EE219 Project 2: Clustering

Report

Yi Jia 805033204

Shuojian Ye 904946811

Part 1

Build tf-idf matrix with min_df=3.

Dimension: (7882, 25535)

Part 2

Apply K-means clustering with k = 2 using the TF-IDF data.

(a) contingency matrix

[3, 3900],

[1684, 2295]])

(b) The five measures:

Homogeneity: 0.248

Completeness: 0.332

V-measure: 0.284

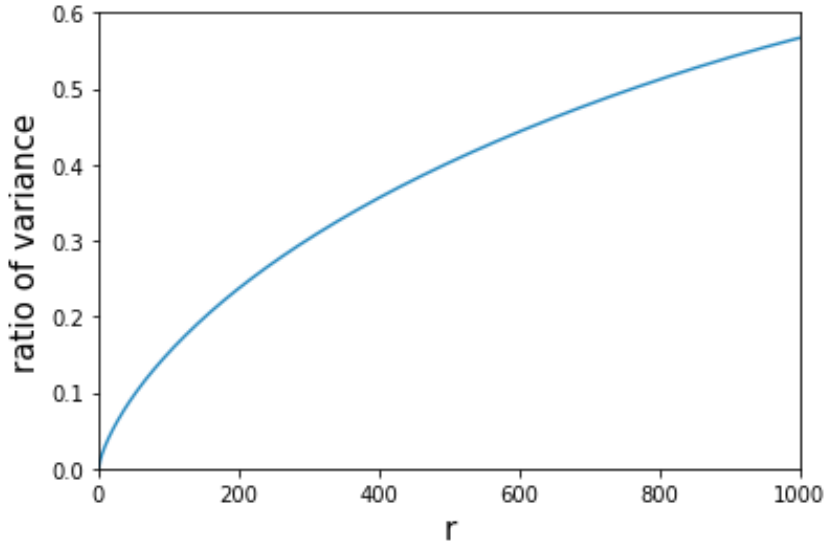
Adjusted Rand-Index: 0.174

Adjusted Mutual Info: 0.248

Part 3

(a)

(i) The ratio of variance vs r: num of top r principle components is as follows. We can see when r reaches 1000, the ratio of variance reaches about 0.6.



(ii) SVD:

We trained k-means cluster with different r values and get the following results.

r	Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Adjusted Mutual Info
1	0.001	0.001	0.001	0.001	0.001
2	**0.536**	**0.539**	**0.542**	**0.621**	**0.536**
3	0.305	0.369	0.334	0.254	0.305
5	0.226	0.313	0.262	0.149	0.225
10	0.240	0.326	0.277	0.164	0.240
20	0.239	0.325	0.275	0.163	0.239
50	0.242	0.326	0.278	0.166	0.242
100	0.244	0.328	0.280	0.168	0.244
300	0.251	0.334	0.286	0.177	0.251

Contingency Matrix:

r=1

[[1817 2086]

[1730 2249]]

r=2

[[157 3746]

[3300 679]]

r=3

[[3888 15]

[1939 2040]]

r=5

[[3898 5]

[2412 1567]]

r=10

[[3 3900]

[1639 2340]]

r=20

[[3900 3]

[2347 1632]]

r=50

[[3900 3]

[2333 1646]]

r=100

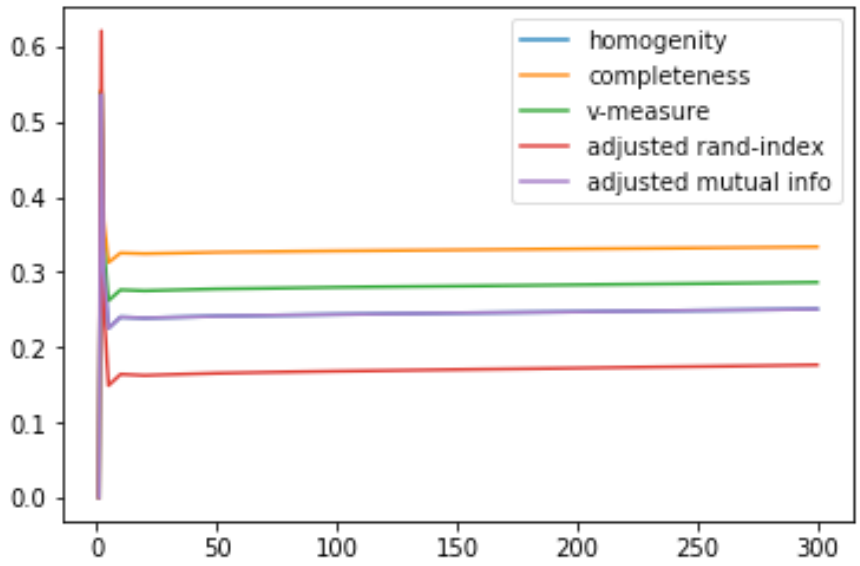
[[3900 3]

[2320 1659]]

r=300

[[3 3900]

[1698 2281]]



We can see from this graph, when r=2, the 5 measure all the best. The k-means cluster performs best at r=2.

NMF:

r	Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Adjusted Mutual Info
1	0.001	0.001	0.001	0.001	0.001
2	**0.622**	**0.622**	**0.622**	**0.728**	**0.622**
3	0.242	0.319	0.275	0.176	0.241
5	0.183	0.281	0.222	0.105	0.183
10	0.188	0.287	0.227	0.106	0.187
20	0.095	0.214	0.131	0.030	0.095
50	0.065	0.190	0.097	0.014	0.065
100	0.004	0.033	0.007	0.000	0.004
300	0.036	0.143	0.058	0.005	0.036

Contingency Matrix:

r=1

[[1817 2086]

[1730 2249]]

r=2

[[3653 250]

[329 3650]]

r=3

[[3888 15]

[2274 1705]]

r=5

[[5 3898]

[1319 2660]]

r=10

[[2 3901]

[1326 2653]]

r=20

[[2 3901]

[724 3255]]

r=50

[[3902 1]

[3472 507]]

r=100

[[3876 27]

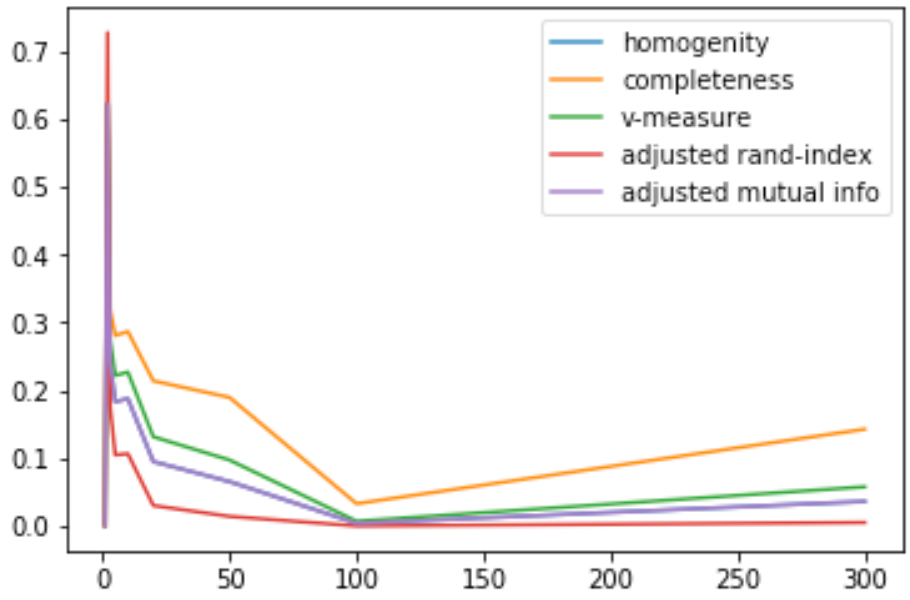
[3881 98]]

r=300

[[3895 8]

[3653 326]]

We can see from this graph, when r=2, the 5 measure all the best. The k-means cluster performs best at r=2.



Question: How do you explain the non-monotonic behavior of the measures as r increases?

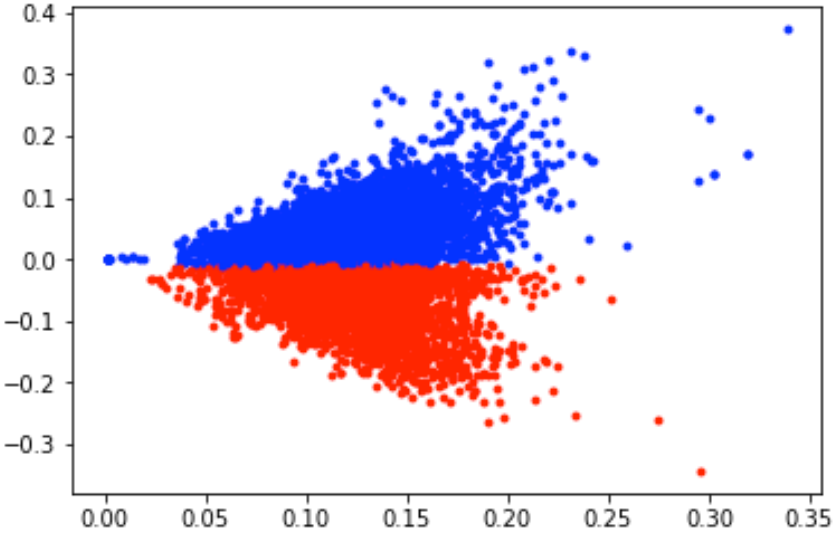
When at first r is small, too few features cannot fully do the clustering, then, when r increases, the feature number increases, thus more feature will lead to better performance. After r reaches

its best performance and r continues to increase, too many useless features that are common to all documents may cause the performance to be decreasing since then.

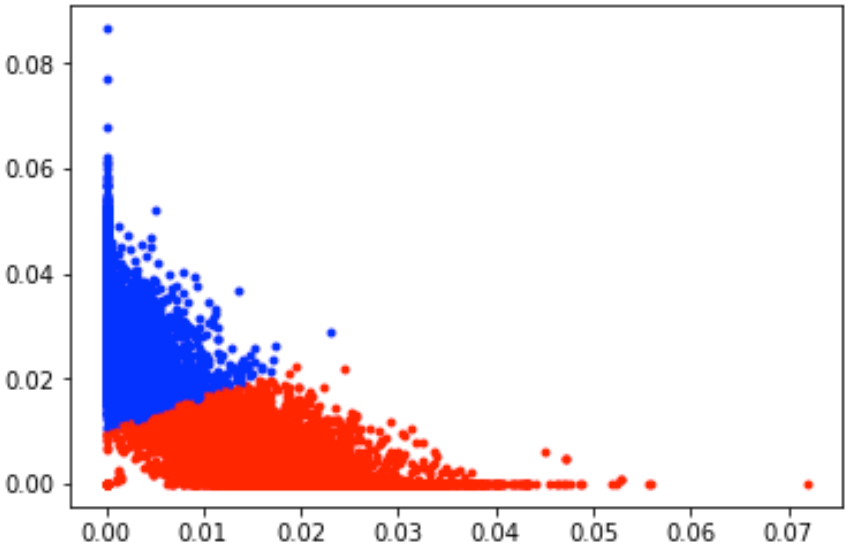
Part 4

(a)

Clustering Result: r=2, SVD



Clustering Result: r=2, NMF



(b)

The table of 5 measures of clusters for different methods

Feature Preprocessing	Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Adjusted Mutual Info
svd	0.536	0.539	0.542	0.621	0.536
normalize+svd	**0.209**	**0.237**	**0.222**	**0.227**	**0.209**
nmf	0.622	0.622	0.622	0.728	0.622
normalize+nmf	**0.672**	**0.675**	**0.674**	**0.765**	**0.672**
log+nmf	0.655	0.660	0.658	0.743	0.655
norm+log+nmf	**0.695**	**0.695**	**0.695**	**0.792**	**0.695**
log+nor+nmf	0.681	0.683	0.682	0.773	0.681

We can see from this table:

1) After normalization+svd, the performance of kmeans cluster is poor.
2) After normalization +nmf, the performance of kmeans cluster is better.
3) After log +nmf, the performance of kmeans cluster is better
4) When combine log and normalization, first normalize data and then do log will achieve best performance.

Question: Can you justify why logarithm transformation may increase the clustering results?

Because logarithm transformation can deal with skewed data.

