# Project 2 : Clustering

Aadithya Venkatanarayanan 404946465
Narendran Raghavan 404945767
Srividhya Balasubramanian 205023825

11th February 2018

## 1 Objective

Clustering is the grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities. This methodology partitions the data implementing a specific join algorithm, most suitable for the desired information analysis.
There are several different ways to implement this partitioning, based on distinct models. The most important ones are: Centralized,Distributed,Connectivity,Group,Graph,Density. The most common clustering method is the centroid-based.In this type of grouping method, every cluster is referenced by a vector of values. Each object is part of the cluster whose value difference is minimal, compared to other clusters.Under centroid based clustering, the popular, notable approach is k-means clustering. It aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.
The goal of this project is to represent the data in an efficient way (i.e) cluster them and produce reasonable results. We also try different preprocess methods to analyze whether there is any increase in performance of the clustering.

## 2 Question 1

TFIDF, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in searches of information retrieval, text mining,user modeling and to determine the importance of a word in the document.The stop words are those that occur too frequent in the document or very rarely and hence are dropped out of the vocabulary. They are generated from the list provided by scikit-learn package. Terms that occur in less than three documents (min_df=3) were removed. Hence this type of representation that does not include too much irrelevant information. TFxIDF vector representation is created using the following definition:

$$TFxIDF(t,d) = tf(t,d) * idf(t)$$

where tf(t,d) represents the frequency of term t in document d.

**O**utput:
Dimensions of Numerical feature vector for training data: (7882, 18469)
Number of terms Extracted for training data: 18469
Dimensions of TF-IDF vector(7882, 18469)

# 3 Question 2a

We apply K-means clustering with k = 2 using the TF-IDF data.In statistics, a contingency table is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables. It is often used to record and analyze the relation between two or more categorical variables.It is also referred to as cross tabulation or cross tab. Now the difference is that Confusion Matrix is used to evaluate the performance of a classifier, and it tells how accurate a classifier is in making predictions about classification, and contingency table is used to evaluate association rules.

**Output**
Contingency Matrix:

$$M = \left(\begin{smallmatrix} 1317 & 2586 \\ 3932 & 47 \end{smallmatrix}\right)$$

# 4 Question 2b

In order to make a concrete comparison of different clustering results, there are various measures of purity for a given partition of the data points with respect to the ground truth. The measures we examine in this project are the homogeneity score, the completeness score, the V-measure, the adjusted Rand score and the adjusted mutual info score.

- Homogeneity is a measure of how "pure" the clusters are. If each cluster contains only data points from a single class, the homogeneity is satisfied
- On the other hand, a clustering result satisfies completeness if all data points of a class are assigned to the same cluster. Both of these scores span between 0 and 1; where 1 stands for perfect clustering
- The V-measure is then defined to be the harmonic average of homogeneity score and completeness score.
- The adjusted Rand Index is similar to accuracy measure, which computes similarity between the clustering labels and ground truth labels. This method counts all pairs of points that both fall either in the same cluster and the same class or in different clusters and different classes
- Finally, the adjusted mutual information score measures the mutual information between the cluster label distribution and the ground truth label distributions.

**Output**
Homogeneity: 0.426
Completeness: 0.464
V-measure: 0.444
Adjusted Rand-Index: 0.432
Adjusted Mutual-Index: 0.426

# 5 Question 3a

High dimensional sparse TF-IDF vectors do not yield a good clustering result because in a high-dimensional space, the Euclidean distance is not a good metric anymore.Also K-means-clustering becomes inefficient 1.when the clusters are not round shaped which will result in inaccurate identification of clusters properly 2.when the clusters have unequal variances.So, we use LSI and NMF for dimensionality reduction.We use

Latent Semantic Indexing (LSI) to minimize mean square residual between original data and reconstruction from its low dimensional approximation. In addition to LSI, the dimensionality is reduced through Non-Negative Matrix Factorization (NMF) and thus the reduced form is used.

**Output:**

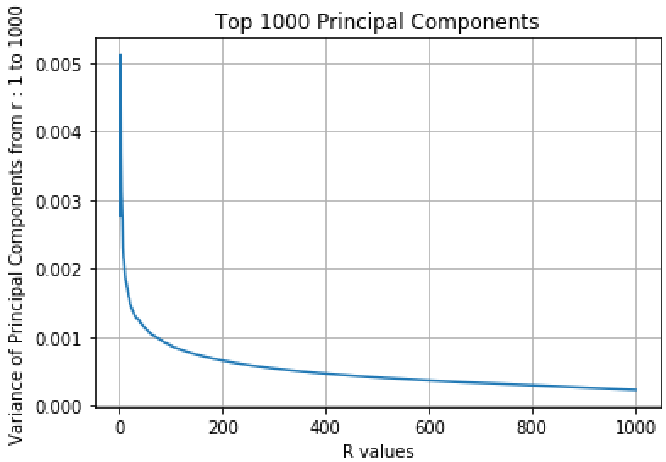
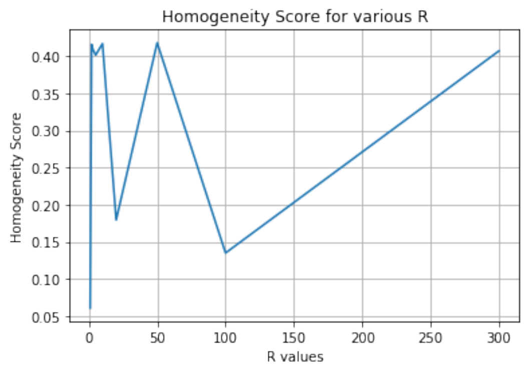



