11  51 335  93   0  66  21   1   2   1
    0   0]
 [  0   1 386  41   0  43   0   0   0   2 484  24   0   3   0   0   1   4
    1   0]
 [  0   0 489  31   0  46   0   0   0   0 378  27   2  17   2   1   0   3
    0   0]
 [  0   0 486  41   0  38   0   0   0   1  12  34   1   0 378   0   2   0
    1   0]
 [  0   0 311  17   0  30   0   0   0   6   5  27   0   0 603   0   0   0
    0   0]
 [  2   5 384  17   0 165   0   0   0   1  29  26   0   1   0   3  13   1
  344   0]
 [ 15   4 518 112   0  17   0   0   2   0 231  44   1  13   2   8   8   2
    7   0]
 [  0   1 672  64   0 123   0   0   0   0  20  34   4   0   0   1   0   0
    0  71]
 [  1   2 491  21   1  57   0   0   0   1  29  33   0   0   2 344   5   0
    0   0]
 [  0   1 259  22 442  77   0   0   0   0   2  10 181   0   0   2   1   0
    0   0]
 [  1   0 355   5   6 496   0   0   0   3  10  21   3   0   2   4   3   1
    0   0]
 [  0   0 326   4   6 412 161   0   0   1   0  25   5   0   0   0   0   0
    0   0]
 [  0   1 371   7   1 366   0   0   0   0   8  15   0   2   2   2   0   0
    0   0]
 [  0   0 295   6 133 130   0   0   0   0   5   5  51   0   0   1   1   0
    0   1]]
```

Figure 25: Contingency matrix for r=100

```
Contingency Matrix:  [[  0   3   2   0  13   2   0 321 115   0   0  59   0 239   7   5  33   0
    0   0]
 [  4   8   3   0  14   2   2 292   0   0   1   0   0  13 376   1  17  44
  195   1]
 [  1   7   1   0  17   0  18 263   1   0   0   2   5   8 246   0  20 380
   16   0]
 [  1   1   4   0  20   1 222 236   0   0   1   0   0   6 386   1  30  51
   22   0]
 [  1   3   1   0  28   0 117 325   0   0   1   0   0   5 458   0  12   5
    7   0]
 [  5   3  12   0  16   2   1 268   0   0   1   0   0   5 568   0  19  59
   27   2]
 [ 67   0   0   0   5   5  78 563   1   0  22   0   0   8 164   0  32  19
   11   0]
 [  3   0   1   0  11  13   3 842   0   0   0   0   0  48  39   0  21   2
    0   7]
 [  0   1   0   0  14   9  15 859   3   0   2   0   0  43  22   0  27   1
    0   0]
 [  2   0   1  80  13 414   0 329   1   0  75   1   0  26  32   0  20   0
    0   0]
 [  7   1   0 383  12  55   0 291   0   0 211   0   0  24  14   0   1   0
    0   0]
 [  2   3 363   1  14   2   0 354   0   0   0   0   0 150  53   0  41   7
    1   0]
 [  1   3   6   0  17   0  16 673   1   0   2   0   0  11 223   0  23   4
    3   1]
 [  0   0   0   0  66   5   0 677   4  71   0   0   0  96  59   1  10   1
    0   0]
 [  1 178   0   0   6   6   0 420   0   0   1   1  22  48  23 268  12   0
    0   1]
 [  0   2   0   0   8   2   0 331 391   0   0   2   0 224  26   2   8   1
    0   0]
 [  3   0   1   1  16   6   0 334   3   0   1   0   1 485   5   0  52   0
    0   2]
 [  2   0   0   1   8   5   0 295   6   0   0   2   0 613   3   0   5   0
    0   0]
 [  0   0   0   0   9   6   1 332   1   0   1   3   0 395   7   0  14   1
    0   5]
 [  0   0   0   0  13   0   0 279 110   2   2  37   0 168   6   3   8   0
    0   0]]
```

Figure 26: Contingency matrix for r=300