The following are detailed results for each cluster:

1) Normalized+Nmf, r=2, clustering result:

Contingency matrix:

[[380 3523]

[3864 115]]

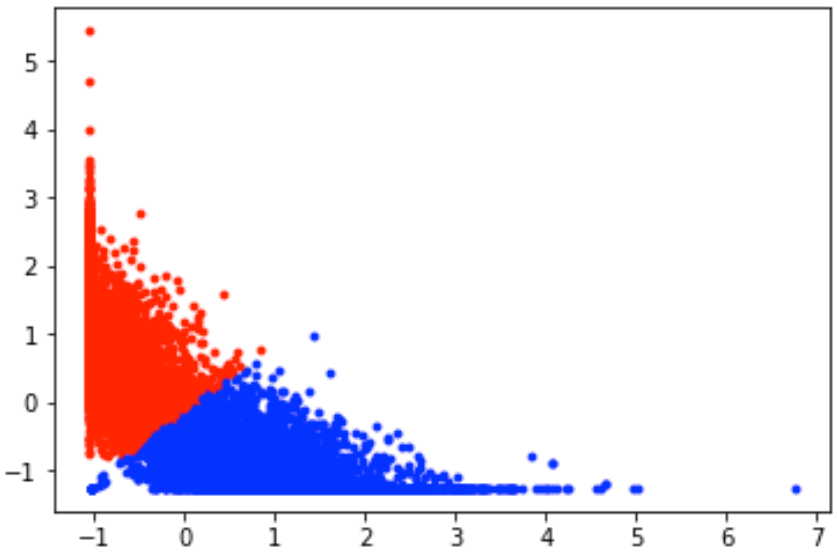
Homogeneity: 0.672

Completeness: 0.675

V-measure: 0.674

Adjusted Rand-Index: 0.765

Adjusted Mutual Info: 0.672



2) Normalized SVD, r=2, clustering result:

Contingency matrix:

[[1791 2112]

[3705 274]]

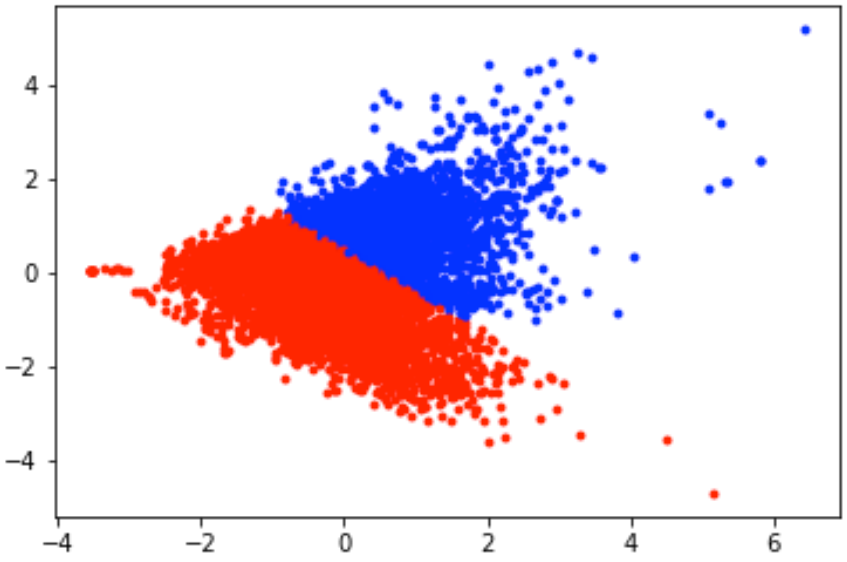
Homogeneity: 0.209

Completeness: 0.237

V-measure: 0.222

Adjusted Rand-Index: 0.227

Adjusted Mutual Info: 0.209



3) logarithm+nmf, r=2, clustering result:

Bias to avoid 0 values before log: 0.001

contingency matrix:

[[450 3453]

 [3884 95]]

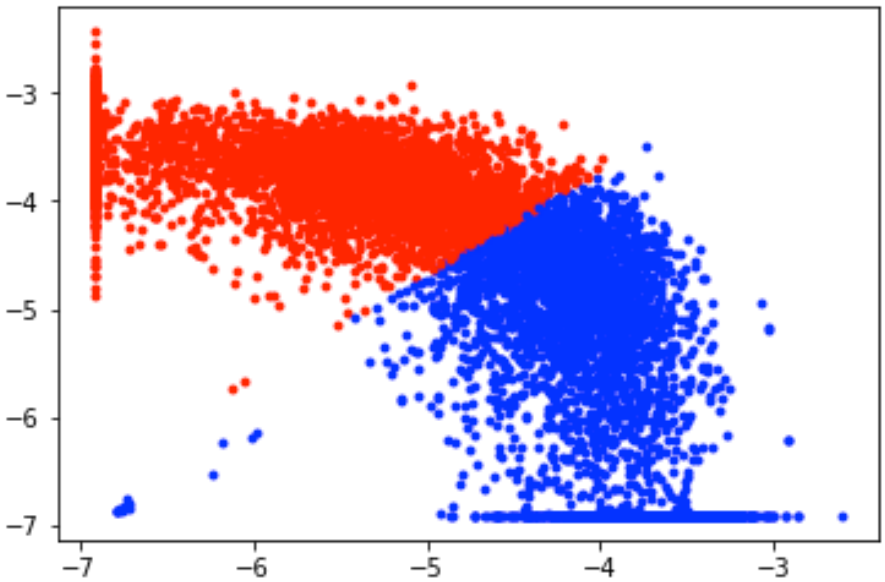
Homogeneity: 0.655

Completeness: 0.660

V-measure: 0.658

Adjusted Rand-Index: 0.743

Adjusted Mutual Info: 0.655



4) Normalize+logarithm+nmf, r=2, clustering result:

Bias to avoid 0 values before log: 0.1

[[3629 274]

[160 3819]]

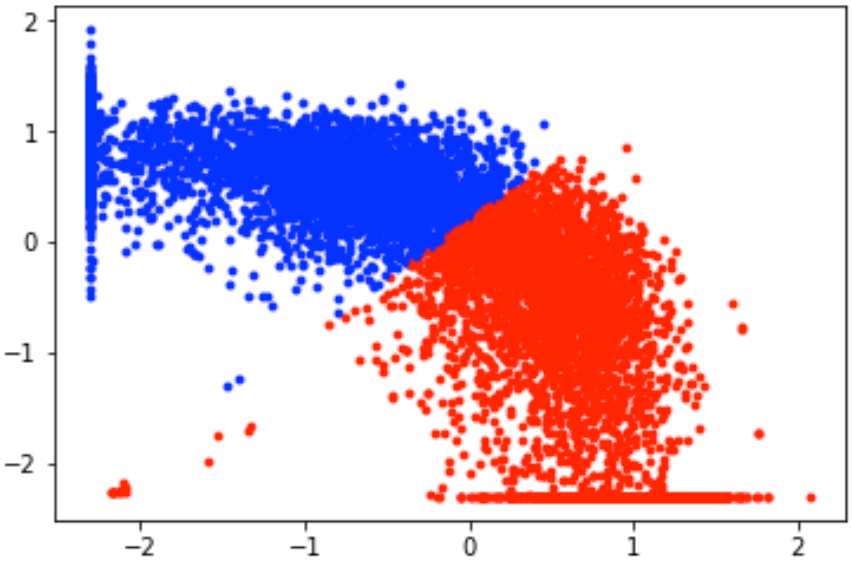
Homogeneity: 0.695

Completeness: 0.695

V-measure: 0.695

Adjusted Rand-Index: 0.792

Adjusted Mutual Info: 0.695



5)logarithm+normalize+nmf, r=2, clustering result:

Bias to avoid 0 values before log: 0.001

[[3544 359]

[116 3863]]

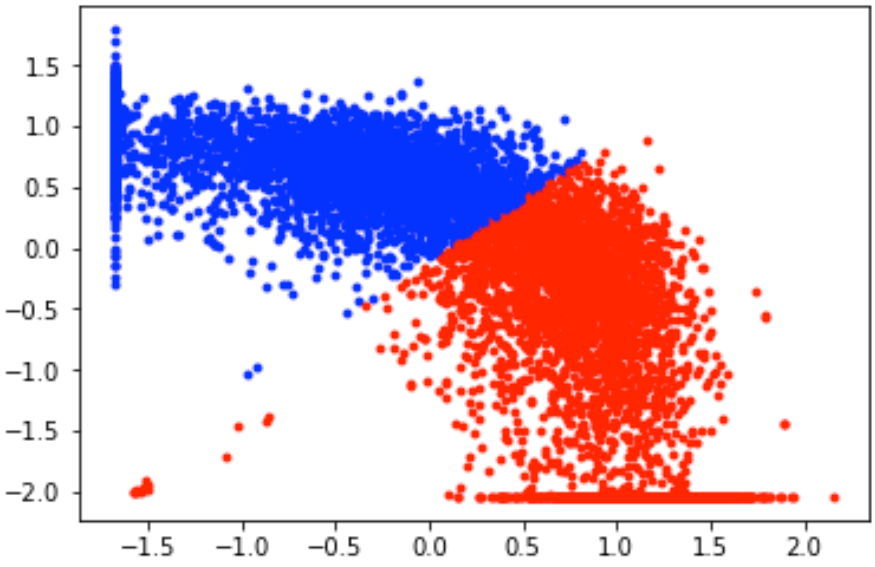
Homogeneity: 0.681

Completeness: 0.683

V-measure: 0.682

Adjusted Rand-Index: 0.773

Adjusted Mutual Info: 0.681



Part 5:

part1:

By using 20 types-dataset with min_df =3

Dimension of the data: 18846, 48401;

part2:

(a) contingency matrix

[[1 13 0 41 15 225 0 0 0 1 0 0 0 0 136 89 0 70 142 66]
[16 9 66 0 2 406 4 34 0 0 0 0 0 0 1 1 260 74 1 99]
[6 11 93 0 6 246 2 471 0 0 0 0 9 0 2 0 52 31 0 56]
[7 29 208 0 6 310 16 48 0 0 0 0 182 0 0 0 11 25 0 140]
[11 32 114 0 18 566 12 8 0 0 0 0 68 0 1 0 8 21 0 104]
[23 6 6 0 21 169 1 58 0 2 0 0 0 0 0 0 556 66 0 80]
[3 46 25 0 15 200 542 10 0 0 0 0 34 7 0 6 3 5 0
79]
[8 50 0 0 18 326 14 2 0 0 0 0 0 0 20 3 2 25 0
522]
[18 17 0 0 12 263 13 0 0 0 0 0 0 0 2 0 0 103 0 568]
[3 8 0 0 8 338 0 0 0 0 0 0 0 484 2 4 3 2 0 142]
[3 11 0 0 55 154 1 0 0 0 0 0 0 735 1 0 0 2 0
37]
[24 7 1 0 9 156 0 6 0 517 0 0 0 0 59 18 11 51 0 132]

[32 8 16 0 8 610 6 6 0 2 0 0 3 0 0 0 10 51 0 232]

[19 9 0 0 9 623 1 1 77 0 0 0 0 0 18 2 6 20 3 202]

[565 1 0 0 10 279 1 1 0 0 0 0 0 0 14 1 6 25 0 84]

[3 3 0 1 11 294 0 1 0 0 0 0 0 0 36 0 4 17 536 91]

[5 15 3 1 4 98 3 0 0 4 0 0 0 0 589 73 0 14 0 101]

[0 17 0 0 2 169 0 0 0 0 192 405 0 0 112 0 6 3 2 32]

[15 53 0 0 0 147 0 0 0 2 0 0 0 1 432 5 0 25 3 92]

[4 20 0 70 6 190 0 0 2 0 0 1 0 0 102 11 0 18 135 69]]

(b) The five measures using k-means:

Homogeneity: 0.348

Completeness: 0.430

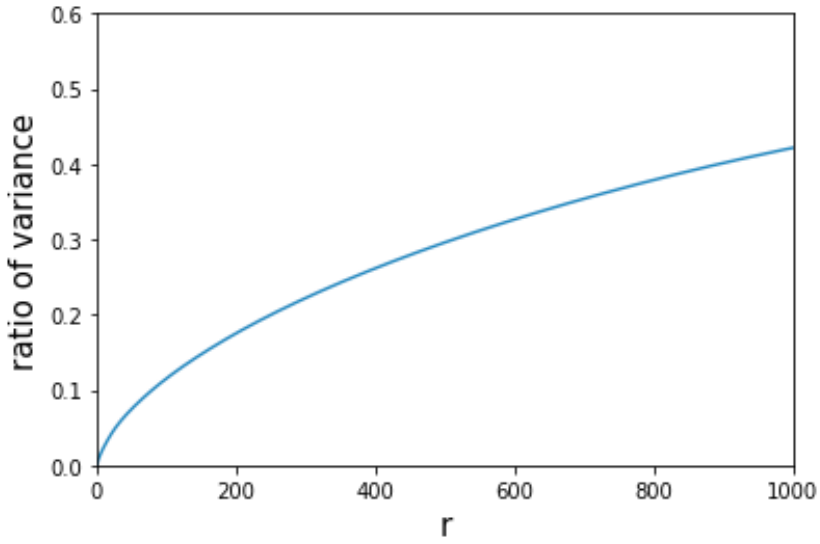
V-measure: 0.385

Adjusted Rand-Index: 0.126

Adjusted Mutual Info: 0.346

part3:

(i) The ratio of variance vs r: num of top r principle components is as follows. We can see when r reaches 1000, the ratio of variance reaches about 0.45.