Figure 1: Top 1000 Principal components

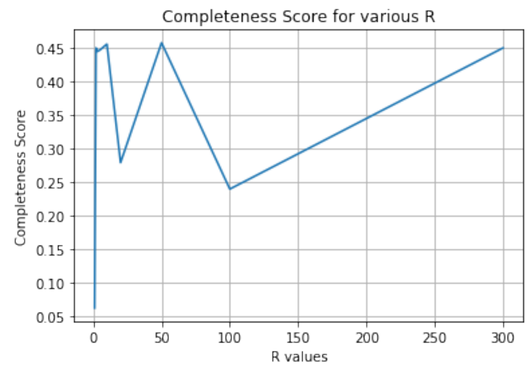
The different metric values that we obtained for various r values are as follows. We have tabulated SVD Truncated Data for various R:

| r | Contingency matrix | Homogeneity | Completeness | V-Measure | Adjusted Rand-Index | Adjusted Mutual Index |
|---|---|---|---|---|---|---|
| 1 | $M = \begin{pmatrix} 1690 & 2213 \\ 2856 & 1123 \end{pmatrix}$ | 0.061 | 0.062 | 0.061 | 0.082 | 0.061 |
| 2 | $M = \begin{pmatrix} 2619 & 1284 \\ 59 & 3920 \end{pmatrix}$ | 0.416 | 0.450 | 0.432 | 0.435 | 0.416 |
| 3 | $M = \begin{pmatrix} 1316 & 2587 \\ 3922 & 57 \end{pmatrix}$ | 0.409 | 0.445 | 0.426 | 0.425 | 0.409 |
| 5 | $M = \begin{pmatrix} 1447 & 2456 \\ 3956 & 23 \end{pmatrix}$ | 0.401 | 0.447 | 0.423 | 0.393 | 0.401 |
| 10 | $M = \begin{pmatrix} 2566 & 1337 \\ 38 & 3941 \end{pmatrix}$ | 0.417 | 0.455 | 0.435 | 0.424 | 0.417 |
| 20 | $M = \begin{pmatrix} 3899 & 4 \\ 2690 & 1289 \end{pmatrix}$ | 0.179 | 0.279 | 0.218 | 0.100 | 0.179 |
| **50** | $M = \begin{pmatrix} 2559 & 1344 \\ 34 & 3945 \end{pmatrix}$ | 0.418 | 0.457 | 0.437 | 0.423 | 0.418 |
| 100 | $M = \begin{pmatrix} 9 & 3894 \\ 1034 & 2945 \end{pmatrix}$ | 0.135 | 0.239 | 0.173 | 0.063 | 0.135 |
| 300 | $M = \begin{pmatrix} 2499 & 1404 \\ 29 & 3950 \end{pmatrix}$ | 0.407 | 0.450 | 0.427 | 0.405 | 0.407 |

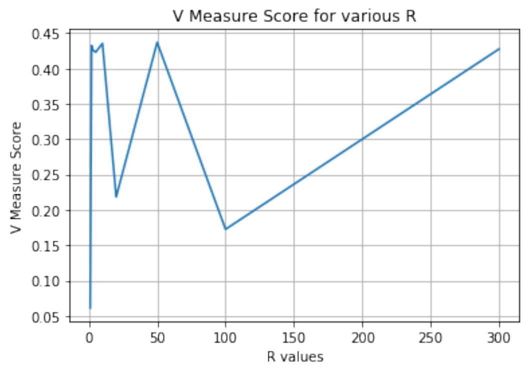
SVD Truncated Data for various R:



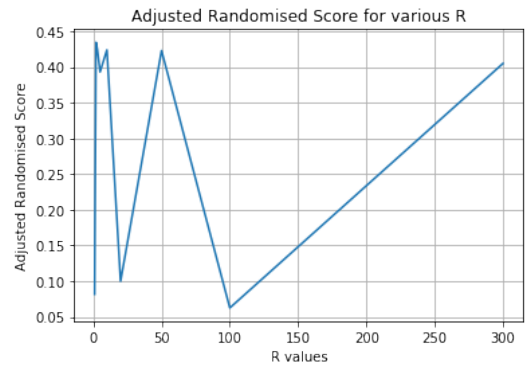

(a) Homogeneity Score



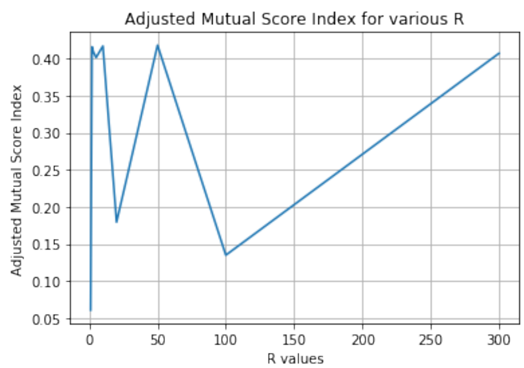

(b) Completeness




(c) V-Measure score




(d) Adjusted Randomised Score index

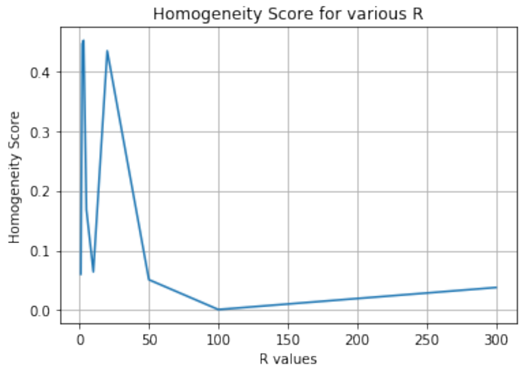



(e) Adjusted mutual Score index

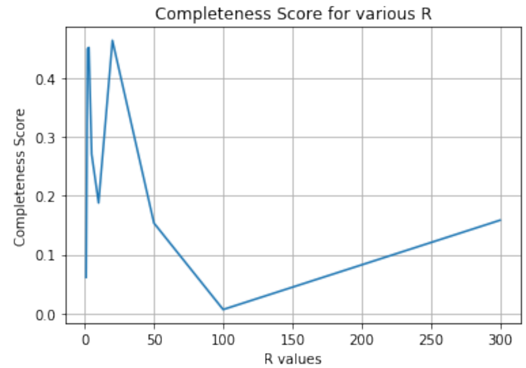
The different metric values that we obtained for various r values are as follows. We have tabulated NMF Truncated Data for various R:

| r | Contingency matrix | Homogeneity | Completeness | V-Measure | Adjusted Rand-Index | Adjusted Mutual Index |
|---|---|---|---|---|---|---|
| 1 | $M = \begin{pmatrix} 2225 & 1678 \\ 1139 & 2840 \end{pmatrix}$ | 0.060 | 0.061 | 0.061 | 0.081 | 0.060 |
| 2 | $M = \begin{pmatrix} 3633 & 270 \\ 793 & 3186 \end{pmatrix}$ | 0.446 | 0.451 | 0.448 | 0.533 | 0.446 |
| 3 | $M = \begin{pmatrix} 3531 & 372 \\ 640 & 3339 \end{pmatrix}$ | 0.452 | 0.453 | 0.452 | 0.552 | 0.452 |
| 5 | $M = \begin{pmatrix} 3900 & 3 \\ 2765 & 1214 \end{pmatrix}$ | 0.168 | 0.271 | 0.208 | 0.088 | 0.168 |
| 10 | $M = \begin{pmatrix} 3405 & 498 \\ 3976 & 3 \end{pmatrix}$ | 0.064 | 0.188 | 0.096 | 0.018 | 0.064 |
| 20 | $M = \begin{pmatrix} 1208 & 2695 \\ 3918 & 61 \end{pmatrix}$ | 0.434 | 0.465 | 0.449 | 0.460 | 0.434 |
| **50** | $M = \begin{pmatrix} 465 & 3438 \\ 17 & 3962 \end{pmatrix}$ | 0.051 | 0.154 | 0.077 | 0.015 | 0.051 |
| 100 | $M = \begin{pmatrix} 3783 & 120 \\ 3904 & 75 \end{pmatrix}$ | 0.001 | 0.006 | 0.002 | 0.000 | 0.001 |
| 300 | $M = \begin{pmatrix} 307 & 3596 \\ 3 & 3976 \end{pmatrix}$ | 0.038 | 0.159 | 0.061 | 0.007 | 0.038 |

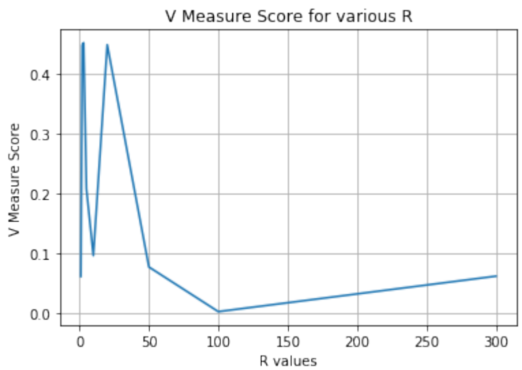
NMF Truncated Data for various R:



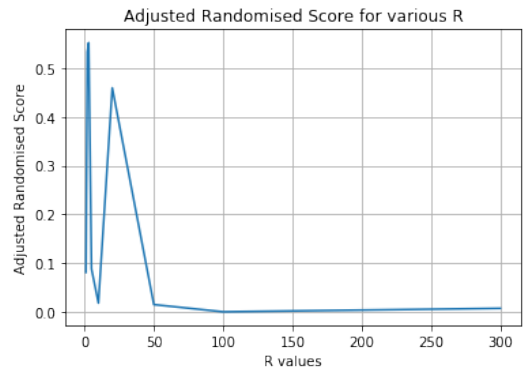

(a) Homogeneity Score



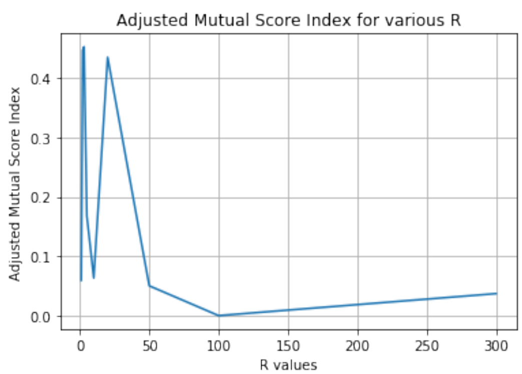

(b) Completeness




(c) V-Measure score



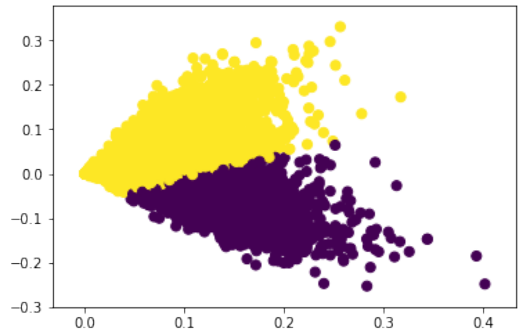

(d) Adjusted Randomised Score index



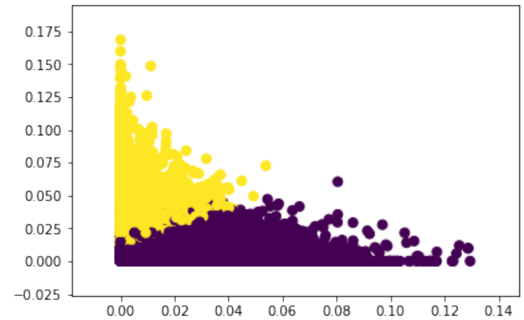

(e) Adjusted mutual Score index

# 6 Question 4a

The plots for Visualization of K-Means clustering:




(a) Visualization of K-Means clustering for 2 clusters using r = 2 obtained in LSI




(b) Visualization of K-Means clustering for 2 clusters using r = 3 obtained in NMF

# 7 Question 4b

**Output:**
**Normalizing data for r = 2 with LSI**
Contingency Matrix:

$$M = \left(\begin{smallmatrix} 2620 & 1283 \\ 59 & 3920 \end{smallmatrix}\right)$$

Homogeneity: 0.416
Completeness: 0.450
V-measure: 0.432
Adjusted Rand-Index: 0.435
Adjusted Mutual-Index: 0.416

**Normalizing data for r = 3 with NMF**
Contingency Matrix:

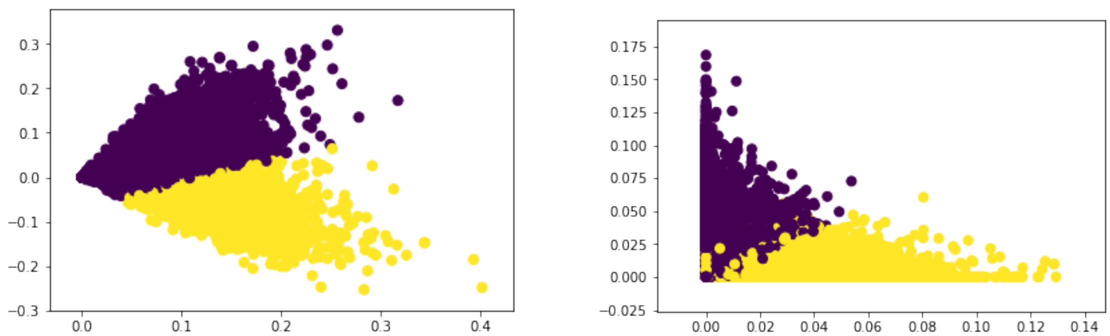
$$M = \left(\begin{smallmatrix} 954 & 2949 \\ 3862 & 117 \end{smallmatrix}\right)$$

Homogeneity: 0.470
Completeness: 0.488
V-measure: 0.479
Adjusted Rand-Index: 0.530
Adjusted Mutual-Index: 0.470

**Log Transformation using r = 3**
Contingency Matrix:

$$M = \left(\begin{smallmatrix} 3213 & 690 \\ 177 & 3802 \end{smallmatrix}\right)$$

Homogeneity: 0.520
Completeness: 0.528
V-measure: 0.524
Adjusted Rand-Index: 0.608
Adjusted Mutual-Index: 0.520




(a) Normalizing data for r = 2 with LSI

(b) Normalizing data for r = 3 with NMF

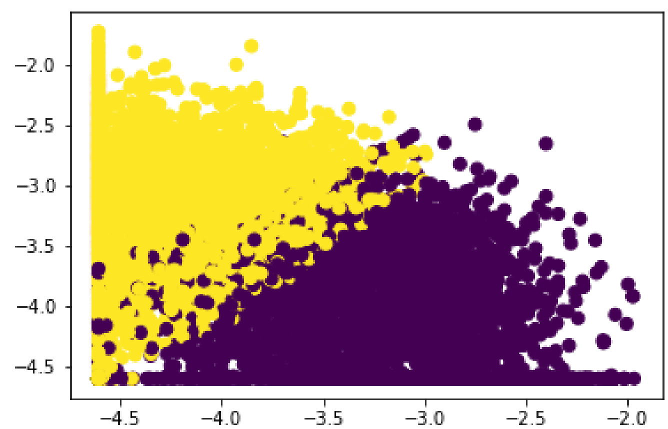



Figure 2: Log Transformation using r = 3

**Normalizing and then taking log transform of NMF reduced data**
Contingency Matrix:

$$M = \begin{pmatrix} 988 & 2915 \\ 3870 & 109 \end{pmatrix}$$

Homogeneity: 0.484
Completeness: 0.485
V-measure: 0.485
Adjusted Rand-Index: 0.591
Adjusted Mutual-Index: 0.484

**Taking log transform and then normalizing of NMF reduced data**
Contingency Matrix:

$$M = \begin{pmatrix} 988 & 2915 \\ 3870 & 109 \end{pmatrix}$$

Homogeneity: 0.484
Completeness: 0.485
V-measure: 0.484
Adjusted Rand-Index: 0.590
Adjusted Mutual-Index: 0.484