We observe the best results for r=10

Index	Homogeneity	Completeness	V Measure	ARI	AMI
1	0.0152	0.0164	0.0158	0.0031	0.0120
2	0.1637	0.1768	0.1700	0.0485	0.1610
3	0.1748	0.1910	0.1825	0.0503	0.1721
5	0.2413	0.2759	0.2574	0.0724	0.2388
10	0.2703	0.3212	0.2936	0.0757	0.2680
20	0.2256	0.2775	0.2489	0.0484	0.2230

Table 24: Stats for NMF against various index values

```
Contingency Matrix:  [[ 51  35  41  51  22  65  36  47  11  24  46  61  60  66  25   0   9  28
   82  39]
 [ 83  28  51  99  12  76  29  46   3  12 102  65  62 100  26   0   5  28
   81  65]
 [ 66  57  41  93  13  62  54  74   1  29  88  87  35  65  24   0   5  40
   73  78]
 [ 87  59  30  90  11  70  22  79   0  26  89  80  37  72  15   0   4  32
   93  86]
 [ 76  41  55  83   6  64  40  56   3  12  83  83  67  83  26   0   2  40
   86  57]
 [109  26  60  75   6  77  22  32   5  10  95  80  73 119  32   0   5  15
   99  48]
 [102  19  72  69   4 120  22  29   1   5  80  53  92 111  50   0   0  11
  102  33]
 [ 97  23  66  66   5 102  68  28   1   8  80  77  82  83  49   0   2  16
   85  52]
 [ 91  17  91  61   3  97  48  31   0   8  69  48 109 113  66   0   0  17
   88  39]
 [ 94  29  89  51  10  97  67  30   0  14  63  66 101  84  56   0   1  15
   84  43]
 [106  19  76  55   8 104  45  32   0   9  69  60 106 104  67   0   2  26
   70  41]
 [ 70  64  38  76  28  59  34  69   5  35  74  65  48  62  38   3  14  58
   69  82]
 [113  25  57  79   6  89  35  31   0   9  83  61  77 120  35   0   4  15
  101  44]
 [105  23  87  68   6  90  40  37   1  13  78  51  92 102  49   0   2  10
   90  46]
 [ 97  20  84  79   9  89  43  43   1  11  71  54 107  79  42   0   1  23
   89  45]
 [ 59  63  32  82  43  50  29  61  11  41  90  74  39  63  23   1  31  60
   77  68]
 [ 72  47  75  67  19  78  26  41   3  31  74  65  50  66  39   0  16  38
   55  48]
 [ 87  31  64  77   9  77  34  44   6  14  78  57  70  92  37   3   7  19
   85  49]
 [ 75  34  47  44  18  64  31  33   6  27  68  45  51  65  19   1   9  18
   57  63]
 [ 45  24  41  41  18  50  30  36   5  22  36  37  46  50  38   1  11  24
   40  33]]
```

Figure 27: Contingency matrix for r=1

```
Contingency Matrix:  [[182   0   1  61  97  35   2   0 172  39   8   1  12   0 143  37   2   5
    2   0]
 [  6 126  81  47   0  49 158  17   1   0  90 143  14   9  19  22  21  69
   27  74]
 [  6 164  64  72   1  16 101  93   2   0  67 122   2   8  14  12  13  46
   26 156]
 [  1 163  82  37   0  11 123  62   1   0  66 186   2   4   6   8  13  53
   45 119]
 [  3 103  95  60   0  41 162  13   0   0  97 147   5   2  10  11  25  71
   59  59]
 [  3  93  77  39   1  21 204  15   1   0 137 206   2   8   7  10  17  77
   23  47]
 [  3  76  63  68   0  64 195  19   0   0 154 122   5   1  15  17  12  98
   13  50]
 [ 83   4  20 132  11 173  30   0  39   1  54   4  29   0 126 138  46  96
    4   0]
 [139   1   9 125  10 152  16   0  53   0  60   3  26   0 192 125  29  54
    2   0]
 [203   3   3 143  45  75  16   0 130   3  24   6  19   0 224  71   6  19
