EE219 Project 2
Clustering
Winter 2018

Group Members:
Asavari Limaye UID: 605224431
Pooja Janagal Nagaraja UID: 405222664
Rupa Mahadevan UID: 005225216
Vaishnavi Ravindran UID: 805227216

1. INTRODUCTION

In this project we deal with clustering of textual data. A cluster refers to a collection of data points aggregated together because of certain similarities. We define a target number k, which refers to the number of clusters we need to make from a dataset. A centroid is the imaginary or real location representing the center of the cluster. Every data point is allocated to each of the clusters by reducing the in-cluster sum of squares. In other words, the K-means algorithm identifies k centroids, and then allocates every data point to the cluster whose centroid is closest to it, with the end goal of making the clusters with as small a spread as possible. The 'means' in the K-means refers to averaging of the data; that is, finding the centroid.

We do the following in this project:

1. We find the proper representations of the provided data, such that the clustering is efficient and gives us clusters close to the real labels of the data.
2. We perform K-means clustering on the dataset, and evaluate the performance of the clustering.
3. We try different preprocessing methods which may improve the performance of the clustering.

2. DATASET

For our dataset, we use scikit-learn's "20 Newsgroups" dataset. This is a collection of over 20K documents that are partitioned across 20 different topics/categories. For questions 1-10, we consider just data from two categories. Later on, we extend our procedure to all 20 categories.

Since the dataset consists of words, these have to be transformed into a representation that can be handled by the k-means algorithm. For this, the dataset is converted into its Term-Frequency Inverse-Document Frequency (TF-IDF) representation. TF-IDF is a vector of all the words in the vocabulary or corpus, where each dimension represents the term frequency of that word in a given document that is scaled appropriately by its IDF, that measures how important the term is to a given document. A term that appears more frequently in a document but less frequently in the remaining documents is given a high weight because this term is considered important and specific to that particular document.

QUESTION 1: Report the dimensions of the TF-IDF matrix you get.

Using skelarn's inbuilt methods, the frequencies of all terms is computed using CountVectorizer. This is fed as an input to the tf-idf method that generates the tf-idf vectors for a given dataset.

The shape of the resulting matrix is: **(7882, 27768)**

In our case, we consider two classes of documents: one belonging to "Computational technology" and the other belonging to "Recreational technology". K-means clustering is performed on these documents using the TF-IDF vectors mentioned previously to group them into two clusters. The algorithm is trained for 1000 iterations. The algorithm is also trained for 30 different initial centroid values. Once the clusters have been determined, the provided ground truth labels of the data-points are compared with the classes predicted by the k-means algorithm. In the following section, we see the metrics used to measure the performance of the clustering algorithm.

QUESTION 2: Report the contingency table of your clustering result.

The contingency table for k-means is a representation of the number of documents that belong to a true class vs the cluster that k-means has classified a document as.

The contingency matrix obtained was:

```
[4, 3899],
[1718, 2261]
```

QUESTION 3: Report the metrics for the clustering result.

There are other metrics that measure how well the clusters formed by the k-means algorithm classifies the data. In this section, we discuss those metrics.

1. Homogeneity score: This is a measure of how pure the clusters formed by the algorithm are. The score is high when all the clusters have data points belonging to just one class.
2. Completeness score: This score indicates whether all the data points belonging to one class have been predicted to belong to the same cluster.
3. V-measure: This is a harmonic mean between the homogeneity score and completeness score.

4. Adjusted Rand Index Score: This score is a measure of how well the k-means algorithm performs. This metric counts all pairs of points that belong to the same cluster and the same true class. It also counts all those pairs of points that belong to different classes and have been placed in different clusters. The score returned is a value between 0-1.
5. Adjusted Mutual Information Index: This score measures the similarity between the actual class or labels and the clusters created.

Metric	Value
Homogeneity Score	0.2535958928926043
Completeness Score	0.334815748824373
V-measure	0.28860033608397917
Adjusted Rand Index Score	0.18076179588914554
Adjusted Mutual Information Score	0.25352755133060884

Table 1 : Different metrics to measure the performance of K-means clustering algorithm using TF-IDF vectors

From Table 1, it can be seen that the k-means algorithm has performed quite poorly and has scores in the range of 0.18 to 0.34. The performance of the algorithm can definitely be improved.

The TF-IDF vector used above has a very high dimension because of which the performance of the clustering algorithm is poor. In a very high dimensional space, Euclidean distance is almost same for all the points and hence it is difficult to obtain good clustering results.

To overcome this issue, dimensionality reduction techniques could be applied on the high dimensional TF-IDF vectors. In this section, two dimensionality reduction techniques, namely, SVD and NMF have been considered.

It is important to examine the number of dimensions that the TF-IDF vectors have to be reduced to because too low a dimension could result in loss of information that impacts the performance of the algorithm again. Hence, in the following section, the right dimension is determined.

QUESTION 4: Report the plot of percent variance for the top r principal components.

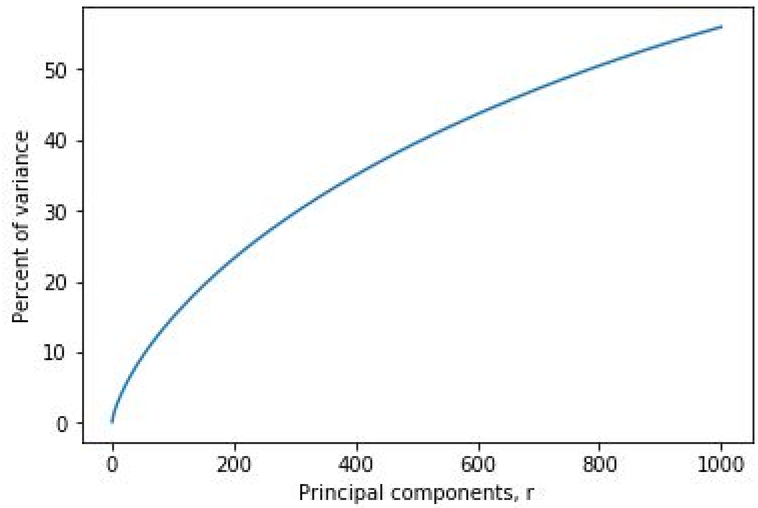


Fig 1: Percent of variance retention vs number of principal components, r

On reducing the dimensions of the TF-IDF vectors to r dimensions, whose range is from 1 to 1000, only a fraction of the variance from the original data is retained. Figure 1 is a plot of the percent of variance retained by the top principal components for SVD.

Out of these 1000 values for r, 9 values have been chosen and k-means clustering algorithm is run for each of these lower dimensional representation. The scores for the clusters created are measured. The value of r for which the best score is obtained is used as the dimension henceforth.

QUESTION 5: Report the best r for SVD and NMF.

The scores generated for different values of r for SVD have been reported in Table 2

Dimensions (r value)	Homogeneity Score	Completeness Score	V-measure	Adjusted Rand Index Score	Adjusted Mutual Information Score
1	0.0003003030178761853	0.0003047688479979988	0.0003025194525487269	0.00033904080274629444	0.00020877974994934834
2	0.5928445154123904	0.6080671630362278	0.6003593587733158	0.648591716893542	0.5928072398752575
3	0.2375614248617169	0.31709966233910336	0.2716276636192888	0.16950318518005686	0.23749161477753092

5	0.18721505194147797	0.2837477581769007	0.22558830610006753	0.10854391112052623	0.18714062680026902
10	0.0954692068245713	0.2183726563458545	0.13285585347288345	0.03563509228205547	0.09538635368498467
20	0.09193784413983222	0.21533669061454289	0.12885930241588447	0.033369850303460145	0.09185466550421705
50	0.06985526414894397	0.1954038397611516	0.10291814034672729	0.020672823875375657	0.06977004503487678
100	0.07239132259060248	0.1959825474777242	0.10572889091606381	0.02238895648403063	0.07230633938157532
300	0.014691214289954626	0.1356024532531799	0.026510294565993825	0.0013732921152795757	0.014600622181283582

Table 2 : Different metrics to measure the performance of K-means clustering algorithm using different reduced dimensions of data, r

The scores generated for different values of r for NMF have been reported in Table 3