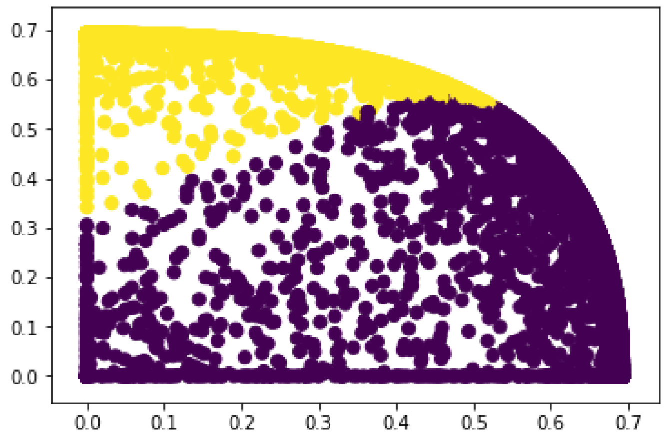



Figure 3: Normalizing and then taking log transform of NMF reduced data

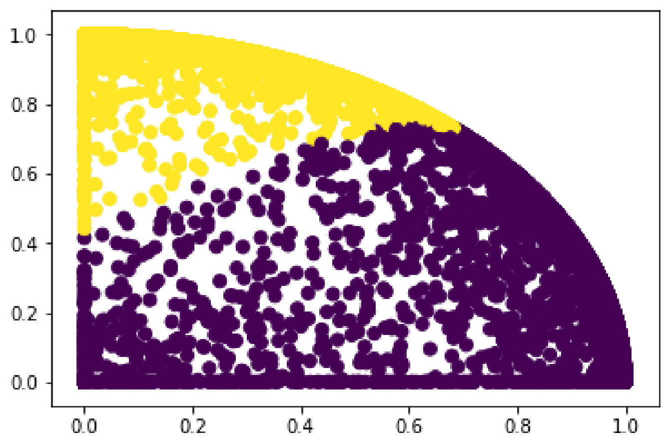



Figure 4: Taking log transform and then normalizing of NMF reduced data

# 8 Question 5

In this part we want to examine how purely we can retrieve all 20 original sub-class labels with clustering. Therefore, we need to include all the documents and the corresponding terms in the data matrix and proper representation through dimensionality reduction of the TF-IDF representation.

**Output:**

**For getting TFIDF matrix and printing its dimensions**

Dimensions of Numerical feature vector for training data: (7882, 18469)

Number of terms Extracted for training data: 18469

Dimensions of TF-IDF vector(7882, 18469)

**Finding best r value with LSI for 20 clusters**

| r | Homogeneity | Completeness | V-Measure | Adjusted Rand-Index | Adjusted Mutual Index |
|---|---|---|---|---|---|
| 1 | 0.086 | 0.021 | 0.034 | 0.010 | 0.021 |
| 2 | 0.615 | 0.149 | 0.240 | 0.076 | 0.148 |
| 3 | 0.582 | 0.145 | 0.232 | 0.080 | 0.145 |
| 4 | 0.589 | 0.147 | 0.235 | 0.074 | 0.146 |
| 5 | 0.601 | 0.152 | 0.243 | 0.078 | 0.152 |
| 6 | 0.568 | 0.146 | 0.232 | 0.075 | 0.146 |
| 7 | 0.547 | 0.144 | 0.228 | 0.072 | 0.143 |
| 8 | 0.609 | 0.158 | 0.250 | 0.084 | 0.157 |
| 9 | 0.600 | 0.159 | 0.251 | 0.090 | 0.158 |
| 10 | 0.584 | 0.155 | 0.245 | 0.087 | 0.155 |

Contingency Matrix :
For r=1
[[263 85 101 280 241 330 141 203 185 322 37 305 74 155 138 286 274 222 257 4]
[309 18 300 106 380 241 221 51 365 224 3 271 178 26 334 158 347 375 72 0]]
For r=2
[[93 236 135 1 411 133 0 61 11 3 449 18 480 0 185 522 158 232 465 310]
[666 0 191 396 87 561 288 0 465 487 137 160 5 131 0 12 3 374 16 0]]
For r=3
[[456 48 125 298 288 110 41 290 1 160 348 580 382 39 147 0 232 260 97 1]
[12 465 850 59 4 1 86 408 589 0 0 122 69 0 657 179 9 0 0 469]]
For r=4
[[231 182 244 0 490 286 83 254 54 435 3 1 42 119 286 4 588 195 400 6]
[17 662 0 161 25 10 506 648 0 24 218 369 0 1 415 251 48 0 0 624]]
For r=5
[[46 4 359 10 699 154 90 394 1 231 92 214 371 0 428 81 204 248 0 277]
[604 303 8 639 92 0 1 14 353 863 0 0 67 194 650 0 17 0 173 1]]
For r=6
[[361 500 167 87 519 151 1 1 11 197 197 1 200 202 168 55 281 147 547 110]
[3 788 0 0 5 3 283 402 666 0 0 496 160 946 0 0 49 29 149 0]]
For r=7
[[319 24 370 220 88 1 4 2 79 569 610 194 253 455 122 10 261 90 145 87]
[3 453 206 17 0 258 81 445 0 5 839 0 49 10 0 650 962 1 0 0]]
For r=8
[[615 6 151 220 33 244 255 155 121 1 3 402 200 529 196 143 123 349 153 4]
[16 645 0 0 0 0 208 1021 0 272 428 10 15 936 0 53 0 3 2 370]]
For r=9
[[268 512 3 316 257 202 161 87 626 156 257 1 71 155 153 1 4 19 261 393]
[192 1131 795 2 0 16 923 0 22 0 0 423 1 0 58 312 93 0 8 3]]
For r=10
[[364 135 273 4 199 192 31 595 1 413 303 1 107 2 126 172 136 130 169 550]
[179 813 3 94 3 15 0 36 338 9 0 399 0 790 0 0 0 0 52 1248]]