    3   1]
 [253   0   3 118  52  72   8   0 165   3  26   4  14   0 234  31   3  12
    0   1]
 [ 96   5  40  75  43  72  18   4  74  15  39  14  68   0  86 140  96  78
   20   8]
 [ 25  32  77  79   1 145 102   0   5   0 106  40   3   0  41  60  41 182
   35  10]
 [140   2  10 113  21 124  22   1  88   2  38   9  25   1 195 108  22  67
    2   0]
 [ 94   6  20 111  16 167  25   1  47   5  46  12  40   0 153 133  28  74
    7   2]
 [174   0   5  57 202  26   2   1 241  92   5   5  19   0  98  46   5  17
    2   0]
 [174   0   1  71 114  52   1   0 191  29  16   0  22   0 170  57   3   8
    1   0]
 [271   1   0  74  86  30   1   0 224  22   8   0   9   0 187  21   2   4
    0   0]
 [202   3   0  62  78  37   2   0 164  24   3   0  18   0 132  37   6   6
    0   1]
 [117   0   0  68  82  28   2   0 127  41   7   1   7   0 111  31   3   2
    1   0]]
```

Figure 28: Contingency matrix for r=2

```
Contingency Matrix:  [[ 97   0  53  13   5  61   9   7   1   2  16  68   0  86 126 106  30  51
   65   3]
 [ 45 158  12   0 165   0 120  42 208  62   3  55  43  12   0  42   0   0
    6   0]
 [ 28 250   5   0 114   1  81  21 166  62   1  80 152   8   1   9   0   0
    6   0]
 [ 20 221   3   0 122   0 100  28 233  94   1  41  95   4   0  15   0   0
    5   0]
 [ 41 128   3   0 179   0 129  54 194 103   2  74  20   9   0  24   0   0
    3   0]
 [ 30 109   3   0 242   0 133  25 280  58   0  55  29   5   0  16   0   0
    3   0]
 [108  95   3   0 178   1 182  41 160  34   4  75  29  10   3  46   0   0
    6   0]
 [161   1  49   0  42   1  57 100   9   7  21 103   0 145   5 178   0   0
  110   1]
 [160   1  52   0  38   4  46  54   2   6  17 104   0 176   6 209   0   0
  121   0]
 [164   3 101   0  20   1  17  20  13   3  47  91   0 184   8 178   0   0
  141   3]
 [139   0 121   0  15   1  16   9   7   1  50  76   0 195   3 189   0   0
  172   5]
 [ 78   6 146   0  30   0  62 115  15  22  67  65   5 121   5 112   0   0
  123  19]
 [114  29  12   0 124   1 195  98  66  55   0  75   0  59   2 126   0   0
   28   0]
 [186   2  54   1  40   1  67  41  13   3  15  76   2 165   9 191   0   2
  120   2]
 [179   6  66   0  31   2  70  73  17  11  23  82   1 141   7 174   0   0
  101   3]
 [ 60   1  22  68  12 213  17   7   6   1   4  68   1  26 202  39  85 149
   15   1]
 [105   0 132   0   8   5   7  21   0   2  90  56   0 151  16 143   0   1
  151  22]
 [122   1 115   2   6   5   4   7   0   0  41  61   0 200  25 148   4   1
  186  12]
 [ 91   3  99   0   2   2   4  10   1   0  55  42   0 149  19 133   0   1
  144  20]
 [ 89   0  30  30   8  66   1   5   1   1  14  66   0  55  87  64  28  37
   44   2]]
```

Figure 29: Contingency matrix for r=3

```
Contingency Matrix:  [[  2 225   0   2  15  40   0   0   8 119   1  25   0 137   1   8 149   0
    0  67]
 [ 25  33   0 241   0   5   0   0 250   0   1  86  73  24  74  19 106   2
   34   0]
 [ 26  11   0 258   0   1   0   4 143   1   0  60 203  14  69   8 114  17
   56   0]