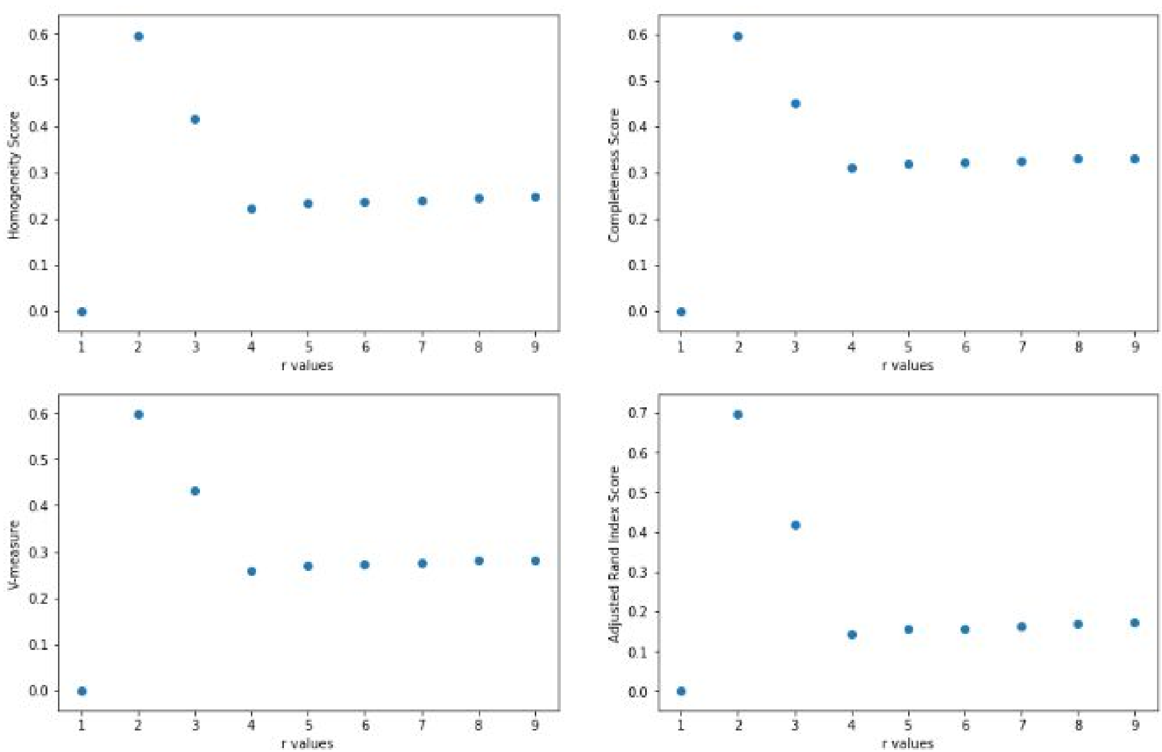
Dimensions (r value)	Homogeneity Score	Completeness Score	V-measure	Adjusted Rand Index Score	Adjusted Mutual Information Score
1	0.0003003030178761853	0.0003047688479979988	0.0003025194525487269	0.00033904080274629444	0.00020877974994934834
2	0.5928445154123904	0.6080671630362278	0.6003593587733158	0.648591716893542	0.5928072398752575
3	0.2375614248617169	0.31709966233910336	0.2716276636192888	0.16950318518005686	0.23749161477753092
5	0.18721505194147797	0.2837477581769007	0.22558830610006753	0.10854391112052623	0.18714062680026902
10	0.0954692068245713	0.2183726563458545	0.13285585347288345	0.03563509228205547	0.09538635368498467
20	0.09193784413983222	0.21533669061454289	0.12885930241588447	0.033369850303460145	0.09185466550421705

50	0.069855264 14894397	0.195403839 7611516	0.102918140 34672729	0.020672823 875375657	0.069770045 03487678
100	0.072391322 59060248	0.195982547 4777242	0.105728890 91606381	0.022388956 48403063	0.072306339 38157532
300	0.014691214 289954626	0.135602453 2531799	0.026510294 565993825	0.001373292 1152795757	0.014600622 181283582

Table 3 : Different metrics to measure the performance of K-means clustering algorithm using different reduced dimensions of data, r

Clearly, from the above two tables, it can be seen that k-means clustering performs the best when r value is set to 2.

The best r value for SVD : 2
The best r value for NMF : 2



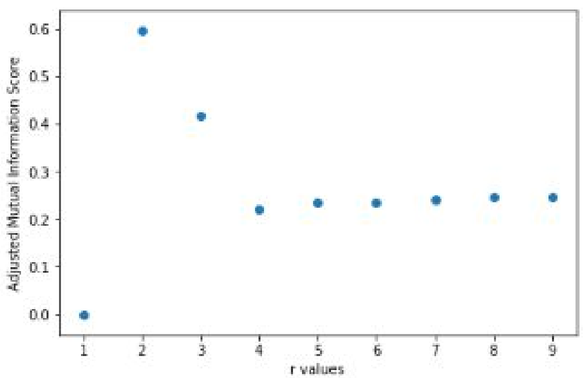
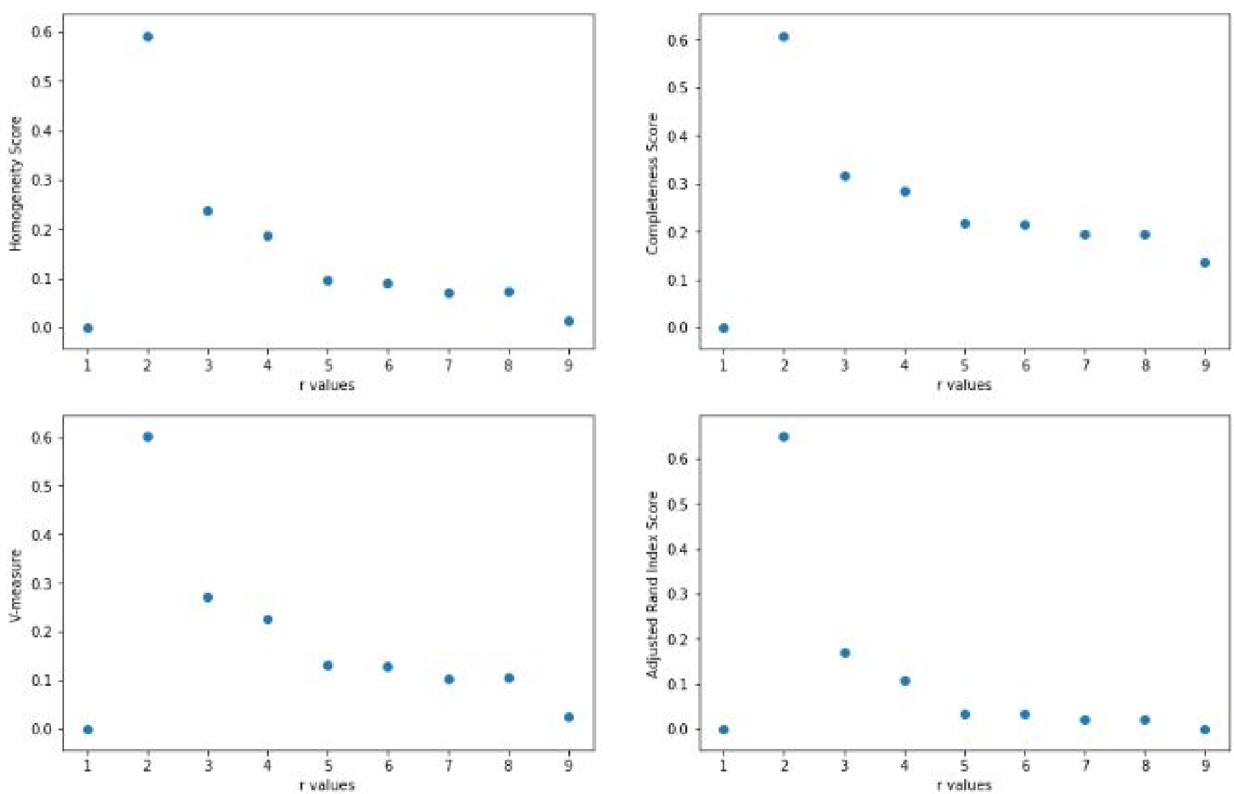


Fig 2: Plot of scores for k-means clustering using different values of r for SVD



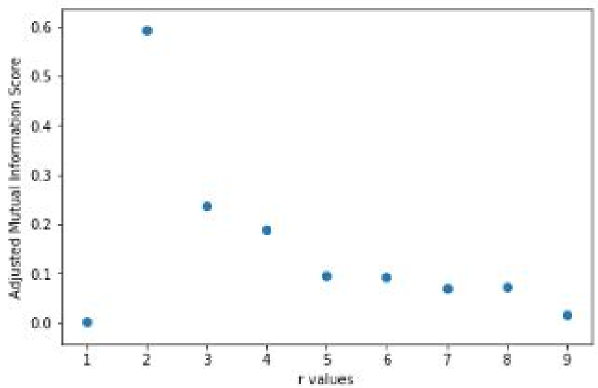


Fig 3 : Plot of scores for k-means clustering using different values of r for NMF

Though in this case, we observe that all the metrics are high for a particular value of r, it is possible that sometimes some of the scores contradict each other. In such cases, r-values can be picked based on the V-score, since it captures both the homogeneity and completeness. Then, the Adjusted Mutual Information Score can be used as a distinguishing factor and finally the Adjusted Rand Index score can be considered as it signifies the accuracy of clustering

QUESTION 6: Non-monotonic behaviour of measures as r increases.