### Finding best r value with NMF for 20 clusters

| r | Homogeneity | Completeness | V-Measure | Adjusted Rand-Index | Adjusted Mutual Index |
|---|---|---|---|---|---|
| 1 | 0.086 | 0.022 | 0.034 | 0.011 | 0.021 |
| 2 | 0.635 | 0.154 | 0.248 | 0.086 | 0.154 |
| 3 | 0.599 | 0.155 | 0.246 | 0.106 | 0.155 |
| 4 | 0.558 | 0.141 | 0.225 | 0.080 | 0.140 |
| 5 | 0.560 | 0.142 | 0.227 | 0.077 | 0.142 |
| 6 | 0.540 | 0.142 | 0.225 | 0.081 | 0.142 |
| 7 | 0.546 | 0.152 | 0.237 | 0.089 | 0.151 |
| 8 | 0.519 | 0.137 | 0.217 | 0.068 | 0.137 |
| 9 | 0.541 | 0.144 | 0.227 | 0.077 | 0.144 |
| 10 | 0.534 | 0.140 | 0.221 | 0.067 | 0.139 |

Contingency Matrix :
For r=1
[[319351141371412951573222327420187230251299329433110182]
[4052222216342813029038617823805838413931802243001З]]
For r=2
[[27810314057715168760231101311341427921612910501]
[1776180151294686668134355250870294672013442 2]]
For r=3
[[273883724069186401465952821899254304661514128 3209]
[23915440146959560058515010981319141682210]]
For r=4
[[1811462019023861223036350416926605735963210919 5649]
[21847040107967781054154911272011004613 3121]]
For r=5
[[240179952443406401018219194221196028113216 5589374]
[735300480570913593341005014543554212155 22]]
For r=6
[[46747366078347211221344172157615215040219602 0815436]
[224902856103028472872400013952409035]]
For r=7
[[614013818386010345256228256696150169339254076 3135271]
[86020320420961131051381055230016120]]
For r=8
[[237291300211245903120615612749412514581325 941251142]
[457788145420500590030231116849242131 47]]
For r=9
[[236255894499102141624393203242181245612135 453159104]
[9820081074086021231491038159050115 40]]
For r=10
[[667270389001226739613292017011431322465561602628]
[114743433517101840144255200735017171243582]]

### Visualization of K-Means clustering for 20 clusters

### Normalizing data

Normalizing data for r = 9 with LSI
Contingency Matrix:
[[616 333 545 175 3 95 160 1 31 1 434 116 174 130 374 14 246 307 1 147]

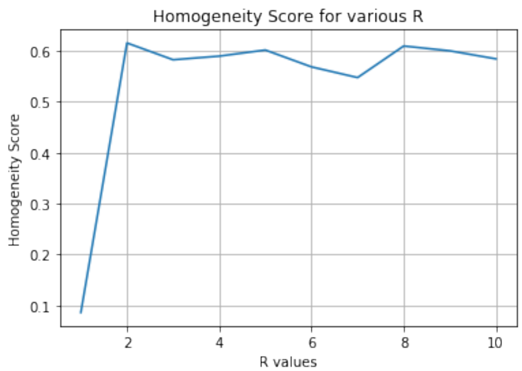
[20 2 1122 18 722 2 930 314 0 407 6 0 53 0 154 174 0 0 55 0]]
Homogeneity: 0.590
Completeness: 0.156
V-measure: 0.247
Adjusted Rand-Index: 0.086
Adjusted Mutual-Index: 0.156

Normalizing data for r = 3 with NMF
Contingency Matrix:
[[327 39 88 499 6 290 32 554 0 40 154 290 141 230 182 16189 58220143]
[425 91 1 4 413 13 867 14 152 0 14 19 7 187 900 766 0 106 0 0]]
Homogeneity: 0.595
Completeness: 0.153
V-measure: 0.243
Adjusted Rand-Index: 0.097
Adjusted Mutual-Index: 0.152

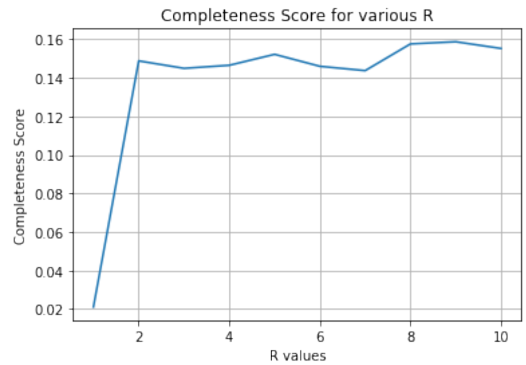
**Log Transformation**
Contingency Matrix:
[[32 504 2 462 190 277 209 314 105 88 1 443 228 147 28 228 172 33 153 287]
[158 3 840 35 0 33 276 0 299 343 554 18 16 12 793 0 215 307 11 66]]
Homogeneity: 0.630
Completeness: 0.151
V-measure: 0.244
Adjusted Rand-Index: 0.089
Adjusted Mutual-Index: 0.151