 [200  21   1  54   0   1   0  89 102   0   3  55  16   9  46   6  74 165
  140   0]
 [274  38   0  30   0   3   0  27  96   0   1  75   4  16  36  10 146 117
   90   0]
 [  6  11   0 276   0   0   0   0 353   0   0  85  70   6  77   9  89   0
    6   0]
 [259  32   4  18   0   3   8  37  85   3  42  54   3  12   6  95 179  85
   50   0]
 [ 58 311   0   6   0  23   0   0  40   3   1  92   0 157  11  34 251   3
    0   0]
 [ 29 378   0   2   0  21   1   0  23   7   1  71   0 152  12  26 262  11
    0   0]
 [  1  52  40   8   0   5 137   0  22   2 236  11   3  25   6 286 159   0
    0   1]
 [  0  32 113   3   0   3 240   0   8   0 277   7   1  16   2 204  93   0
    0   0]
 [ 13 218   0  14   0 130   0   0  32   3   2  91   9 327  29   5 118   0
    0   0]
 [ 90 177   0  21   0   0   0   1 105   1   3 214   4  62  54  26 198  15
   13   0]
 [  0 348   0  13   1  20   0   0  57   3   0 104   2 184   8  15 232   0
    0   3]
 [  5 292   0  16   0  24   0   0  44   5   3 122   2 175  17  38 244   0
    0   0]
 [  1 104   0   6  74  13   0   0  13 360   0  24   2  43   3  10 129   0
    0 215]
 [  2 291   0   0   1 139   0   0  10   9   2  24   0 277   2  14 138   0
    0   1]
 [  0 338   0   1   2  80   0   0   5  11   0  20   0 326   0  11 144   0
    0   2]
 [  3 273   0   3   0  87   0   0   3   4   1  23   1 243   1  15 117   0
    0   1]
 [  0 150   0   2  30  26   0   0   8 108   3  10   0  89   1   4 140   0
    0  57]]
```

Figure 30: Contingency matrix for r=5

```
Contingency Matrix:  [[  2  10   6   0   1  51   0   0 126 223   0   0 164   0   0  53   2   2
  159   0]
 [  3 198  77   0   0   0  43   0   0 266   0 240  17   0   5   1  37   8
   78   0]
 [  1  79  21   0   0   0 256   0   1 187  17 328  12   2   2   0  34   1
   44   0]
 [  4 110  78   1   0   0  23   0   0 156 145 115   2  74   6   0 221   2
   44   1]
 [  1 127 100   1   0   0   4   0   0 257  95  41   6  15   3   0 234   6
   73   0]
 [  8 112  33   1   0   0  63   0   0 281   1 413   5   0   3   1  10   4
   53   0]
 [  0  40 424   9   0   0   7   0   1 127  52  14   7  16 185   0  50  21
   17   5]
 [  1  46 248   0   1   0   0   0   3 354   3   5  41   0  12   6  31   4
  235   0]
 [  0  30  96   2   0   0   0   0   7 403  10   2  48   0   1   1  21   7
  368   0]
 [  1  36  10 215   0   0   0   0   2 230   0   1  22   0   6   3   1 347
   63  57]
 [  0  13  16 296   0   0   0   0   0 133   0   0  10   0  20   0   0 298
   36 177]
 [332  22  13   0   9   0   4   0   1 258   0  21 155   0   2  46  11   1
  116   0]
 [  7 106 141   3   0   0   2   0   1 429  11  22  23   1   4   0  54  10
  170   0]
 [  0  71  38   0   1   3   1  71   3 409   0   0 100   0   3   2   0   3
  285   0]
 [  0  32 139   0   1   0   0   0   3 432   0   9  88   0  12  10   2  10
  249   0]
 [  0  27  13   0   0 161   1   0 383 186   0   1 120   0   0  24   1   0
   80   0]
 [  0   7  10   0  55   1   0   0   7 165   0   0 326   0   3 238   2   7
   89   0]