As discussed in the previous section, the number of clusters and dimension of the dataset have a significant impact on the performance of k-means clustering algorithm. Since in this case, the number of clusters is fixed and only the dimensionality of data changes, as the dimensionality increases, the performance of the algorithm becomes poorer. In high dimensional space, euclidean distance is almost the same for any two vectors. However, we see that at very low dimensions, the algorithm performs poorly since all the information that is needed to distinguish between the classes is not present in the data.

4. VISUALIZATION

In this part, we plot/visualize the the clustering results of k-means with the best value of r found in the previous step, which is r=2 for both SVD and NMF. We do so by first reducing the data to 2D using SVD (n_components=2) and then plot with the clustering results and ground truths for both. Since the data-points are just grouped into clusters, but the clusters aren't matched to labels, there is no one-to-one mapping from assigned clusters and ground truth data labels. This is the reason why the colours in some of the plots may appear switched, or mismatched.

QUESTION 7: Visualize the clustering results for:

• SVD with its best r
• NMF with its best r

<u>Results:</u>

SVD with best r = 2

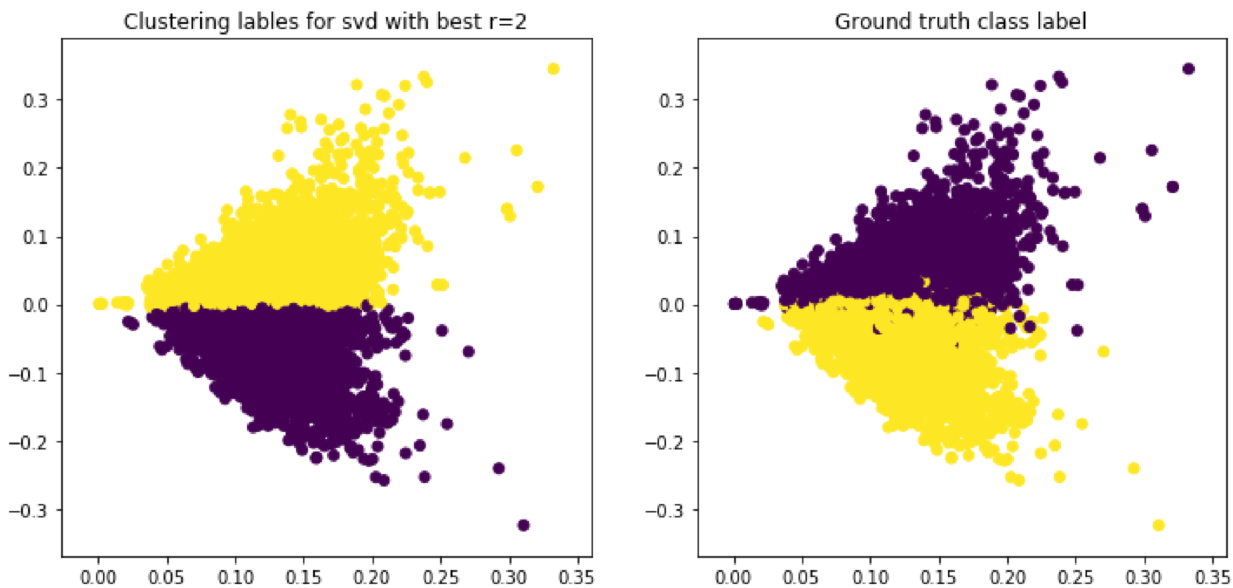


Fig 4 : Plot of clustering results' labels for svd with best r=2 against ground truth labels.

NMF with best r = 2

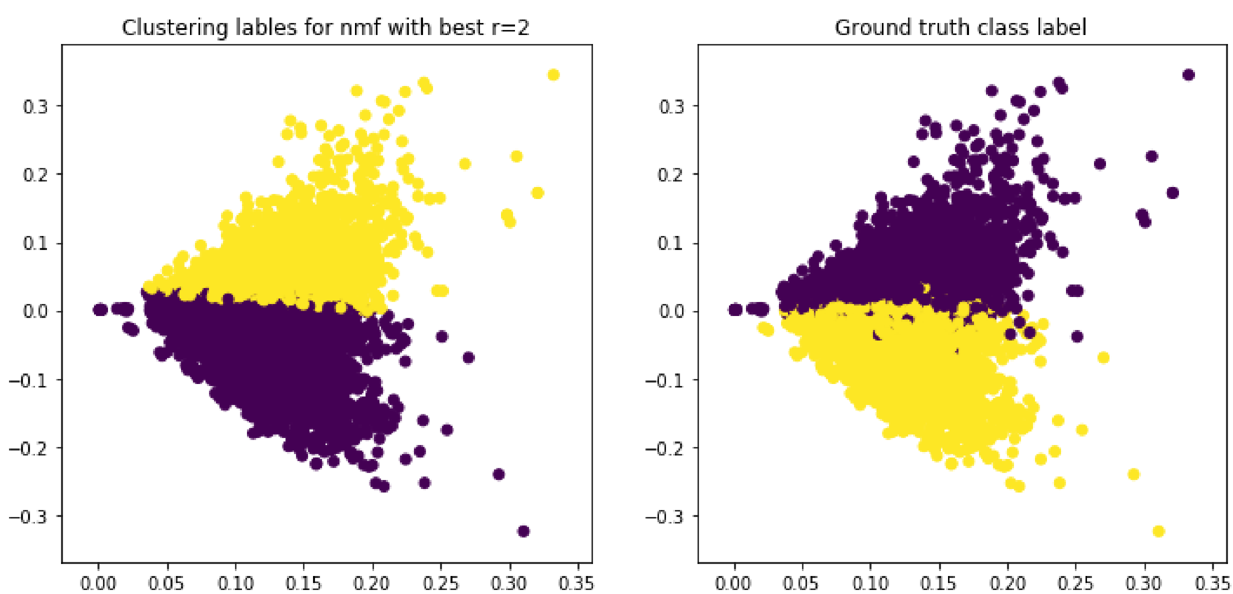


Fig 5 : Plot of clustering results' labels for nmf with best r=2 against ground truth labels.

In this second part of visualization, we try out 4 transformation methods and evaluate the corresponding clustering results. The transformations are performed on SVD-reduced data and NMF-reduced data using the best r. For these visualizations also we by first reduce the data to 2d using SVD (n_components=2).

QUESTION 8: Visualize the transformed data

AND

QUESTION 10: Report the new clustering measures (except for the contingency matrix) for the clustering results of the transformed data.

i. <u>Normalization: Scaling features s.t. each feature has unit variance:</u>

a. For SVD:

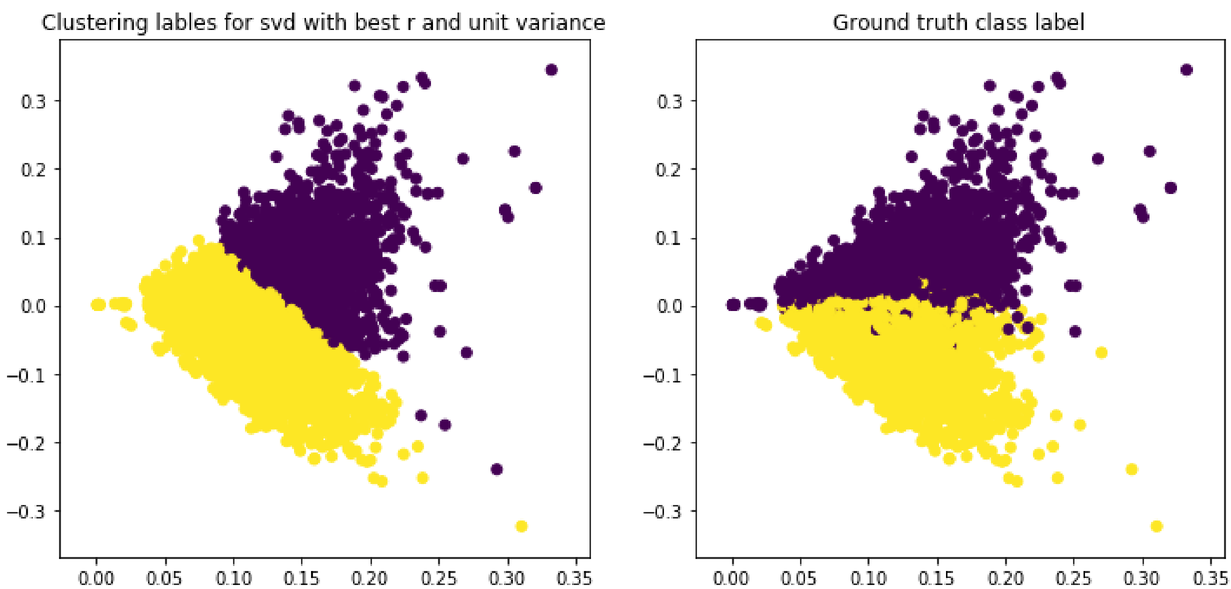


Fig 6 : Plot of clustering results' labels for svd with best r=2 & unit variance against ground truth labels.

Table 4 : Metric values for clustering results' for svd with best r=2 & unit variance

Metric	Value
Homogeneity	0.23609131805072042
Completeness	0.26450291273288246
V-measure	0.24949085488903866
Adjusted Rand Index	0.2556510317794412
Adjusted Mutual Information Score	0.2360213789796711

b. For NMF:

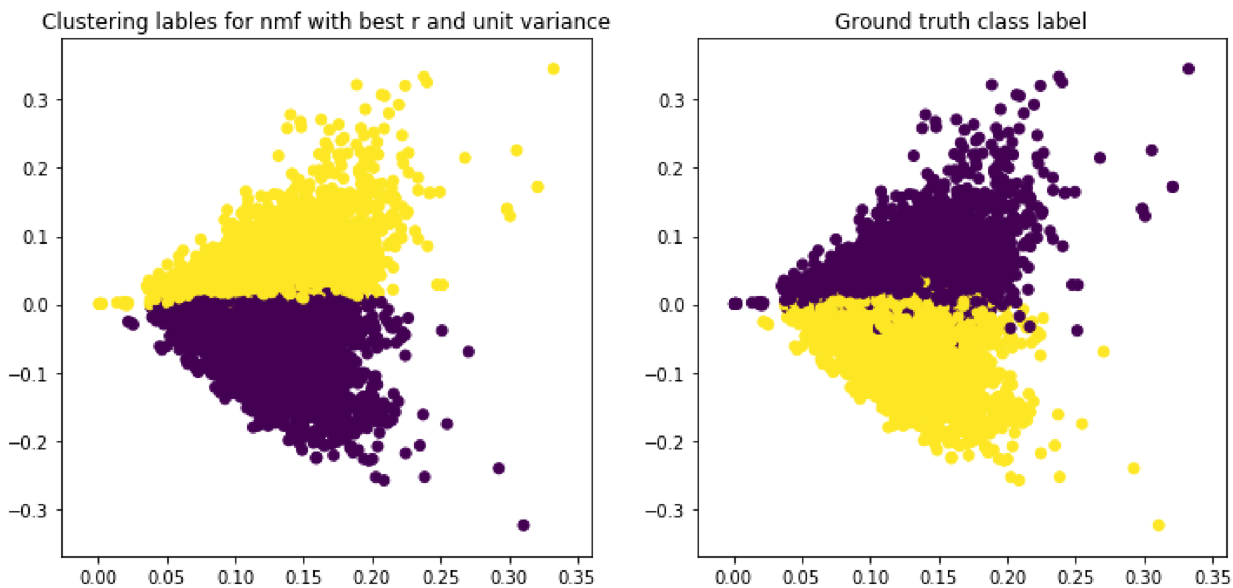


Fig 7 : Plot of clustering results' labels for nmf with best r=2 & unit variance against ground truth labels.

Table 5 : Metric values for clustering results' for nmf with best r=2 & unit variance

Metric	Value
Homogeneity	0.6828038321574016
Completeness	0.6856459752144646
V-measure	0.6842219522524521
Adjusted Rand Index	0.7734426774605906

Adjusted Mutual Information Score	0.6827747927166996

ii. Logarithm transformation:

$$\mathbf{f}(\mathbf{x}) = \mathbf{sign}(\mathbf{x}) \cdot \left(\log(|\mathbf{x}| + c) - \log c\right), \quad \left(\mathbf{sign}(\mathbf{x})\right)_i \equiv \begin{cases} 1 & x_i > 0 \\ 0 & x_i = 0 \\ -1 & x_i < 0 \end{cases}$$

With C=0.01

a. **For SVD:**

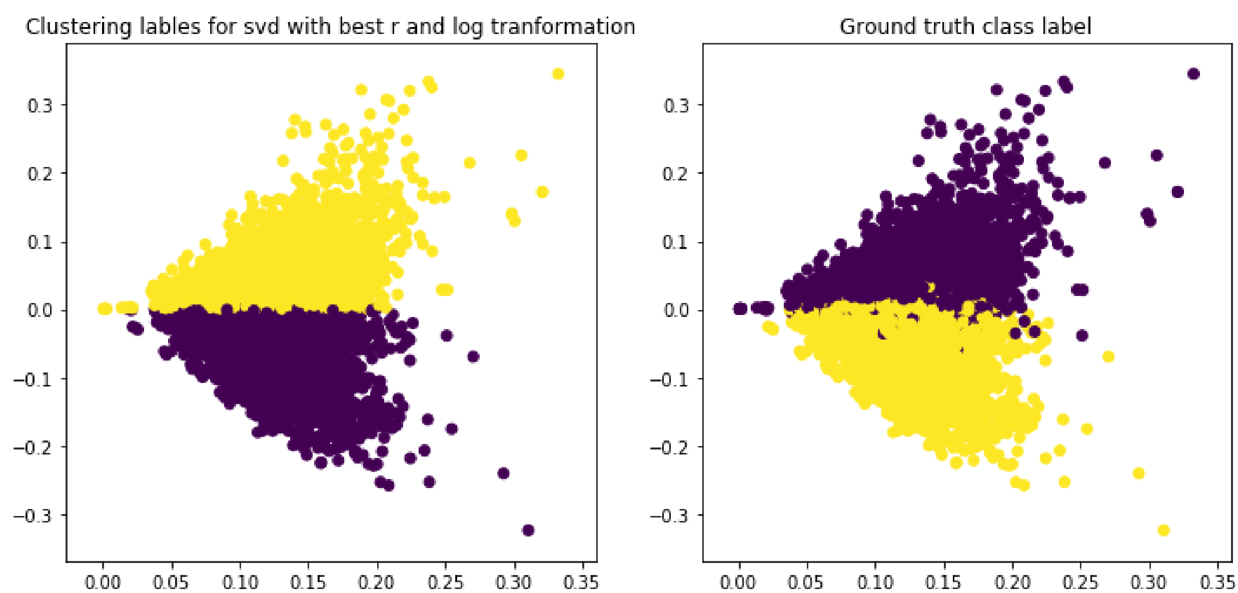


Fig 8 : Plot of clustering results' labels for svd with best r=2 & log transformation against ground truth labels.