(ii)

for r = [1,2,3,5,10,20,50,100,300]

I). NMF r=1:

Homogeneity: 0.028

Completeness: 0.030

V-measure: 0.029

Adjusted_Mutual_info_score: 0.024

Adjusted Rand-Index: 0.006

confusion matrix:

[[43 46 77 10 70 57   0 80   0 21 10 78 61 51 21 72 73 26   3   0]
 [112   8 59 61 25 96   5 46 47 102   2 84 17 100   4 30 59 94   0 22]
 [84 25 84 39 40 95   0 65 16 48   0 117 33 91 10 50 102 61   1 24]
 [96 16 84 50 45 92   0 68 18 62   1 100 39 85   3 58 84 78   0   3]
 [86 17 68 48 29 94   0 78 24 60   2 97 33 110   3 52 83 73   3   3]
 [118   4 58 104 18 78   1 34 75 103   1 72 14 82   4 31 53 115   6 17]
 [110   7 59 73 22 82   0 60 53 112   0 84 15 107   0 36 62 88   1   4]
 [83 20 77 64 34 112   0 56 20 70   1 107 40 89   9 55 83 67   1   2]
 [115   6 71 41 35 127   0 56 13 62   0 122 24 107   3 31 90 92   0   1]
 [89 26 99 30 40 106   0 76 24 54   2 87 22 98   6 64 83 77   0 11]
 [74 13 83 48 29 111   0 51 52 63   1 92 17 99   1 45 102 93   0 25]
 [44 57 105 13 83 74   3 114   6 18 17 69 81 54 31 109 76 34   2   1]
 [117   6 58 50 15 113   0 69 29 86   0 98 13 123   6 41 72 85   0   3]
 [107   4 71 92 24 97   3 44 26 105   1 81   9 106   4 41 66 106   0   3]
 [98 23 78 38 34 102   0 64 20 65   2 105 23 110   8 59 88 65   1   4]
 [61 46 76 42 58 99   1 66 27 49 13 87 54 82 45 56 83 48   0   4]
 [42 67 87 18 79 50   0 95   2 17 11 57 75 52 51 93 80 29   4   1]
 [71 17 66 25 56 103   0 90 30 57   4 80 24 82 11 66 95 55   6   2]
 [47 49 76   9 68 60   1 68   5 24   7 80 45 49 28 63 54 34   7   1]
 [40 28 60   7 63 56   0 57   6 28   7 48 32 34 15 64 46 30   5   2]]

ii). NMF r=2

Homogeneity: 0.171

Completeness: 0.184

V-measure: 0.177

Adjusted_Mutual_info_score: 0.169

Adjusted Rand-Index: 0.049

confusion matrix:

[[36 93   5 124 52   3   1 31   7 41 15 53   3 66   3 95 73   0 76 22]
 [37   0 210   0   3 183 37   0 98 13 13   2 113   0 208   0   0   7   7 42]
 [20   0 229   0   2 101 117   0 45   6 17   1 239   0 176   0   0   1   2 29]
 [10   0 222   0   0 125 123   0 35   3 22   1 186   0 211   0   0   1   1 42]
 [18   0 238   0   0 130 62   0 44   4 30   1 177   0 202   0   0   2   0 55]
 [42   0 165   0   1 250 43   0 130   5 11   1 97   0 192   0   0   7   1 43]
 [32   0 188   0   1 173 78   0 106   5   8   1 155   0 205   0   0   1   2 20]
 [139   0 104   9 18 137 14   0 97 54 64 14 45   2 114   1   0   0 27 151]
 [189   0 78   8   9 143   9   0 67 68 32 24 33   0 113   0   0   0 62 161]
 [152   0 98   1 20 107   6   0 85 86 56   7 50   2 132   0   0   0 36 156]
 [118   0 128   1 14 165   9   0 120 47 40   8 47   2 130   0   0   0 16 154]
 [101   5 30 84 111 14 12   2 31 153 65 44 21 56 26   3 10   0 145 78]
 [105   0 166   1   0 178 12   0 81 14 37 10 79   0 180   0   0   0 10 111]
 [176   1 38 37 22 80   4   0 150 49 13 117 19   9 91   4   1   4 106 69]
 [142   0 108   6 23 123 10   0 80 69 48 19 39   5 122   1   0   2 59 131]
 [27 196   1 80 13   8   0 85 24   9   0 117   1 43   4 166 158   0 60   5]
 [42 49   8 139 93 13   1   6 19 71 22 66   6 120   7 58 57   0 102 31]
 [33 151 20 125 27 20   0   9 20 26 13 104   3 37 12 182 54   0 77 27]
 [61 45   5 96 68   8   4   7 17 48 36 72   6 70   4 70 27   0 109 22]
 [30 72   8 72 17   4   1 40 23 25 10 58   7 46   6 69 79   0 37 24]]

iii)NMF r=3

Homogeneity: 0.204

Completeness: 0.212

V-measure: 0.208

Adjusted_Mutual_info_score: 0.201

Adjusted Rand-Index: 0.065

confusion matrix:

[[99   3 51   0   4 134 14   0 35 22 123 22 40   0   4 22   1   9 117 99]
 [1 81   4 91 79   0   8 168   0 49   0 98 18 14 10 40 192 117   3   0]

[0 153 2 209 79 0 2 193 0 10 0 49 9 66 6 27 98 81 1 0]

[0 110 1 205 81 0 12 208 0 13 0 24 17 70 3 26 127 85 0 0]

[0 115 0 81 139 0 33 146 0 46 0 42 34 5 5 48 120 149 0 0]

[0 69 0 96 63 0 5 234 0 21 0 81 12 23 3 32 275 74 0 0]

[3 69 0 47 118 0 30 67 0 133 0 84 78 8 2 132 113 87 4 0]

[29 23 33 0 69 1 36 17 0 139 3 124 81 0 12 141 122 140 20 0]

[25 15 41 4 42 0 16 23 0 89 2 155 80 0 22 130 158 166 28 0]

[88 3 1 0 29 0 90 3 0 237 5 41 244 0 1 146 27 67 12 0]

[51 1 3 0 11 0 129 1 0 256 4 37 292 0 0 180 7 19 8 0]

[22 48 84 127 34 4 5 96 1 21 7 28 12 15 258 18 107 87 12 5]

[0 62 5 29 116 0 13 77 0 73 0 92 33 0 15 86 216 162 5 0]

[51 4 72 0 11 8 27 10 0 98 16 196 81 2 7 196 52 40 118 1]

[36 33 38 8 56 0 16 39 0 109 9 116 66 1 18 142 99 155 46 0]

[51 0 64 0 1 240 0 1 102 17 90 31 5 0 1 27 10 2 122 233]

[140 0 155 1 9 77 16 2 5 25 125 37 49 0 16 37 24 21 117 54]

[111 1 42 0 1 207 19 1 8 35 113 28 50 0 2 45 2 9 176 90]

[112 0 110 0 16 83 28 1 5 30 77 43 38 0 8 51 11 13 115 34]

[62 0 47 0 3 106 13 1 53 20 56 21 34 0 1 19 9 5 81 97]]

iv). NMF r=5

Homogeneity: 0.284

Completeness: 0.298

V-measure: 0.291

Adjusted_Mutual_info_score: 0.282

Adjusted Rand-Index: 0.101

confusion matrix:

[[133 3 0 1 182 22 0 6 0 23 70 49 0 0 6 58 155 0 53 38]

[1 149 0 91 0 5 26 12 0 58 92 17 0 199 245 6 2 0 70 0]

[0 179 0 95 1 5 145 4 0 21 34 7 0 332 127 1 0 0 34 0]

[0 143 3 133 0 11 137 10 0 33 38 24 3 255 149 6 0 1 36 0]

[0 217 2 91 0 6 19 12 0 24 105 54 1 160 189 3 0 0 80 0]

[0 113 0 147 0 8 28 5 0 66 53 13 0 186 307 5 0 0 57 0]

[0 134 18 49 1 17 31 37 1 97 196 65 7 61 85 7 1 1 167 0]

[0 3 3 15 0 110 0 37 0 260 179 62 0 0 26 135 1 0 159 0]

[0 2 19 20 10 143 0 110 0 283 118 30 2 0 27 81 4 10 137 0]

[0 0 247 1 0 16 0 211 66 44 80 23 238 0 2 10 0 0 56 0]

[0 1 157 0 0 2 0 128 253 10 15 10 413 0 4 0 0 0 6 0]

[0 6 1 55 0 312 3 4 0 71 30 9 0 8 11 112 2 349 18 0]

[0 54 3 84 1 39 1 52 0 182 137 26 1 46 152 32 0 0 174 0]

[2 2 1 6 17 46 2 36 0 269 162 56 0 2 39 67 26 0 256 1]

[0 12 26 15 6 109 1 54 0 229 130 33 0 5 37 140 12 6 172 0]

[341 0 1 2 163 6 0 1 0 34 24 3 1 1 14 2 195 0 53 156]

[0 0 0 2 4 288 0 10 0 99 62 36 0 0 1 236 23 114 35 0]

[17 0 2 0 139 37 0 22 0 92 94 63 0 0 7 81 277 0 107 2]

[3 0 7 1 18 196 0 21 0 122 57 44 0 0 1 159 57 34 55 0]

[98 1 0 2 79 37 0 8 0 41 45 35 0 0 2 32 140 0 50 58]]

v). NMF r=10

Homogeneity: 0.284

Completeness: 0.313

V-measure: 0.298

Adjusted_Mutual_info_score: 0.282

Adjusted Rand-Index: 0.115

confusion matrix:

[[0 4 0 26 14 1 46 6 84 105 4 114 64 17 0 1 0 125 188 0]

[1 12 175 0 27 8 0 8 87 63 0 111 80 274 2 0 0 9 1 115]

[2 17 247 0 7 2 0 14 37 49 0 50 40 147 0 0 0 4 0 369]

[4 17 179 0 11 2 0 21 39 84 0 59 30 184 12 0 3 0 0 337]

[1 9 260 0 8 6 0 22 107 57 0 126 25 195 13 0 1 2 0 131]

[12 14 115 0 30 18 0 4 64 85 0 92 79 349 3 0 0 2 0 121]

[1 10 132 0 5 1 0 30 175 92 0 227 1 182 24 0 12 6 1 76]

[0 20 3 5 2 4 0 34 93 262 0 250 28 153 19 0 1 116 0 0]

[0 23 1 0 30 2 0 13 89 367 0 170 85 146 5 0 2 57 6 0]

[0 8 0 0 1 1 0 6 154 109 0 253 3 86 3 0 354 16 0 0]

[0 7 0 1 1 0 0 6 79 30 0 94 2 65 6 0 703 4 1 0]

[417 21 7 19 27 8 0 6 34 59 5 44 26 40 7 243 0 24 0 4]