 [  0   4   4   0 201   2   0   0   7 136   0   0 227   0   1 318   0   2
   38   0]
 [  0   8  20   1  26   0   0   0   4 149   0   1 306   0   0 157   3   5
   95   0]
 [  0   5   6   1   5  55   0   2 108 191   0   1 116   0   0  47   0   4
   87   0]]
```

Figure 31: Contingency matrix for r=10

```
Contingency Matrix:  [[  0   1   0  10   2 282  75   0 142   1   0 100   0  53   1  37   3  75
   17   0]
 [ 34   5   0 160   0 395  29   1  91   0   0   6   7  82  24   0   0   0
   86  53]
 [310   6   0  75   0 288  25  22  65   0   0   7   2  45  31   0   0   1
   41  67]
 [ 35  17   3  98   2 239  11 191  67   0   0   2   7  86   6   0   0   0
   37 181]
 [  4  13   1  94   0 304  22 109  97   0   0   4   5 111   3   0   0   0
   72 124]
 [ 47  13   1 109   0 311  11   1  49   1   0   3   6  52 321   0   0   0
   52  11]
 [ 15   5  25  76  36 396   7  70  44   0   0   5 159  66   1   0   0   0
   14  56]
 [  0   3   0  40 314 346  22   3 117   0   0  21   6  93   3   0   0   0
   22   0]
 [  0   1   2  35  70 440  38  15 162   0   0  16   1 183   2   0   0   2
   29   0]
 [  0   0 322  37   1 347  42   0 120   1   0   6  10  88   0   0   0   1
   19   0]
 [  0   0 544  15   1 258  31   0  61   0   0   5  33  40   0   0   0   0
   11   0]
 [  6 317   0  21   0 221  32   0  79 178   0  59   2  57   1   0   0   0
   18   0]
 [  3  38   4  74  18 360  10  14 145   0   0   5   5 234   1   0   0   1
   62  10]
 [  1   2   0  56   0 330  34   0 146   0  71  35   2 267   0   0   0   4
   42   0]
 [  0   1   0  32   1 422  40   0 155   0   0  34   9 264   0   1   0   0
   28   0]
 [  1   0   0  22   0 244  35   0  83   0   0  49   0  37   0 357   2 148
   19   0]
 [  0   3   0   8   3 298  48   0 113   0   0 363   5  47   1   3   2   3
   12   1]
 [  0   2   0   5   0 283  31   0  91   0   0  73   3  22   0   1 407   4
   18   0]
 [  0   0   1   7   2 272  49   1 128   0   0 244   2  58   0   0   1   1
    9   0]
 [  0   1   2   5   2 227  36   0  77   0   1  80   0  37   0 104   1  45
   10   0]]
```

Figure 32: Contingency matrix for r=20

We observe the best results for r=10

Non Linear Logarithmic Transformation :

- Homogeneity: 0.308791294432
- Completeness: 0.312854824595
- V-measure: 0.310809778422
- Adjusted Rand-Index: 0.148654837272
- Adjusted Mutual-Index: 0.306560380024

Normalized SVD Data :

- Homogeneity: 0.279789283219
- Completeness: 0.326832036631
- V-measure: 0.301486605465
- Adjusted Rand-Index: 0.083093485568
- Adjusted Mutual-Index: 0.277448098245

Performing Logarithmic transformation first and then normalizing data:

- Homogeneity: 0.301911394335
- Completeness: 0.30585182322
- V-measure: 0.303868834905
- Adjusted Rand-Index: 0.146377258257
- Adjusted Mutual-Index: 0.299658294371

Normalizing the data first then performing logarithmic transformation :

- Homogeneity: 0.057676261021
- Completeness: 0.25036945454
- V-measure: 0.0937547466643
- Adjusted Rand-Index: 0.0309288799046
- Adjusted Mutual-Index: 0.057517290138