Table 6 : Metric values for clustering results' for svd with best r=2 & log transformation

Metric	Value
Homogeneity	0.6103154102550904
Completeness	0.6102847108358388
V-measure	0.6103000601594027
Adjusted Rand Index	0.7173615346457451

Adjusted Mutual Information Score	0.6102490340872908

b. For NMF:

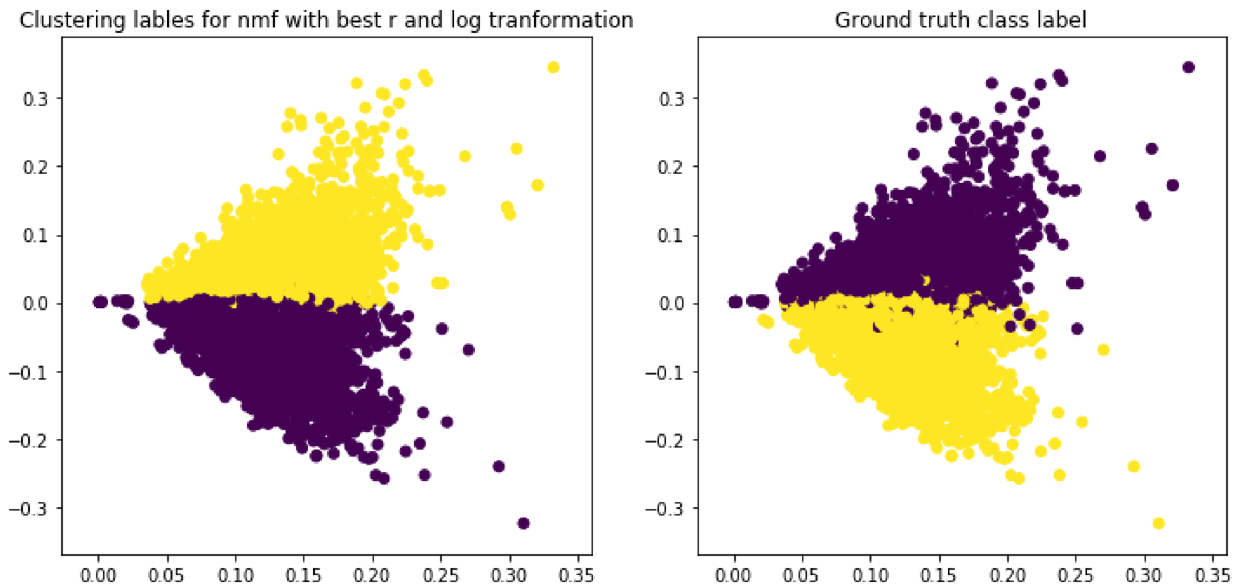


Fig 9 : Plot of clustering results' labels for nmf with best r=2 & log transformation against ground truth labels.

Table 6 : Metric values for clustering results' for nmf with best r=2 & log transformation

Metric	Value
Homogeneity	0.7008978537788074
Completeness	0.7021953876542337
V-measure	0.7015460207584749
Adjusted Rand Index	0.7950170594682267
Adjusted Mutual Information Score	0.7008704708755917

iii. Normalization first then Log transformation: We try the combination of unit variance followed by log transformation in that order for svd and nmf:

a. For SVD:

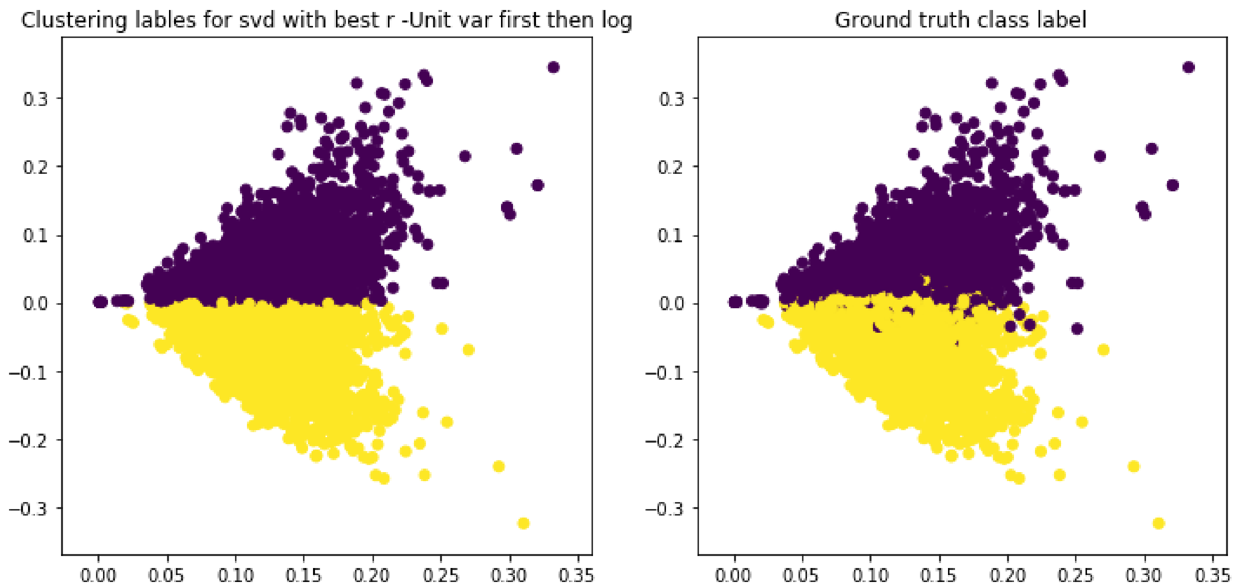


Fig 10 : Plot of clustering results' labels for svd with best r=2, unit variance first then log transformation against ground truth labels.

Table 7 : Metric values for clustering results' for svd with best r=2 & unit variance first then log transformation .

Metric	Value
Homogeneity	0.6094447324650026
Completeness	0.609408642445872
V-measure	0.609426686921128
Adjusted Rand Index	0.7165020096856823
Adjusted Mutual Information Score	0.6093728858160544

b. For NMF:

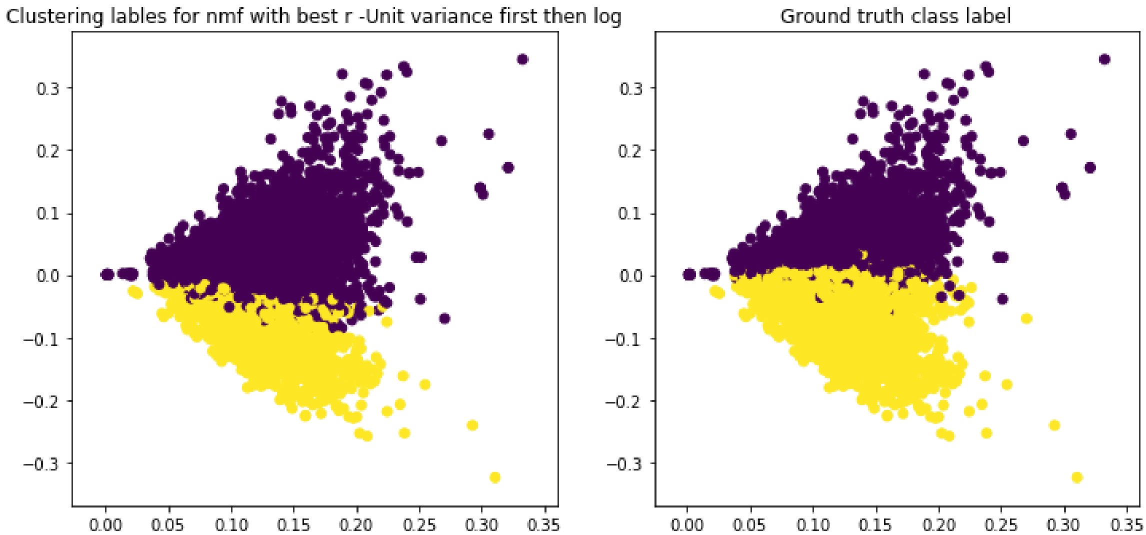


Fig 11 : Plot of clustering results' labels for nmf with best r=2, unit variance first then log transformation against ground truth labels.

Table 8 : Metric values for clustering results' for nmf with best r=2 & unit variance first then log transformation .