**Normalizing and then taking log transform of NMF reduced data**
Contingency Matrix:
[[8 62194 211 28 34 168 344 193 321 327 124 41 613 13 143 64 0 62 494]
[457 17 3 841 832 0 0 99 0 9 411 4 94 92 0 16 931 168 1 4]]
Homogeneity: 0.635
Completeness: 0.167
V-measure: 0.265
Adjusted Rand-Index: 0.182
Adjusted Mutual-Index: 0.167

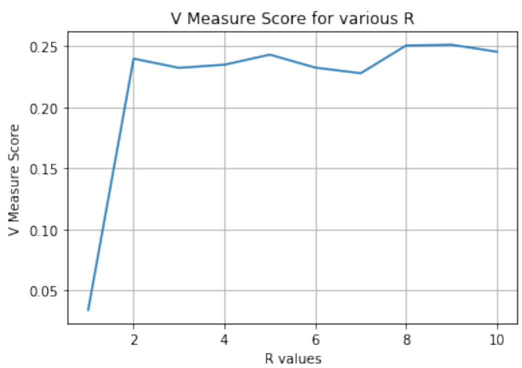
**Taking log transform and then normalizing of NMF reduced data**
Contingency Matrix:
[[8 621 94 211 28 34 168 344 193 321 327 124 41 613 13 143 64 0 62 494]
[457 17 3 841 832 0 0 99 0 9 411 4 94 92 0 16 931 168 1 4]]
Homogeneity: 0.628
Completeness: 0.169
V-measure: 0.266
Adjusted Rand-Index: 0.197
Adjusted Mutual-Index: 0.168