[9 40 66 0 22 14 0 6 99 118 0 213 39 306 5 0 1 17 0 29]
[0 31 2 7 5 3 1 7 152 111 0 281 21 197 2 0 0 164 4 2]
[0 399 2 8 6 276 0 2 41 39 0 90 17 87 0 0 0 19 0 1]
[0 4 0 4 5 2 204 3 27 46 0 98 23 67 0 0 0 61 452 1]
[3 5 0 282 8 3 0 12 63 125 76 59 8 15 4 2 0 244 1 0]
[0 0 1 351 2 0 2 17 66 8 266 54 4 12 1 0 0 154 2 0]
[1 28 0 148 18 4 0 46 45 67 29 70 68 15 14 1 0 215 6 0]
[1 5 0 31 2 0 69 16 52 60 4 74 21 44 2 0 0 108 139 0]]

vi). NMF r=20

Homogeneity: 0.303

Completeness: 0.383

V-measure: 0.338

Adjusted_Mutual_info_score: 0.300

Adjusted Rand-Index: 0.093

confusion matrix:

[[5 0 162 0 150 1 71 1 0 0 0 0 1 0 136 0 2 0 17 253]
[44 2 1 0 213 2 78 16 0 2 296 0 0 0 0 0 1 0 0 318]
[44 0 0 0 110 13 35 2 0 2 634 0 0 0 0 0 0 7 0 138]
[55 3 0 0 170 5 28 8 0 3 205 2 0 0 0 0 0 186 0 317]
[57 3 0 0 377 1 24 9 0 13 70 0 0 0 0 0 0 72 0 337]
[21 0 0 0 141 2 85 26 0 2 372 0 4 0 0 0 0 0 0 335]
[37 69 0 0 427 8 5 4 0 32 50 14 0 0 0 0 3 50 0 276]
[34 543 0 0 136 1 21 8 0 27 3 0 0 0 0 0 3 0 6 208]
[118 272 0 0 150 1 83 16 0 5 0 0 0 0 0 0 0 0 0 351]
[46 0 0 0 310 1 3 4 0 7 0 403 0 0 0 0 4 0 1 215]
[189 0 0 0 123 0 2 6 0 6 0 598 0 0 0 0 0 0 1 74]
[16 0 0 0 92 40 54 7 0 8 13 0 448 1 0 0 16 0 55 241]
[93 50 0 0 278 1 50 35 0 7 39 0 2 0 0 0 0 3 0 426]
[36 0 3 0 240 3 19 19 0 5 3 0 0 0 0 78 2 0 10 572]
[32 1 0 0 148 104 26 397 0 0 4 0 0 0 0 0 1 0 11 263]
[17 1 521 0 108 0 18 3 0 1 1 0 0 0 0 0 0 0 10 317]
[12 3 0 0 87 2 13 5 0 5 0 0 3 0 0 0 69 0 564 147]

[19 0 3 400 129 0 3 0 170 1 0 0 0 0 0 0 0 0 32 183]

[6 4 3 0 124 1 12 13 0 28 0 0 2 124 0 0 5 0 234 219]

[12 1 172 0 113 4 15 0 0 2 1 0 0 1 16 2 0 0 75 214]]

vii). NMF r=50

Homogeneity: 0.214

Completeness: 0.387

V-measure: 0.275

Adjusted_Mutual_info_score: 0.211

Adjusted Rand-Index: 0.040

confusion matrix:

[[0 0 0 207 1 3 348 0 0 0 0 0 1 0 47 84 38 70 0 0]

[0 0 1 803 1 7 15 2 0 0 0 0 10 0 5 2 50 76 1 0]

[0 0 11 779 1 8 15 2 2 0 0 0 61 0 14 0 40 46 6 0]

[0 0 2 821 2 21 1 3 0 0 0 3 17 0 21 3 43 40 5 0]

[0 0 1 808 1 15 12 14 0 0 0 0 37 0 25 0 24 25 1 0]

[0 0 1 514 15 3 4 2 0 0 0 0 338 0 2 1 24 80 4 0]

[0 2 6 879 0 12 7 33 0 0 0 15 9 0 3 0 0 5 4 0]

[0 3 0 855 0 7 23 31 0 0 0 0 10 0 23 0 10 24 4 0]

[0 0 1 841 0 8 24 6 0 0 0 0 7 0 9 0 3 94 3 0]

[0 0 2 473 0 1 27 7 0 0 0 457 10 0 10 2 1 3 1 0]

[0 0 0 203 0 4 8 6 0 0 0 730 4 0 8 0 0 3 33 0]

[0 5 24 492 290 0 70 9 0 0 25 0 13 0 0 0 9 53 0 1]

[0 0 13 815 6 1 18 7 13 0 0 1 7 0 8 0 41 48 6 0]

[0 0 2 590 1 2 251 5 14 76 0 0 7 0 8 0 14 20 0 0]

[0 1 406 447 0 1 61 0 0 0 0 0 5 0 3 0 36 27 0 0]

[0 0 0 324 0 0 598 0 19 0 0 0 12 0 10 0 8 26 0 0]

[0 277 0 506 3 7 72 5 0 0 0 0 3 0 19 0 0 15 3 0]

[358 0 0 234 0 14 114 1 0 0 0 0 15 164 32 0 3 5 0 0]

[0 8 1 382 0 20 186 27 0 0 0 2 5 0 7 0 0 12 1 124]

[1 5 4 270 0 15 262 2 0 2 0 0 0 0 18 13 4 17 15 0]]

viii). NMF r=100

Homogeneity: 0.181

Completeness: 0.272

V-measure: 0.217

Adjusted_Mutual_info_score: 0.178

Adjusted Rand-Index: 0.035

confusion matrix:

```
[[  3   1   1   2   0 197  40  37  63   0   0   9  43   0 323  80   0   0   0   0]
 [ 21   1   1   1   2 736   0   0 102   0   0  16  72   0   0  13   2   0   6   0]
 [ 11   0   1   0  13 672   0   0 143   0   0  22  89   0   0  23   2   0   9   0]
 [ 20   0   0   0   3 637   0   0 160   2   0  26  95   0   0  13   3   0  23   0]
 [ 11   0   1   0   1 752   0   0  88   0   0  10  55   0   0  19  14   0  12   0]
 [ 36   0  14   0   2 604   0   0 191   0   0  28  98   0   0  10   2   0   3   0]
 [  7  25   0   4   8 686   0   0  82  13   0  15  84   0   0   8  33   0  10   0]
 [ 21 325   0   3   0 274   0   0 112   0   0  51 146   0   0  28  27   0   3   0]
 [ 21  13   0   0   1 458   0   0 216   0   0  63 172   0   0  39   6   0   7   0]
 [  3   0   0   4   1 265   0   0  77 513   0  22  80   0   0  21   7   0   1   0]
 [  7   0   0   0   0 499   0   0  79 305   0   9  47   0   0  44   6   0   3   0]
 [  7   0 330  16  36 215   0   1 106   0   1  71 110   0   0  91   7   0   0   0]
 [ 26  11   3   0   1 651   0   0 142   0   0  30  73   0   0  28   6   0   2  11]
 [ 16   0   0   2   3 515   0   0 185   0   0  22  85  76   1  76   5   0   4   0]
 [ 99   0   0   1  94 480   0   0  90   0   0   2  62   0   0  86   0   0   1  72]
 [  6   0   0   0   0 224   1 423  64   0   0   9  52   0 124  92   1   0   1   0]
 [  5   4   2  68   3 324   1   0 163   0   0  23  95   0   0 207   5   0  10   0]
 [  4   0   0   0   0 406   0   0  76   0   0  17  42   0  38 102   1 238  16   0]
 [ 13   1   1   5   1 192   0  10  99   0 128  24  75   0   1 177  26   0  22   0]
 [  2   1   0   1   4 165  70 106  89   0   0  13  53   2  49  53   2   2  16   0]]
```

ix). NMF r=300

Homogeneity: 0.058

Completeness: 0.079

V-measure: 0.067

Adjusted_Mutual_info_score: 0.055

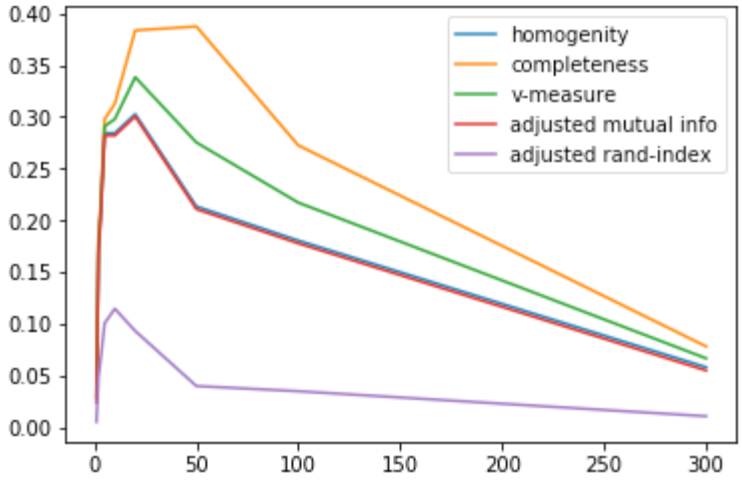
Adjusted Rand-Index: 0.011

confusion matrix:

```
[[  0   0  53  84 177  28  16   0  33   3   0   0   4 116 108   0 173   0   0   4]
 [  2   9  65  56 190   2  31   0  54   6   0   0  23   4  60   0 345  21   0 105]
 [  1   3 100  75 209   7  32   0  71   3   0   0  15   6  73   0 260  35   1  94]
 [  2  15 104  38 158   2  37   0  90  14   0   0   8   1  83   4 291   9  84  42]
 [ 13   6  46  75 235   0  27   0  47   4   0  15 164  13  50   0 251   1   0  16]
 [  2   2 120  25 126   1  51   0  80  15   0   0   3   6 101   0 368  25   0  63]
 [ 30   7  46  48 240   3  34   0  51   8   0   0  16   4  49   8 405   1   8  17]
 [ 23   0 149  70 182  18  56   0 121  16   0   0   0  21 118   0 211   0   0   5]
 [  5   4 171  48 168   8  63   0 138  40   0   0   0   8 143   0 198   0   0   2]
 [  7   1  92 195 226   7  40   0  75   5   0   0   0  22  95   0 226   0   0   3]
 [  6   1  62  54 266   3  16   0  34   3   0   0   0  29  48   0 476   0   0   1]
 [  7   1 134  48 120  37  52   0  81  15   0   0   3  51 175   0 236   6   0  25]
 [  6   7  99  75 234   6  21   0  58   8  18   0   2   7  84   0 306  17   1  35]
 [  3   5 113 117 191  19  25   0  71  16   0   0   1  35 133   0 247   2   0  12]
 [  0   3 100 104 202  13  25   0  56   5   0   0   1  28 106   0 262  14   0  68]
 [  1   1  46  69 158  26  21   0  37   0   0   0   1 189 148   0 298   1   0   1]
 [  5  15 104  76 161  50  49   0  93  12   0   0   1 107  96   0 138   0   0   3]
 [  0   3  70  96 171  11   9 123  27  11   0   0   0  75  68   0 269   2   0   5]
 [ 14   0  97  50 106  55  32   0  79  18   0   0   0  83  89   0 136   6   0  10]
 [  2  12  62  52 100  27  32   0  44   7   0   0   0  67  79   0 141   0   0   3]]
```

r	Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Adjusted Mutual Info
1	0.028	0.030	0.029	0.006	0.024
2	0.171	0.184	0.177	0.049	0.169
3	0.204	0.212	0.208	0.065	0.201
5	0.284	0.298	0.291	0.101	0.282
10	0.284	0.313	0.298	0.115	0.282
20	**0.303**	**0.383**	**0.338**	**0.093**	**0.300**
50	0.214	0.387	0.275	0.040	0.211
100	0.181	0.272	0.217	0.035	0.178
300	0.058	0.079	0.067	0.055	0.011

Then, in order to find the best performance for the 9 r-values, we drew a graph showing the tendency of each 5 measurement as shown below:



According to graph, I choose to use 20 as my best r value, since at the point r=20, the five measures tends to reach their highest value in average.

x). SVD r=1

Homogeneity: 0.028

Completeness: 0.030

V-measure: 0.029

Adjusted_Mutual_info_score: 0.024

Adjusted Rand-Index: 0.006

confusion matrix:

[[67 42 56 81 12 21 2 52 81 62 0 20 11 45 70 0 29 0 73 75]
[26 118 15 83 62 4 3 96 53 97 47 105 2 8 26 22 98 2 68 38]
[41 87 27 123 39 10 1 93 83 94 17 48 0 24 44 24 65 0 100 65]
[53 103 36 105 50 3 0 82 80 91 18 64 1 16 42 3 82 0 91 62]
[40 91 32 104 49 3 3 108 68 95 24 60 2 17 28 3 74 0 88 74]
[33 117 14 76 106 4 7 83 56 76 75 110 1 4 14 17 113 0 58 24]
[31 118 14 82 77 0 1 100 56 84 53 111 0 7 20 4 92 0 70 55]
[50 88 40 107 65 9 1 92 77 111 20 75 1 19 31 2 67 0 87 48]
[30 110 23 120 42 3 0 112 76 129 13 63 0 6 34 1 101 0 87 46]
[55 90 22 92 31 6 0 110 102 95 24 54 2 26 33 11 79 0 94 68]
[43 76 16 101 50 1 0 100 77 110 53 67 1 13 28 25 91 0 107 40]
[95 47 77 75 14 32 3 52 105 78 6 17 16 56 78 1 36 78 123]

[36 126 13 100 51   6   0 126 58 109 29 91   0   5 12   3 80   0 76 63]
[34 107   8 78 92   4   1 103 64 100 26 110   1   4 22   3 111   2 80 40]
[48 102 23 100 39   8   1 113 79 105 20 66   2 22 32   4 68   0 94 61]
[53 62 48 88 42 45   0 88 72 100 27 50 13 46 62   4 49   1 88 59]
[89 45 74 62 20 51   4 52 82 48   2 20 11 65 70   1 26   0 94 94]
[59 74 22 88 26 11   6 87 79 100 30 59   4 17 50   2 54   0 92 80]
[69 47 43 77   9 28   6 52 78 61   5 26   8 48 56   1 36   1 62 62]
[58 42 31 58   7 15   4 34 57 53   6 29   8 27 57   2 32   0 46 62]]

xi). SVD r=2

Homogeneity: 0.217

Completeness: 0.234

V-measure: 0.225

Adjusted_Mutual_info_score: 0.215

Adjusted Rand-Index: 0.068

confusion matrix:

[[1 54 46 127 44 34   6   0   0 18 16 144 136   0 73 76   2   0   0 22]
 [149   0 18   2 21   0 118 67 19   0 48   0   0 151   0   5 129   7 230   9]
 [141   0   1   0   8   0 59 84 161   0 25   0   0 271   0   4 64   1 162   4]
 [165   0   3   0   2   0 52 85 155   0 27   0   1 212   0   0 61   1 214   4]
 [237   0   5   0   4   0 95 126 30   0 60   0   0 126   0   0 72   2 198   8]
 [177   0   4   0   5   0 145 58 25   0 26   0   0 97   0   0 200   7 241   3]
 [139   0 10   1 30   0 127 56 80   0 31   0   1 140   0   1 148   1 204   6]
 [52   2 135 13 80   0 231 19   0   0 214   0   6   2   0 53 103   0 16 64]
 [42   0 204   8 127   0 234 14   0   0 150   0 16   0   0 79 56   0 21 45]
 [26   3 206 22 111   0 164   6   0   0 161   0   8   1   0 143 73   0 11 59]
 [18   0 219 23 139   0 163   6   0   0 144   0 11   0   0 115 121   0   7 33]
 [38 26 132 61 52   0 80 31   5   0 183   0 14   4   1 174 19   0   9 162]
 [160   0 69   0 46   0 262 53   1   0 148   0   1 27   0   6 103   0 100   8]
 [11   4 211 38 240   1 136   4   1   0 81   3 39   1   0 101 92   4 12 11]
 [50   4 170 15 94   0 205 19   2   0 177   0 10   1   0 87 71   1 27 54]
 [3 28 31 38 66 122   7   0   0 54   2 220 207   0 176 24 15   0   3   1]
 [1 107 81 202 52   7 27   1   0   0 33 29 107   0   1 192 15   0   1 54]

[4 48 51 100 93 11 26 0 0 2 36 179 235 0 41 81 13 0 1 19]
[3 62 106 157 69 4 10 2 0 0 28 33 116 0 7 116 7 0 1 54]
[1 30 40 67 47 44 12 2 0 28 21 89 114 0 74 41 9 0 0 9]]

xii). SVD r=3

Homogeneity: 0.238

Completeness: 0.247

V-measure: 0.243

Adjusted_Mutual_info_score: 0.236

Adjusted Rand-Index: 0.082

confusion matrix:

[[6 12 42 111 3 77 0 31 99 0 12 2 133 0 73 19 169 0 10 0]
[15 37 10 0 131 3 70 0 7 1 133 100 0 139 13 8 0 211 75 20]
[11 14 4 0 73 0 225 0 1 0 58 90 1 201 8 3 0 167 34 95]
[14 23 1 0 96 0 114 0 1 0 43 95 0 265 5 13 0 166 34 112]
[10 49 8 0 70 0 82 0 0 0 61 177 2 136 13 32 0 231 79 13]
[22 16 2 0 250 0 44 0 1 0 144 77 0 183 1 4 0 175 45 24]
[4 18 16 0 95 0 90 0 1 0 109 216 2 62 11 33 0 116 180 22]
[37 125 81 0 177 13 1 0 42 0 117 47 22 11 126 31 1 34 125 0]
[63 108 64 0 238 8 0 0 87 0 104 36 17 7 139 16 0 34 75 0]
[7 49 286 0 34 4 0 0 28 0 60 18 71 2 170 92 0 9 164 0]
[0 10 339 0 14 3 0 0 4 0 76 9 41 0 133 163 1 1 205 0]
[342 79 6 0 77 44 6 0 24 313 15 13 6 24 19 3 0 9 5 6]
[38 105 13 0 214 0 13 0 13 0 102 91 1 56 63 11 0 156 105 3]
[17 35 82 2 89 20 1 0 141 0 197 7 41 1 185 54 5 10 101 2]
[64 135 50 0 171 20 5 0 60 1 109 34 21 14 168 8 1 24 101 1]
[2 0 18 248 8 28 0 102 111 0 39 1 65 0 80 0 272 3 20 0]
[63 47 51 15 28 205 0 2 178 16 22 3 146 0 92 10 18 1 13 0]
[5 11 60 70 4 34 0 6 127 0 46 1 155 0 107 37 240 3 34 0]
[32 26 41 12 20 151 0 3 178 5 25 4 98 0 105 28 34 2 11 0]
[2 3 34 96 15 76 0 51 93 0 18 2 65 0 63 21 80 0 9 0]]

xiii). SVD r=5

Homogeneity: 0.312

Completeness: 0.329

V-measure: 0.320

Adjusted_Mutual_info_score: 0.310

Adjusted Rand-Index: 0.120

confusion matrix:

[[63 88   1 278 10   1 74   5   0 32 32 54 10 119   0   1 24   0   7   0]
 [0 11 279   0   0   0 79   0 22   0 89 17   9   1 137   0 18 125 186   0]
 [1   4 148   0   0   0 24   0 123   0 52   5   7   0 326   0 11 140 144   0]
 [0   0 174   0   0   0 23   3 128   0 49 10 21   0 184   1 27 223 139   0]
 [1   3 210   0   0   0 64   1 17   0 40 30 30   0 140   0 17 140 270   0]
 [1   1 378   1   0   0 38   0 27   0 96 11   4   0 95   2 27 183 124   0]
 [1   6 85   2   0   0 187 12 25   0 131 65 38   0 92   0 44 53 232   2]
 [10 88 22   1   0   0 202 32   0   0 267 153 24   0   0   1 154   8 28   0]
 [3 58 16 15   0   0 145 64   0   0 264 99   9   0   0   0 304   8 11   0]
 [0 15   2   0   0   0 105 344   0   0 59 203   5   0   0   0 81   0   7 173]
 [1   5   2   1   0   0 21 360   0   0 48 100   5   0   0   0 30   0   2 424]
 [58 62 22   0 34 279 25   0   3   0 46 10   5   0   5 357 21 45 19   0]
 [4 17 137   1   0   0 184 11   2   0 242 36   6   0 30   2 92 73 147   0]
 [30 138 22 15   0   0 303   4   2   1 267 123   4   1   1   1 52   2 21   3]
 [9 142 27   4   1   0 211 67   1   0 234 80   1   0   3 11 154 16 26   0]
 [14 14   8 380   1   0 79   1   0 129 47   8   1 301   1   0   9   2   2   0]
 [203 229   2   6 160   3 55   1   0   0 48 43 10   0   0 102 47   0   1   0]
 [366 79   1 11 263   0 99   0   0   2 30 43 37   3   0   0   4   0   2   0]
 [133 262   1 20 87   1 80   1   0   0 47 32 35   3   0 25 42   0   6   0]
 [25 69   1 175 16   0 60   2   0 58 47 34 15 93   0   0 28   0   5   0]]

xiv). SVD r=10

Homogeneity: 0.342

Completeness: 0.379

V-measure: 0.359

Adjusted_Mutual_info_score: 0.339

Adjusted Rand-Index: 0.146

confusion matrix:

[[0 0 14 0 1 151 50 1 10 1 0 0 36 0 126 3 72 73 0 261]
[75 0 156 1 0 41 0 16 415 0 1 0 78 0 1 9 154 3 22 1]
[350 0 98 0 0 10 0 3 354 0 0 2 51 0 0 9 68 2 38 0]
[45 0 122 4 0 19 0 5 146 0 4 103 118 1 0 27 98 290 0]
[9 0 301 1 0 33 0 9 107 0 0 20 87 1 0 31 127 1 236 0]
[100 0 82 0 0 37 0 24 521 0 7 0 90 0 0 4 122 0 1 0]
[17 0 384 13 0 45 0 5 38 0 0 18 125 3 2 48 166 2 108 1]
[0 0 233 11 0 45 0 8 11 0 0 0 356 0 4 46 239 35 2 0]
[0 0 116 29 0 74 0 16 27 0 0 0 444 0 27 16 225 16 1 5]
[0 0 88 422 0 129 0 2 2 0 0 0 118 133 5 8 84 3 0 0]
[0 0 30 451 0 31 0 3 0 0 0 0 35 396 0 11 40 1 0 1]
[8 0 19 0 0 48 0 12 43 229 464 0 59 0 16 7 46 40 0 0]
[7 0 219 8 0 61 0 34 100 0 8 1 175 0 0 7 329 0 35 0]
[2 76 83 3 0 164 1 19 13 0 0 0 137 0 5 6 409 64 0 8]
[0 0 58 1 0 68 0 534 13 0 0 0 65 0 1 1 224 21 1 0]
[1 0 21 0 0 53 214 3 7 0 0 0 44 0 2 1 139 21 1 490]
[0 0 19 0 2 106 0 4 1 2 4 0 111 0 123 12 63 462 0 1]
[0 0 13 0 471 110 2 0 2 0 0 0 17 0 3 18 72 227 0 5]
[0 0 27 5 3 79 2 13 2 1 1 0 74 0 87 48 99 326 0 8]
[0 2 7 0 0 82 71 4 3 0 1 0 44 0 53 18 88 69 0 186]]

xv). SVD r=20

Homogeneity: 0.291

Completeness: 0.388

V-measure: 0.333

Adjusted_Mutual_info_score: 0.289

Adjusted Rand-Index: 0.094

confusion matrix:

[[136 1 1 0 0 0 0 0 0 1 71 1 62 0 23 178 138 0 149 38]
[5 17 1 0 0 0 0 0 0 0 145 2 95 2 7 473 0 225 1 0]
[3 3 0 11 1 0 0 0 0 0 78 11 33 2 14 200 0 629 0 0]
[0 9 0 199 4 0 0 0 0 0 215 4 32 3 30 319 0 167 0 0]

[4 10 0 78 1 0 0 0 0 0 154 1 26 14 25 599 0 50 0 1]

[0 28 0 0 0 0 0 0 0 4 207 1 95 2 4 337 0 310 0 0]

[3 4 3 56 14 0 0 0 0 0 203 7 5 34 14 585 0 46 0 1]

[73 12 2 0 0 0 0 0 0 0 488 0 27 31 22 333 0 2 0 0]

[24 21 0 0 2 0 0 0 0 0 588 1 104 7 9 240 0 0 0 0]

[16 4 4 0 442 0 0 0 0 0 152 1 2 7 1 365 0 0 0 0]

[1 3 0 0 790 0 0 0 0 0 48 0 3 6 5 143 0 0 0 0]

[70 7 17 0 0 0 1 0 0 491 167 32 51 8 0 132 0 15 0 0]

[3 36 0 4 1 0 0 0 0 2 323 1 57 6 2 515 0 34 0 0]