Metric	Value
Homogeneity	0.31296490432638224
Completeness	0.38268979660509567
V-measure	0.3443331164752487
Adjusted Rand Index	0.2485171789504215
Adjusted Mutual Information Score	0.31290200097622145

iv. Log transformation first then Normalization: We try the combination of log transformation followed by unit variance in that order for svd and nmf:

a. For SVD:

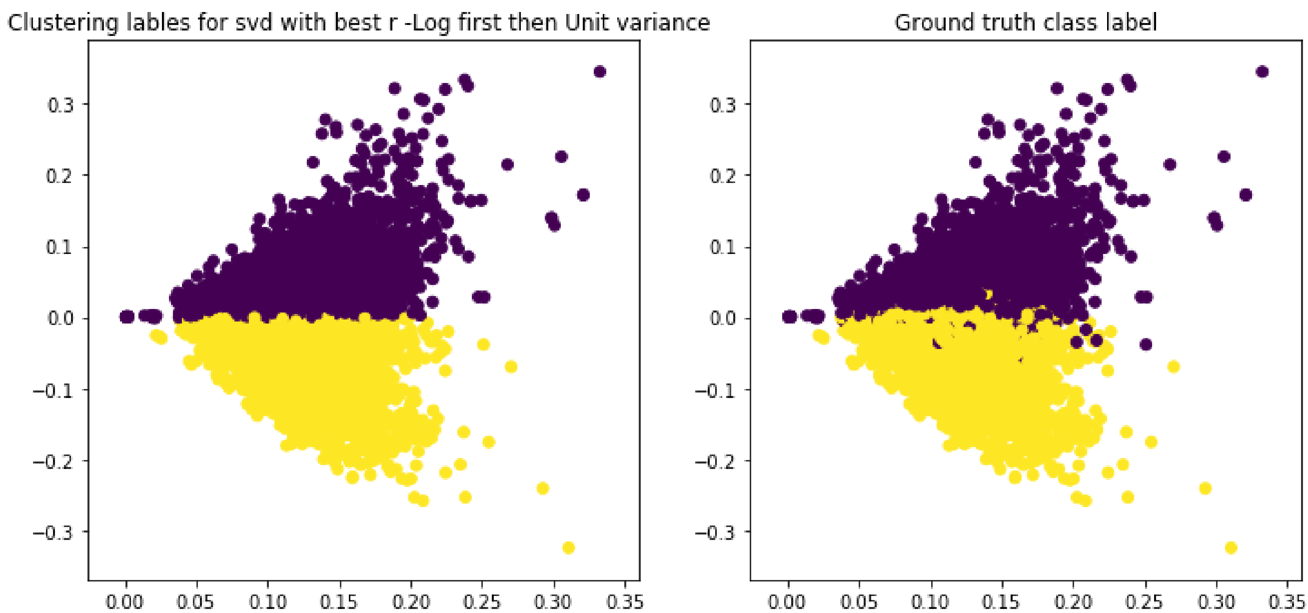


Fig 12 : Plot of clustering results' labels for svd with best r=2, log transformation first then unit variance against ground truth labels.

Table 9: Metric values for clustering results' for svd with best r=2 & log transformation first then unit variance.

Metric	Value
Homogeneity	0.6095947774726546
Completeness	0.6095541488812544
V-measure	0.609574462499973
Adjusted Rand Index	0.7165020095866556
Adjusted Mutual Information Score	0.6095184058375314

b. For NMF:

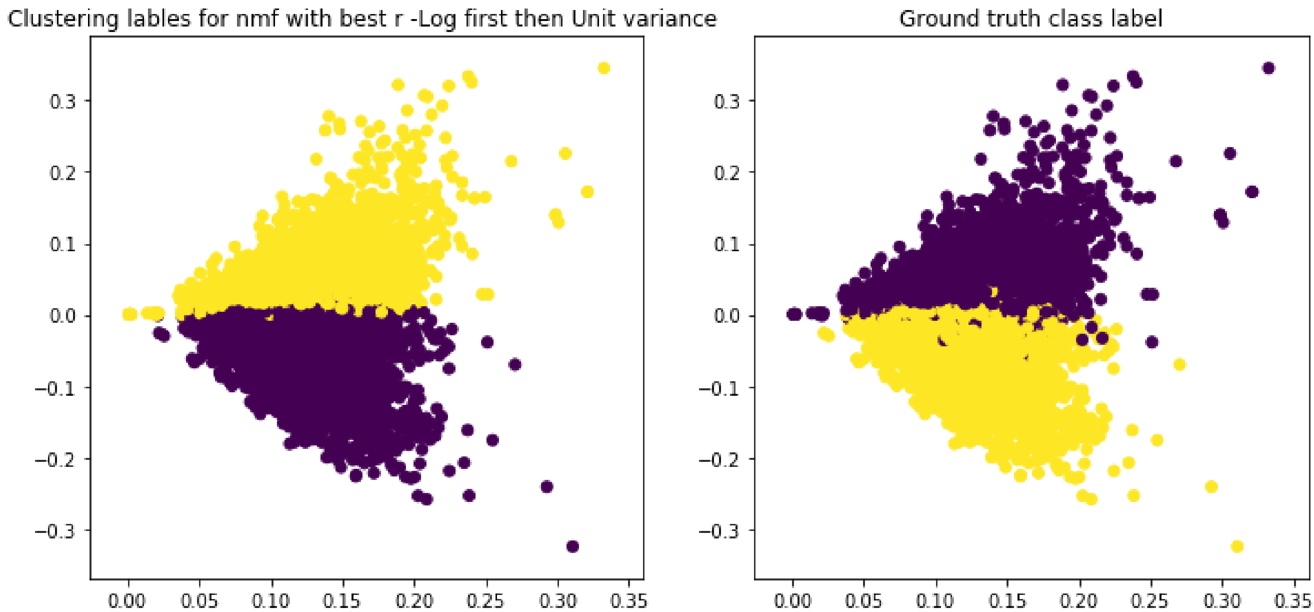


Fig 13 : Plot of clustering results' labels for nmf with best r=2, log transformation first then unit variance against ground truth labels.

Table 10: Metric values for clustering results' for nmf with best r=2 & log transformation first then unit variance.

Metric	Value
Homogeneity	0.7029330839809769
Completeness	0.7040919076936536
V-measure	0.703512018634563
Adjusted Rand Index	0.7972814558331544
Adjusted Mutual Information Score	0.7029058874057212

QUESTION 9: Can you justify why the "logarithm transformation" may improve the clustering results?

The reason log transformation is seen to improve our results is because k-means favours roughly spherical-shaped clusters. Data with heavily skewed variables may lead to very elongated clusters that are not well captured by this method. Taking the log of a variable will reduce the skewness and typically makes the distribution closer to normal. If the log-transformed data is close to normally distributed, then the performance of k-means increases. This is seen in our experiments too where metric values for log transformed data for both SVD and NMF is much better compared to just performing normalization.

QUESTION 11: Repeat the following for 20 categories using the same parameters as in 2-class case:

- ***Transform corpus to TF-IDF matrix;***
- ***Directly perform K-means and report the 5 measures and the contingency matrix;***

Table 11 :

Metric	Value
Homogeneity	0.35942082651801804
Completeness	0.45111242050273204
V-measure	0.4000803165708632
Adjusted Rand Index	0.13663613501490818
Adjusted Mutual Information Score	0.35731878968094594

Contingency Matrix:

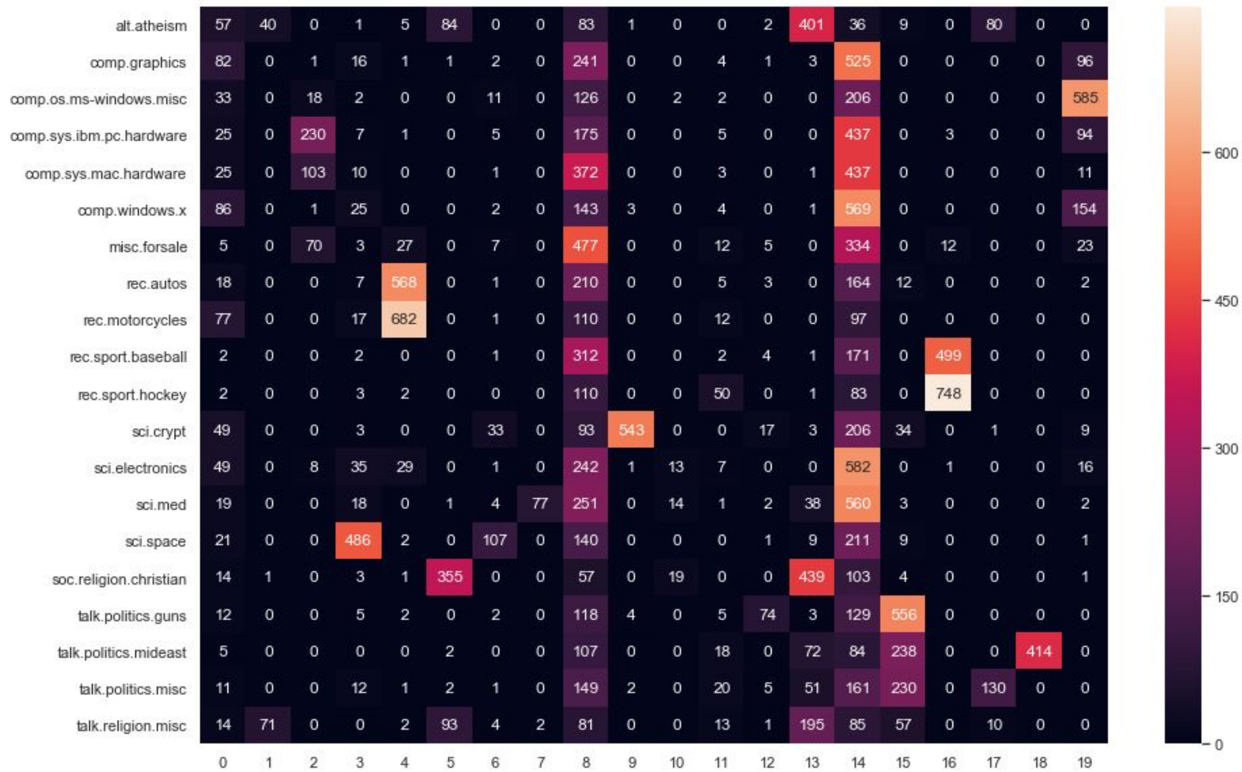


Fig 14 : Contingency Matrix for Question 11

QUESTION 12: Try different dimensions for both truncated SVD and NMF dimensionality reduction techniques and the different transformations of the obtained feature vectors as outlined in above parts.