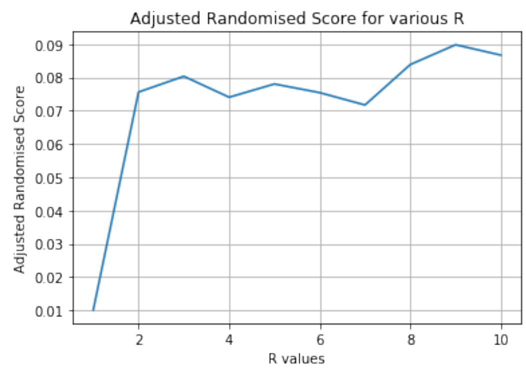

(a) Homogeneity Score



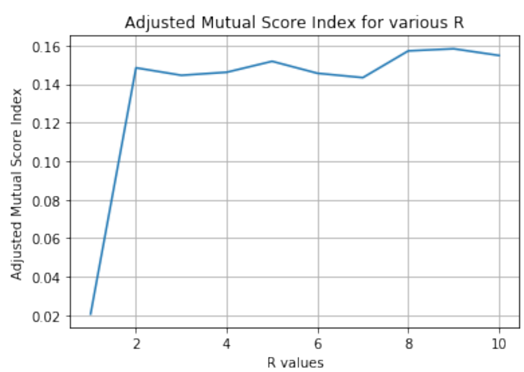

(b) Completeness



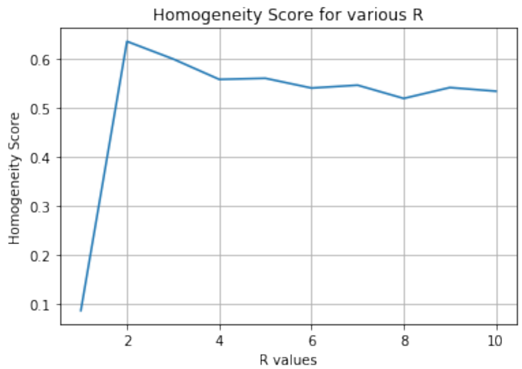

(c) V-Measure score



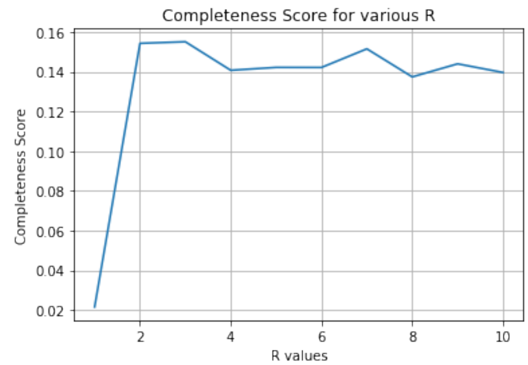

(d) Adjusted Randomised Score index



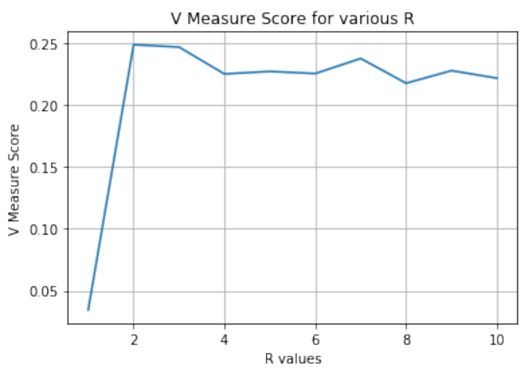

(e) Adjusted mutual Score index



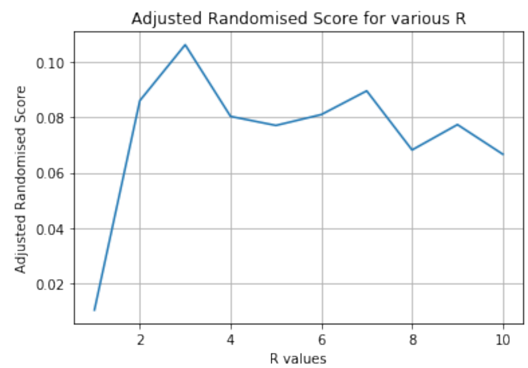

(a) Homogeneity Score



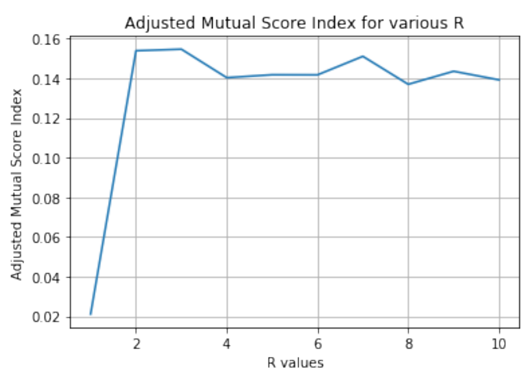

(b) Completeness



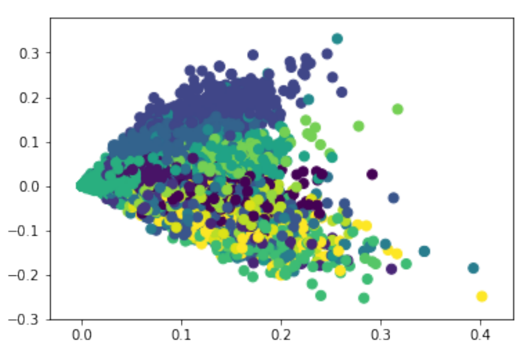

(c) V-Measure score



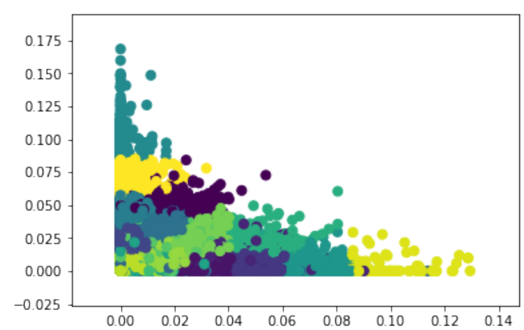

(d) Adjusted Randomised Score index



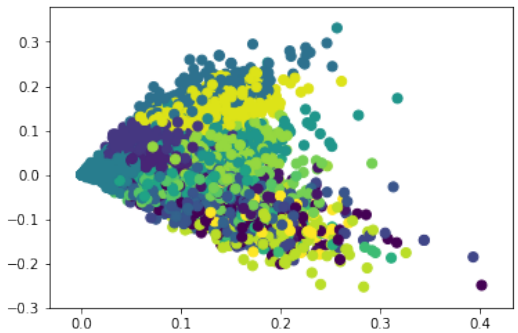

(e) Adjusted mutual Score index



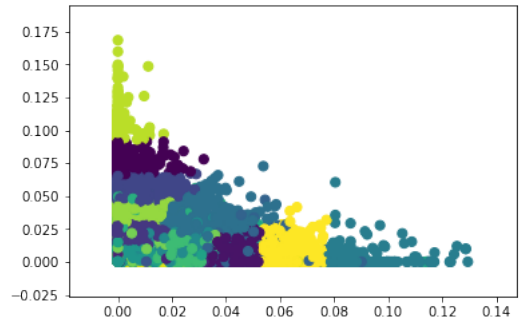

(a) Visualization of K-Means clustering for 20 clusters using r = 9 obtained in LSI



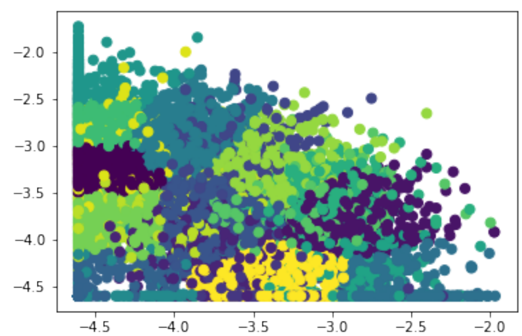

(b) Visualization of K-Means clustering for 20 clusters using r = 3 obtained in NMF



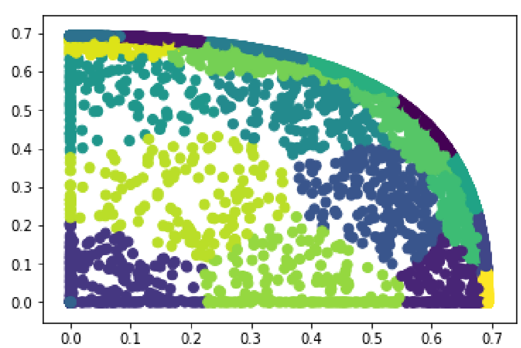
(a) Normalizing data for r = 9 with LSI



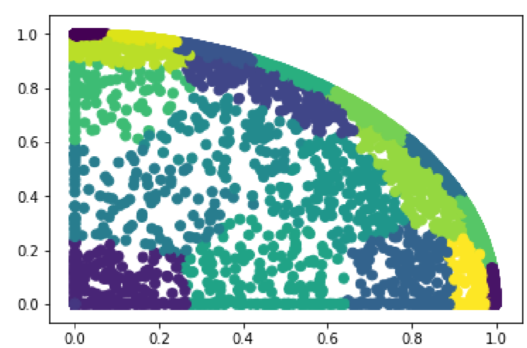
(b) Normalizing data for r = 3 with NMF



(a) Log Transformation



(a) Normalizing and then taking log transform of NMF reduced data for r = 3



(b) Taking log transform and then normalizing of NMF reduced data r=3