[103 21 2 0 2 77 0 0 0 0 298 2 21 5 4 449 0 4 2 0]

[45 415 1 0 0 0 0 0 0 0 123 105 24 0 1 271 0 2 0 0]

[83 3 0 0 0 0 0 0 0 0 115 0 18 1 2 241 0 1 532 1]

[557 6 80 0 0 0 0 0 0 4 130 1 14 5 13 100 0 0 0 0]

[118 0 0 0 0 0 0 169 387 0 54 0 3 1 21 183 0 0 4 0]

[313 17 5 0 0 0 127 0 0 2 107 1 10 25 23 142 0 0 3 0]

[126 0 1 0 0 2 0 0 0 0 87 4 14 2 21 148 18 0 136 69]]

xvi). SVD r=50

Homogeneity: 0.272

Completeness: 0.378

V-measure: 0.317

Adjusted_Mutual_info_score: 0.270

Adjusted Rand-Index: 0.080

confusion matrix:

[[0 0 169 0 0 1 0 0 0 41 74 0 2 179 23 253 0 1 42 1]

[0 127 1 2 0 0 0 0 0 0 84 2 1 149 7 581 0 2 0 17]

[0 548 0 2 0 0 0 0 0 0 37 10 0 88 15 275 0 6 0 4]

[0 448 0 3 0 0 3 0 0 0 31 4 0 155 31 289 0 6 0 12]

[0 168 0 14 0 0 1 0 0 0 24 1 0 123 23 578 0 18 0 13]

[0 114 0 2 0 5 0 0 0 0 98 2 0 224 4 491 0 20 0 28]

[0 120 0 34 0 0 14 0 0 0 5 8 5 172 14 584 0 15 0 4]

[0 2 0 31 0 0 0 0 0 0 26 1 3 511 21 354 0 18 14 9]

[0 0 0 6 0 0 0 0 0 0 105 1 0 571 9 272 0 12 1 19]

[0 0 0 7 0 0 451 0 0 0 2 1 4 156 2 356 0 9 2 4]

[0 0 0 6 0 0 743 0 0 0 2 0 0 44 6 139 0 55 0 4]

[0 8 0 8 0 481 0 0 0 0 52 33 16 175 0 155 1 8 46 8]

[0 34 0 7 0 2 1 12 0 0 55 1 0 277 2 558 0 8 0 27]

[77 2 3 5 0 0 0 0 0 0 21 3 2 290 4 544 0 9 8 22]

[0 1 0 0 0 0 0 78 0 0 25 105 1 127 2 357 0 4 16 271]

[0 1 545 1 0 0 0 0 0 1 17 0 0 110 2 290 0 10 16 4]

[0 0 0 5 0 4 0 0 0 1 13 1 69 118 11 107 0 4 571 6]

[0 0 4 1 397 0 0 0 173 0 3 0 0 51 18 231 0 2 60 0]

[0 0 5 27 0 2 1 0 0 0 11 1 5 119 23 177 128 0 259 17]

[2 0 147 2 1 0 0 0 0 69 20 4 0 103 22 180 1 6 71 0]]

xvii). SVD r=100

Homogeneity: 0.327

Completeness: 0.493

V-measure: 0.393

Adjusted_Mutual_info_score: 0.325

Adjusted Rand-Index: 0.090

confusion matrix:

[[0 0 0 0 28 15 6 0 86 348 0 0 183 1 88 0 2 1 0 41]

[0 184 2 2 7 2 2 0 0 754 0 0 1 0 1 0 1 17 0 0]

[10 638 10 2 14 6 1 0 1 297 2 0 0 0 0 0 0 4 0 0]

[197 148 4 3 31 6 5 0 0 571 0 0 0 1 1 0 0 12 3 0]

[73 22 1 14 25 19 2 0 0 794 0 0 0 0 0 0 0 12 1 0]

[0 292 1 2 4 20 1 0 0 634 0 0 0 4 0 0 0 30 0 0]

[54 33 8 33 14 15 48 0 0 747 0 0 0 0 0 0 5 4 14 0]

[0 3 0 30 20 18 610 0 15 281 0 0 0 0 0 0 3 10 0 0]

[0 0 1 6 9 11 770 0 2 178 0 0 0 0 0 0 0 19 0 0]

[0 0 1 7 2 9 7 0 1 416 0 0 0 0 0 0 4 5 542 0]

[0 0 0 6 6 56 3 0 1 144 0 0 0 0 0 0 0 5 778 0]

[0 13 33 8 0 9 1 0 54 315 0 0 0 531 0 1 16 10 0 0]

[4 18 1 6 2 7 42 0 10 850 13 0 0 2 0 0 0 28 1 0]

[0 2 3 5 4 9 3 0 9 837 14 77 4 0 0 0 2 21 0 0]

[0 2 108 0 2 8 3 0 83 569 0 0 0 0 0 0 1 211 0 0]

[0 1 0 0 2 11 7 0 21 357 20 0 571 0 0 0 0 6 0 1]

[0 0 2 5 11 4 13 0 596 197 0 0 1 4 0 0 70 6 0 1]

[0 0 0 1 19 2 1 388 304 221 0 0 4 0 0 0 0 0 0 0]

[0 0 1 27 24 0 4 0 296 259 0 0 6 2 0 130 5 17 4 0]

[0 0 4 2 22 6 3 0 84 268 0 2 153 0 14 0 1 0 0 69]]

xviii). SVD r=300

Homogeneity: 0.313

Completeness: 0.429

V-measure: 0.362

Adjusted_Mutual_info_score: 0.310

Adjusted Rand-Index: 0.086

confusion matrix:

[[60 0 0 0 1 118 229 1 4 175 80 130 0 0 0 0 0 0 1 0]

[81 0 2 175 0 156 538 0 1 1 0 0 0 0 0 0 1 0 16 2]

[31 0 2 607 0 81 235 0 0 0 0 0 14 2 0 0 0 0 2 11]

[26 0 3 155 0 203 372 0 0 0 0 0 208 0 0 0 1 1 8 5]

[22 0 14 21 0 142 669 0 0 0 0 0 83 0 0 0 0 1 10 1]

[92 0 2 267 0 210 386 3 0 0 0 0 0 0 0 0 0 0 27 1]

[5 0 33 37 1 180 631 0 3 0 0 0 59 0 0 0 6 8 4 8]

[24 0 31 3 3 528 389 0 3 0 0 0 0 0 0 0 0 0 8 1]

[106 0 6 0 0 575 291 0 0 0 0 0 0 0 0 0 0 0 17 1]

[2 0 7 0 0 92 226 0 4 0 0 0 0 0 0 0 641 18 3 1]

[2 0 6 0 0 45 158 0 0 0 0 0 0 0 0 0 12 772 4 0]

[48 1 9 10 5 190 176 495 17 0 0 0 0 0 0 0 0 0 7 33]

[50 0 6 17 0 285 567 1 0 0 0 0 7 13 0 0 0 1 36 1]

[22 0 5 2 0 299 621 0 2 3 0 0 0 14 0 0 0 0 19 3]

[25 0 0 2 1 137 288 0 1 0 0 0 0 0 0 0 0 0 431 102]

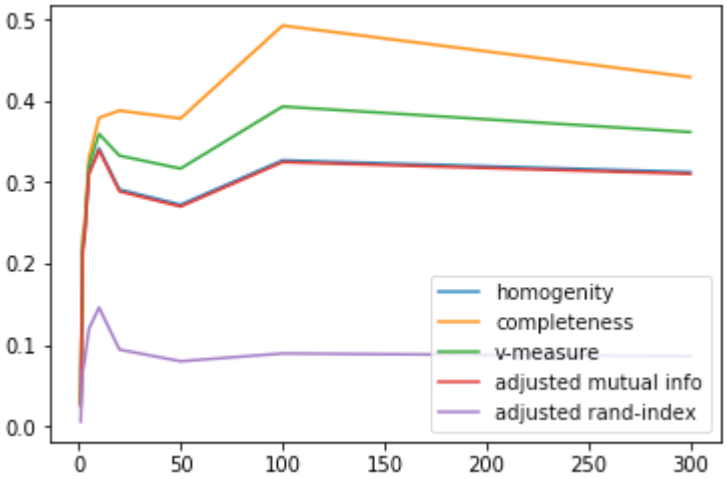
[15 0 0 1 0 112 270 0 3 570 4 0 0 19 0 0 0 0 3 0]

[12 0 5 0 263 154 157 3 310 1 0 0 0 0 0 0 0 0 4 1]

[3 0 1 0 0 58 224 0 0 4 67 0 0 0 184 398 1 0 0 0]

[11 129 28 0 14 207 289 2 69 6 0 0 0 0 0 0 0 4 15 1]

[18 0 2 0 5 128 213 0 47 195 3 13 0 0 0 0 0 0 0 4]]



r	Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Adjusted Mutual Info
1	0.028	0.030	0.029	0.006	0.024
2	0.217	0.234	0.225	0.068	0.215
3	0.238	0.247	0.243	0.082	0.236
5	0.312	0.329	0.320	0.120	0.310
10	0.342	0.379	0.359	0.146	0.339
20	0.291	0.388	0.333	0.094	0.289
50	0.272	0.378	0.317	0.080	0.280
100	**0.327**	**0.493**	**0.393**	**0.090**	**0.325**
300	0.313	0.429	0.362	0.086	0.310

Then, in order to find the best performance for the 9 r-values, we drew a graph showing the tendency of each 5 measurement as shown below:

I chose r=100 as the best value for classification since each of the five measures are relatively higher than others at this value.

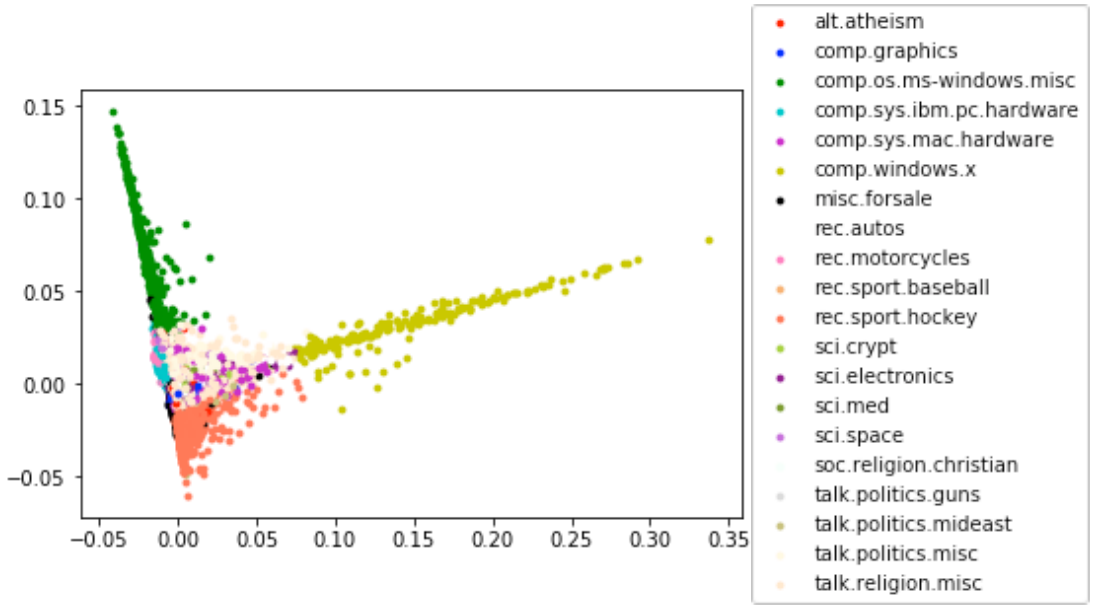
Part4:

To draw the graphs, we used PCA to reduce the dimension into 2D and then used the scatter plot

a.

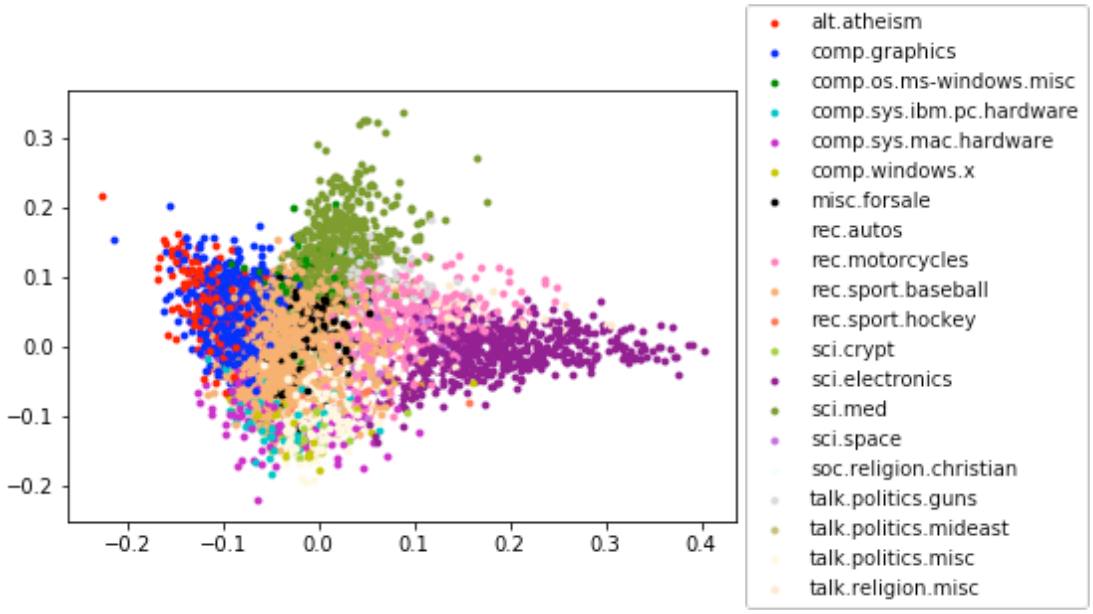
For NMF with r = 20:

the cluster graph is shown as below:



For SVD with r =100

the cluster graph is shown as below:



b.

for the rest of the cases, we used r=20 for nmf and r=100 for svd

NMF norm:

Homogeneity: 0.297

Completeness: 0.379

V-measure: 0.333

Adjusted_Mutual_info_score: 0.294

Adjusted Rand-Index: 0.088

confusion matrix:

[[73 126 0 0 0 0 0 2 278 0 1 2 0 1 151 1 19 137 0 8]
[80 123 0 0 0 0 2 16 383 0 2 1 0 0 1 3 0 0 308 54]
[35 70 10 0 0 0 2 2 185 0 12 0 0 0 0 0 0 0 609 60]
[29 177 190 0 0 2 4 9 298 0 4 0 0 0 0 3 0 0 209 57]
[24 133 74 0 0 0 15 9 549 0 1 0 0 0 0 3 0 0 98 57]
[85 166 0 0 0 0 2 27 259 0 2 0 0 4 0 0 0 0 419 24]
[5 136 53 0 0 14 34 4 567 0 8 3 0 0 0 63 0 0 52 36]
[19 161 0 0 0 0 27 10 205 0 0 3 0 0 0 524 4 0 3 34]
[82 315 0 0 0 0 6 15 217 0 1 0 0 0 0 236 0 0 0 124]
[2 151 0 0 0 388 7 4 389 0 1 4 0 0 0 0 2 0 0 46]
[2 44 0 0 0 606 6 3 160 0 0 0 0 0 0 0 0 0 0 178]
[53 178 0 0 0 0 8 7 159 1 33 16 0 459 0 0 42 0 19 16]

[50 213 4 0 0 1 7 35 476 0 1 0 0 2 0 57 0 0 42 96]

[22 264 0 0 78 0 5 20 530 0 3 2 0 0 3 5 15 0 3 40]

[25 118 0 0 0 0 0 412 286 0 97 1 0 0 0 1 9 0 3 35]

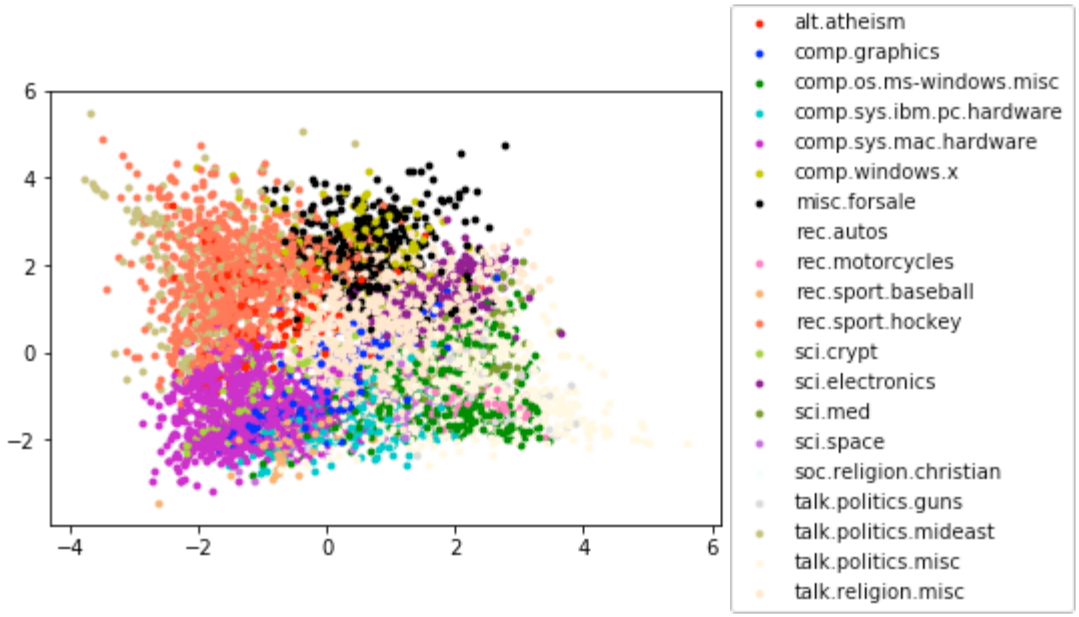
[21 105 0 0 0 0 1 4 315 0 0 0 0 0 518 1 12 0 1 19]

[14 138 0 0 0 0 5 6 131 0 1 69 0 3 0 4 522 0 0 17]

[3 54 0 394 0 0 1 0 248 0 0 0 170 0 3 3 36 0 0 28]

[11 122 0 0 0 0 28 16 222 127 1 5 0 2 4 7 223 0 0 7]

[16 130 0 0 2 1 2 0 218 1 4 0 0 0 153 1 68 15 1 16]]



SVD norm:

Homogeneity: 0.257

Completeness: 0.404

V-measure: 0.314

Adjusted_Mutual_info_score: 0.255

Adjusted Rand-Index: 0.072

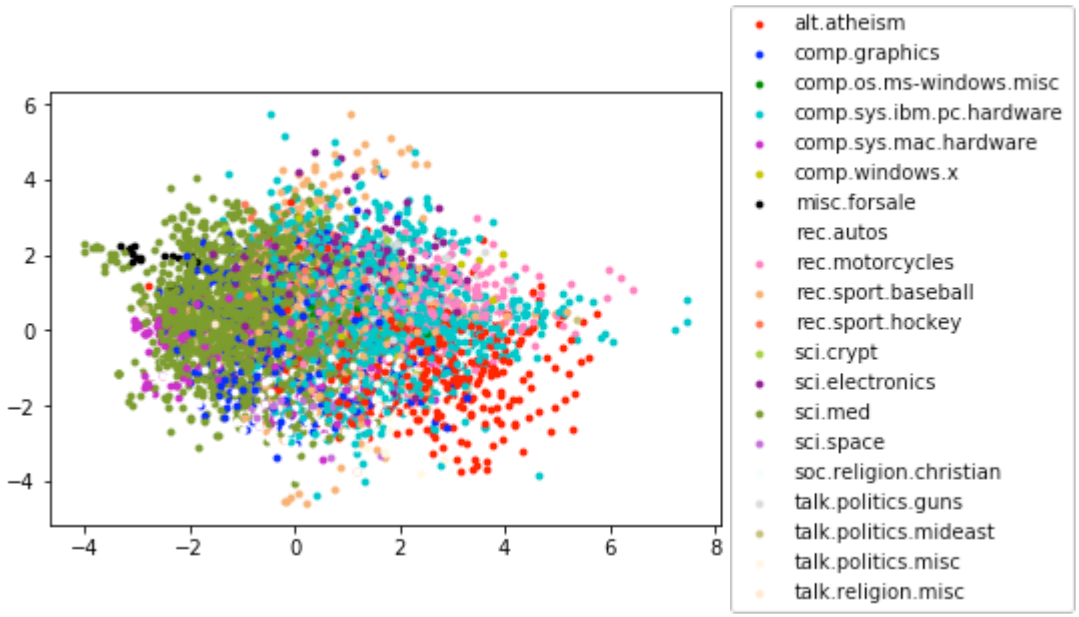
confusion matrix:

[[139 2 0 153 0 0 0 0 4 125 0 0 27 306 0 2 0 0 41 0]

[0 75 1 72 0 4 8 0 1 11 0 0 19 779 0 1 2 0 0 0]

[0 178 0 336 0 12 0 1 2 10 0 0 19 423 2 0 2 0 0 0]

```
[   0 245    0 182    0    2    0    0    0   58    0    0  8 484    0    0    3    0    0    0]
[   0 116    0 168    0    8    0    1    0   33    0    0 21 602    0    0   14    0    0    0]
[   0 383    0   66    0    0   19    3    0    3    0    0  3 509    0    0    2    0    0    0]
[   0   55    0 102    4   10    0    3    1   30    0    0  6 726    0    5   33    0    0    0]
[   0   40    2 153    0    9    0   21    7   11    0    0 47 666    0    3   31    0    0    0]
[   0   22    0   18    0    6    0 670    3   13    0    0  8 250    0    0    6    0    0    0]
[   0   32    0   29   28   13    0    0    0 564    0    0 29 288    0    4    7    0    0    0]
[   0    6    0   24 784    0    0    0    0   20    0    0  2 157    0    0    6    0    0    0]
[   0   15    0 715    0    1    0    0   15    7    1    0  6 206    0   16    9    0    0    0]
[   0   60    3   78    0    2    0    2    0    6    0    0 11 803   13    0    6    0    0    0]
[   2   15    0   33    0    2    0    0    3    8    0    0  6 843   14    2    5    0    0   57]
[   0   57   69 305    0    3    0    0    7   23    0    0 49 473    0    1    0    0    0    0]
[598   12    0   13    0   25    0    0    3    3    0    0 10 313   19    0    0    0    1    0]
[   1   23    0   73    0   19    0    0 483    8    0    0 27 204    0   67    4    0    1    0]
[   2   21    0   34    0    0    0    0  2 339    0   56 31 279    0    0    1 175    0    0]
[   7   11    0 223    1    1    0    0   77    5 131    0 22 263    0    5   29    0    0    0]
[161    7    0   39    0    1    0    0   49   39    0    0 13 246    0    1    2    0   70    0]]
```



NMF log:

In this part, we used bias = 0.01 on the data since the log will occur error on data value 0.

Homogeneity: 0.390

Completeness: 0.430

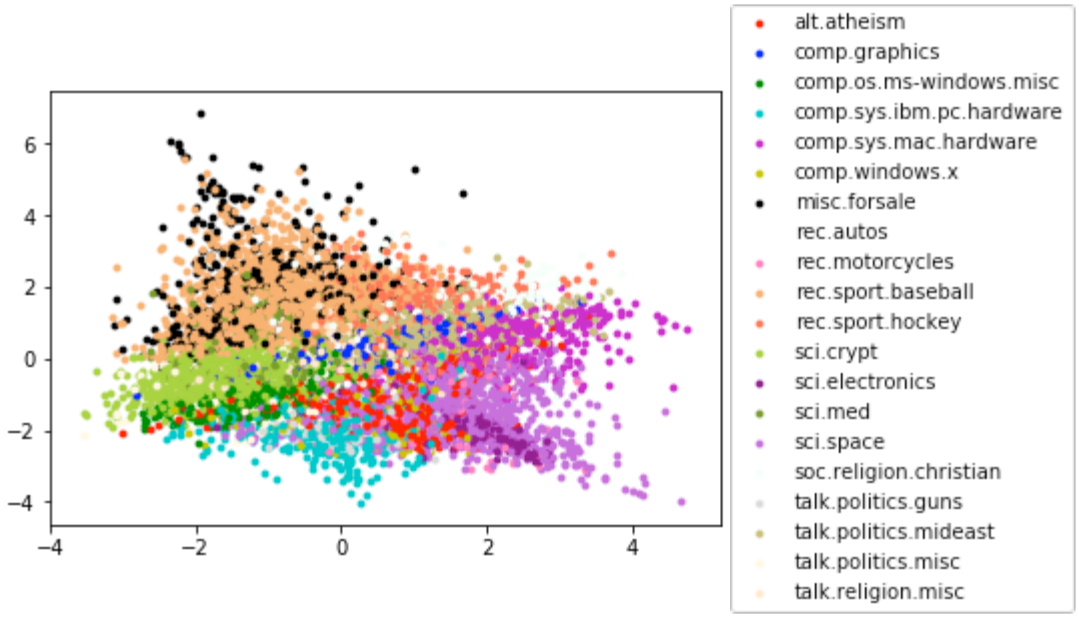
V-measure: 0.409

Adjusted_Mutual_info_score: 0.388

Adjusted Rand-Index: 0.215

confusion matrix:

```
[[279   1   1   4   1 256   0   0   2   0  73  76   1   6  39   3   3  50   0   4]
 [  1   3   0   1   2  13   5   1   1 466 108 176   1  51   3   9   0 105   3  24]
 [  0   2   0   0   1   1  20   0   0 694  47  61   0  61   3  32   0  55   3   5]
 [  0  15   2   0   3   4 322   0   0 282  41 109   0  58   0  16   0  93  19  18]
 [  0  19   1   1   2   2 178   2   0 186  35 319   0  67   1  13   0  99  23  15]
 [  0   0   0   0   8   3   1   0   0 580 113 110   0  27   0   6   0 105   4  31]
 [  1 110  19   0   1   2 115   1   7  79   6 391   0  50   3  34   0 109  44   3]
 [  0 643   0   0   0   6   1   0   4   3  28 104   0  37  18   4   0  87  36  19]
 [  6 408   0   0   0   5   6   0   0   0 101 110   0 130  21   5   0 172  12  20]
 [  1  10 552   3   0  18   0   2   5   0   3 213   0  53   7   8   0 103   9   7]
 [  0   1 674   0   0   7   0   4   0   0   2  56   0 199   3   7   0  29  10   7]
 [  0   1   0   0 601   3   0   0  25  30  61  47   0  14  69  70   1  45  13  11]
 [  1 127   0   0   8  12  19   0   3  90  66 284   0 115   3  10   1 179  11  55]
 [  8  17   0   0   0  16   0 116   5  12  31 284   0  48 148  19   0 242   7  37]
 [  1   6   1   0   0   6   2   0   1   5  34 110   0  40  20 149   0  66   2 544]
 [783   2   0   6   0   3   1   0   0   4  28  73   1  12  19   4   6  42   3  10]
 [  1   3   0   3   5  12   0   1  95   0  16  45   0  12 652  12   0  40   7   6]
 [  6   0   0 510   0   1   0   0   4   0   2  65 256  21  56   2   0  15   2   0]
 [ 10   6   1   4   3   5   0   0   6   0   9  82   1   6 359   7 161  51  34  30]
 [280   2   1   6   0  62   0   3   1   1  13  85   0  10  97  10   2  52   2   1]]
```



NMF log-norm:In this part, we used the bias 0.01

In this part, we used the bias 0.01

Homogeneity: 0.383

Completeness: 0.418

V-measure: 0.400

Adjusted_Mutual_info_score: 0.381

Adjusted Rand-Index: 0.211

confusion matrix:

[[5 0 58 36 68 1 253 0 3 8 6 73 2 1 4 277 1 1 1 1]
[3 418 103 3 192 4 1 7 12 61 30 109 1 0 0 20 2 1 1 5]
[0 619 53 4 95 4 0 32 37 79 5 51 0 0 0 1 1 0 0 4]
[0 225 106 0 110 16 0 360 12 55 20 36 0 2 0 5 6 3 0 26]
[1 145 88 1 318 24 0 212 12 67 14 37 0 1 0 8 3 3 0 29]
[0 524 125 0 122 1 0 1 7 35 33 116 0 0 0 4 13 0 0 7]
[0 66 98 3 397 106 1 126 33 50 4 6 7 21 0 3 1 2 0 51]
[0 3 93 14 108 620 0 1 4 42 22 30 7 0 0 6 0 0 0 40]
[0 0 193 17 109 388 6 7 5 130 21 101 0 0 0 6 0 0 0 13]
[3 0 100 4 200 7 0 0 7 47 8 3 5 572 0 21 0 8 0 9]
[0 0 27 1 55 1 0 0 5 160 7 2 0 713 0 9 0 6 0 13]
[0 21 42 48 40 0 0 0 65 12 10 58 26 0 2 3 648 0 0 16]
[0 83 149 1 293 137 2 23 11 119 56 66 3 0 1 13 15 0 0 12]

[0 16 205 161 286 25 8 0 19 50 39 29 5 0 0 16 0 121 0 10]

[0 7 60 22 109 6 1 2 148 39 550 33 1 1 0 6 0 0 0 2]

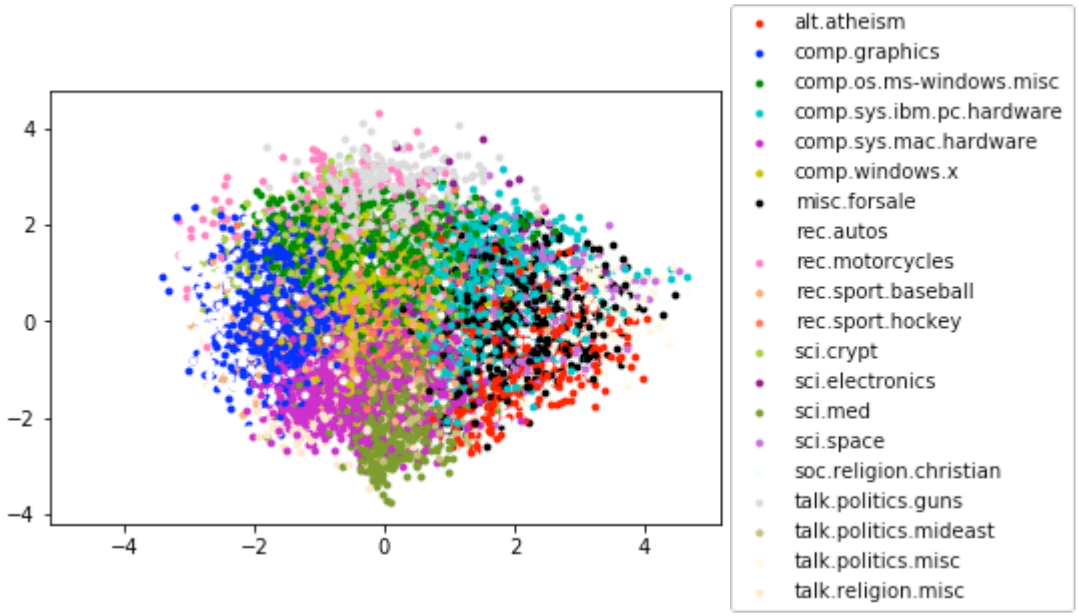
[7 4 39 19 76 2 726 1 4 17 12 41 1 0 34 5 0 0 1 8]

[3 0 64 602 44 3 1 0 16 18 7 20 96 0 0 17 8 3 0 8]

[519 0 17 42 58 0 4 0 2 24 0 2 4 0 0 1 0 0 260 7]

[4 0 65 326 75 6 9 0 7 9 31 8 6 1 178 5 3 0 1 41]

[8 1 62 87 83 2 257 0 11 18 1 11 1 1 7 68 0 4 1 5]]



NMF norm-log:

In this part, we used bias = 1, where we got the value through trying different bias values

Homogeneity: 0.340

Completeness: 0.382

V-measure: 0.359

Adjusted_Mutual_info_score: 0.338

Adjusted Rand-Index: 0.160

confusion matrix:

[[9 213 0 2 1 28 1 1 83 4 141 2 76 56 0 180 0 1 1 0]

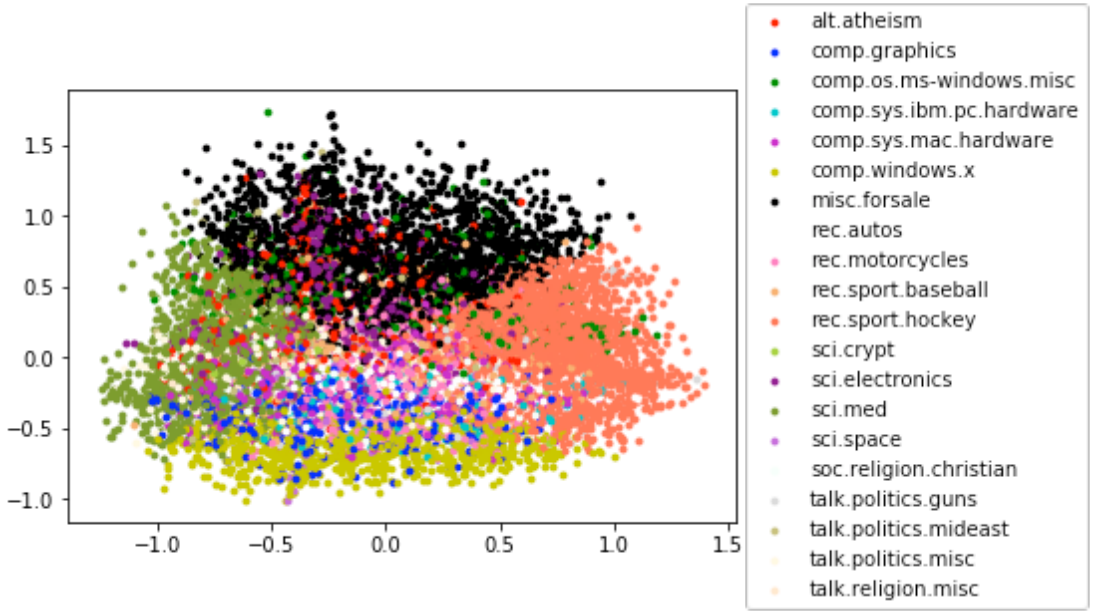
[60 1 2 0 8 2 408 0 31 19 231 0 96 103 0 7 2 2 1 0]

[67 0 20 0 2 1 629 0 20 4 118 0 39 65 0 0 2 17 1 0]

[57 0 256 0 15 0 245 3 28 19 155 0 31 149 0 2 13 8 1 0]

[61 0 132 0 24 0 173 1 44 13 369 0 32 94 0 1 17 2 0 0]

```
[ 32   0   0   0   2   0 493   0  41  30 134   0 110 131   0   2   2   2   9   0]
[ 36   0  82   0 115   0  70  18  27   4 454   0   5 116   0   0  37  11   0   0]
[ 36   0   1   0 567   8   2   0  46  16 137   0  20 119   0   3  34   1   0   0]
[126   0   3   0 321   5   0   0  84  18 134   0  81 215   0   0   8   1   0   0]
[ 47   0   0   0   8   2   0 477  84   5 255   0   2  99   0   6   8   1   0   0]
[178   0   0   0   1   1   1 650  14   5 100   0   2  37   0   0   7   1   0   2]
[ 16   0   0   0   1  51  26   0  46  10  78   1  51  85   0   0  14  45 567   0]
[107   0  13   0 135   1  91   0  57  46 294   0  62 154   0   8   9   3   4   0]
[ 47   6   0   0  40  94  26   0 130  37 283   0  27 180   0   1   5   7   0 107]
[ 42   0   2   0  10  15   8   1  45 503 136   0  30  61   0   1   1 132   0   0]
[ 21 654   1   2   6  17   8   0  22  13 148   5  34  60   0   1   4   1   0   0]
[ 16   1   0   0   3 601   0   0  66   6  77   0  17  99   0   8   6   6   4   0]
[ 34   5   0 457   1  52   1   0  45   0 104   0   3  16 219   0   2   1   0   0]
[  7   7   0   2  10 274   0   1  68  27 118 150   8  61   0   1  36   2   3   0]
[ 17 192   0   2   3  80   0   1  89   1 103   2  12  78   0  39   2   4   1   2]]
```



According to the data, we found that the measurement for only using the logarithm is the best. However, the result of using normalization after logarithm becomes slightly worse, where I used a bias of 0.01 for the data to reach the best value that I can get, but this may not be the best choice among exhaustive test. Further, as we reversed the order of the use of logarithm and normalization, the five measurements show a much worse performance.