We perform dimensionality reduction using SVD and NMF for different dimensions and obtain the best value of ‘r’ for both.
We try the dimensions : [1, 2, 3, 5, 10, 20, 50, 100, 300]

<u>SVD:</u>
The following table shows the value of ‘r’ which gave the best results for each metric:

Table 12: Best values of ‘r’ according to 5 metrics

Metric	**Best Value of R**
Homogeneity Score	10
Completeness Score	300
V Measure Score	300
Adjusted Rand Score	10
Adjusted Mutual Info Score	10

The five scores above show that the best values of ‘r’ are 10 or 300. We apply different transformations with ‘r’ as 10 and 300 and compare the metrics.

Transformations:

1. Normalization: Scaling features such that each feature has unit variance with ‘r’ as 10:

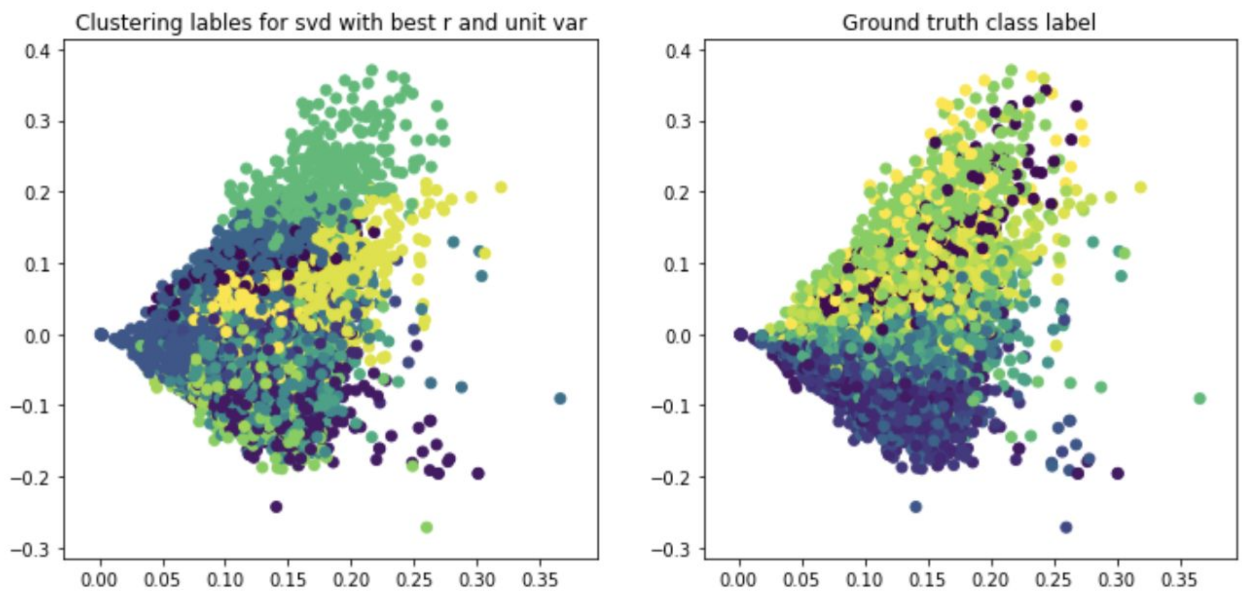


Fig 15 : Clustering Labels vs Ground Truth Labels for r=10 and unit variance

Table 13: Scores for r=10 and unit variance

Metric	Value
Homogeneity	0.30922378407411316
completeness	0.3478736400484909

V-measure	0.32741203786962125
adjusted Rand Index	0.12359410008124433
adjusted mutual information score	0.30697974680838025

2. Logarithm transformation: Applying logarithm transformation with ‘r’ as 10.

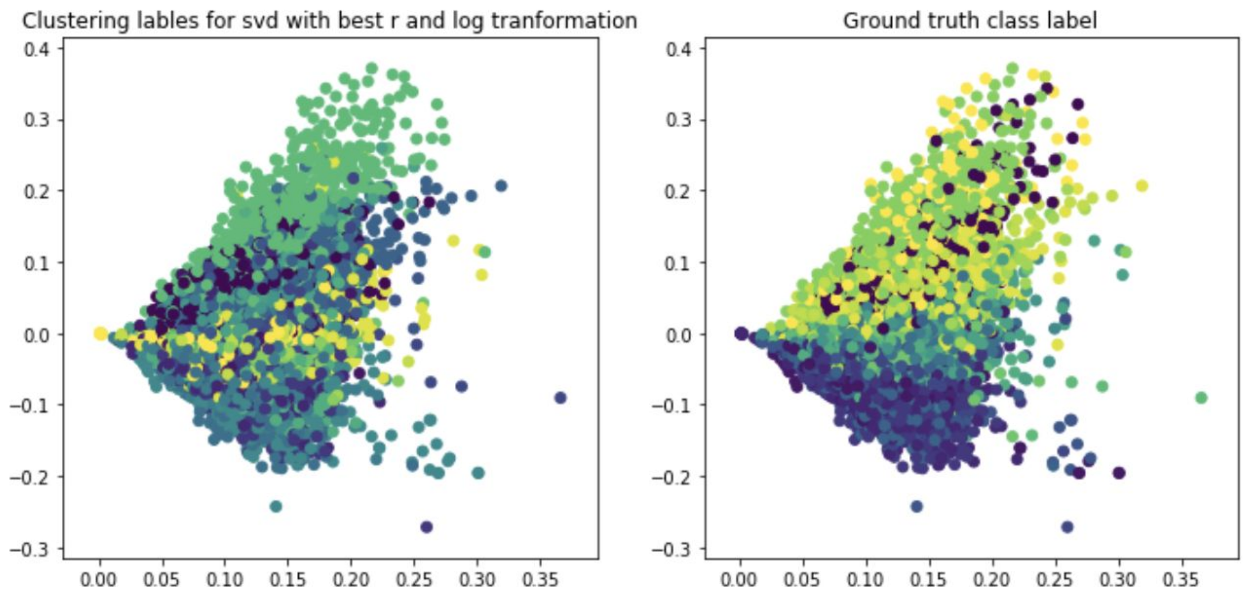


Fig 16 :Clustering Labels vs Ground Truth Labels for r=10 and log transformation

Table 14: Scores for r=10 and log transformation

Metric	Value
Homogeneity	0.3199457278233682
completeness	0.3243669936635296
V-measure	0.3221411914079869
adjusted Rand Index	0.1638402353136279
adjusted mutual information score	0.31775054764242583

3. Normalization first then Log transformation: We try the combination of unit variance followed by log transformation in that order with ‘r’ as 10:

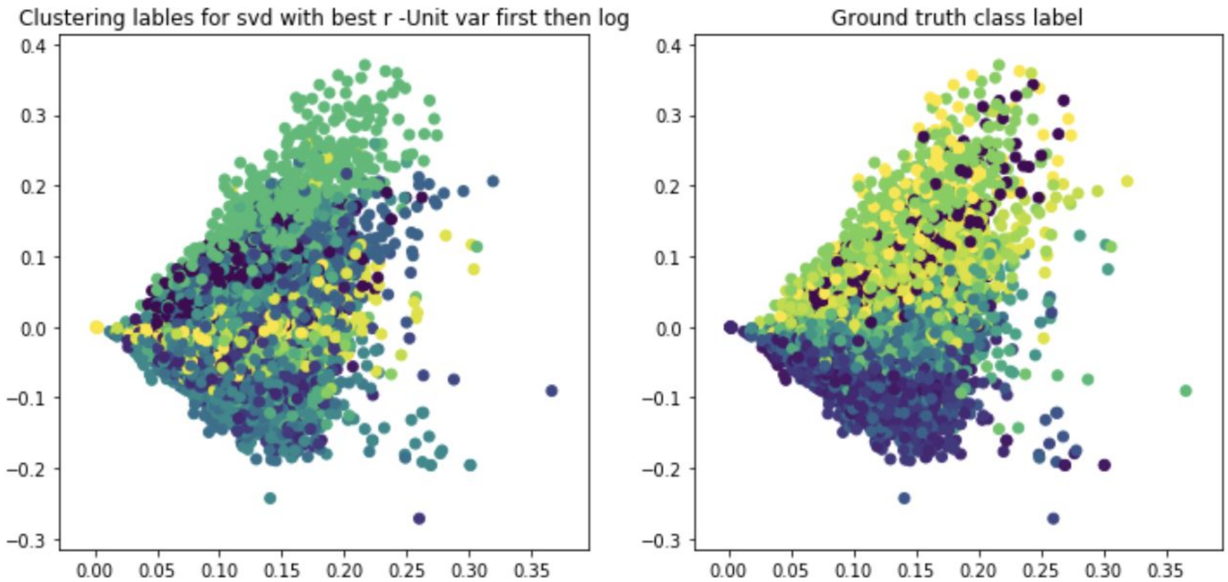


Fig 17 :Clustering Labels vs Ground Truth Labels for r=10 and unit-variance and log transformation in that order

Table 15:Scores for r=10 and unit-variance and log transformation in that order

Metric	Value
Homogeneity	0.3199457278233682
completeness	0.3243669936635296
V-measure	0.3221411914079869
adjusted Rand Index	0.1638402353136279
adjusted mutual information score	0.31775054764242583

4. Log transformation first then Normalization : We try the combination of log transformation followed by unit variance in that order with ‘r’ as 10:

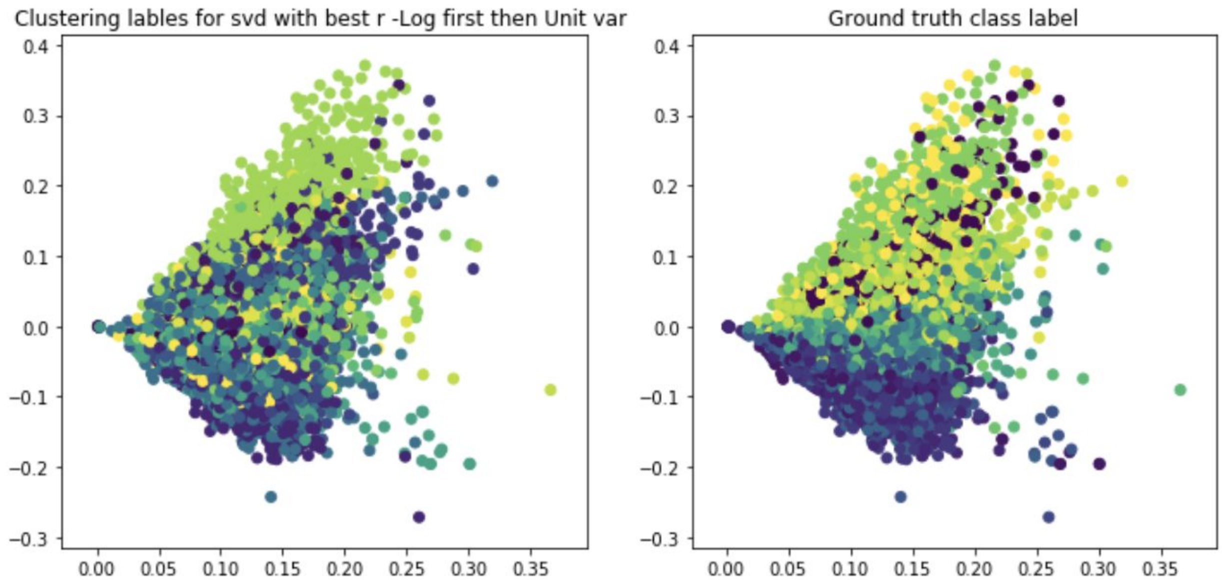


Fig 18 : Clustering Labels vs Ground Truth Labels for r=10 and log transformation and unit-variance in that order

Table 16: Scores for r=10 and log transformation and unit-variance in that order

Metric	Value
Homogeneity	0.3365704182774187
completeness	0.3385488059268524
V-measure	0.33755671334177506
adjusted Rand Index	0.1802102387600942
adjusted mutual information score	0.33442961544381966

5. Logarithm transformation: Applying logarithm transformation with ‘r’ as 300.

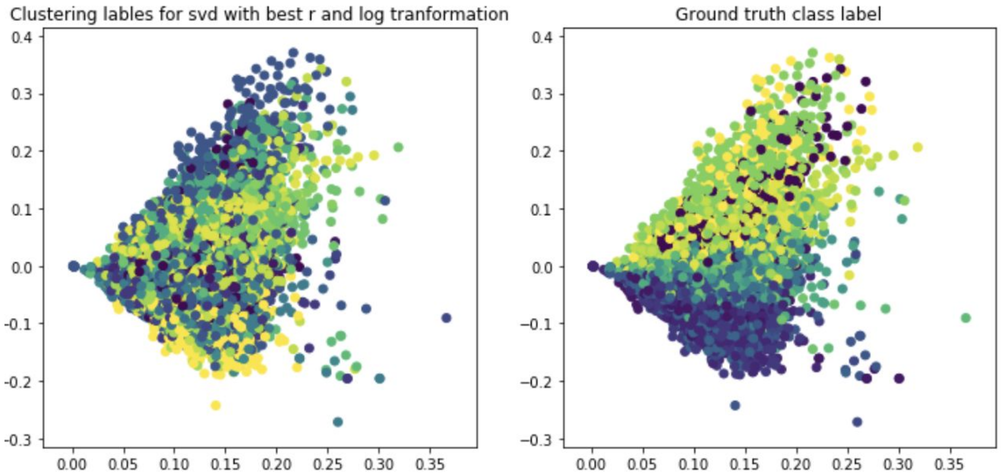


Fig 19 : Clustering Labels vs Ground Truth Labels for r=300 and log transformation

Table 17: Scores for r=300 and log transformation

Metric	Value
Homogeneity	0.36587642013962635
completeness	0.38373629525984254
V-measure	0.37459359774198764
adjusted Rand Index	0.19658435288926981
adjusted mutual information score	0.3638283057898593

6. After having tried out transformations for r=10 and r=300, we tried transformations on r=100 and received the best results for logarithm transformation.

Logarithm transformation: Applying logarithm transformation with ‘r’ as 100.

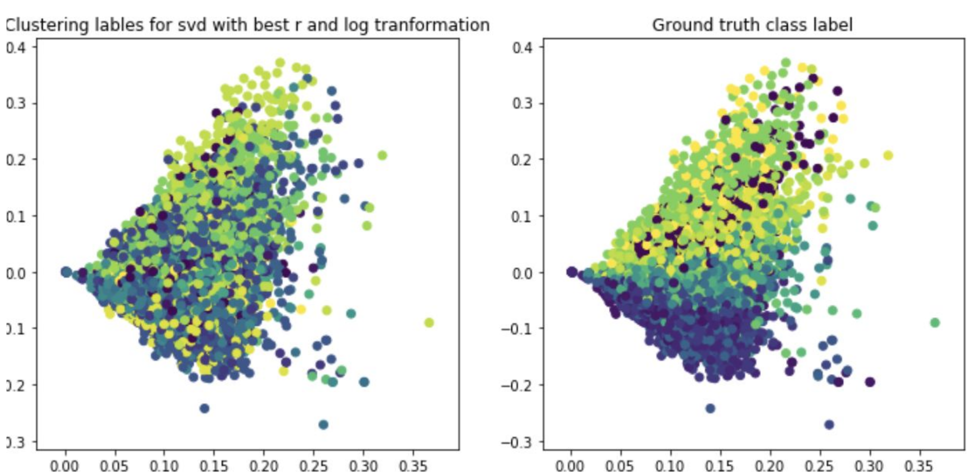


Fig 20: Clustering Labels vs Ground Truth Labels for r=100 and log transformation

Table 18: Scores for r=100 and log transformation

Metric	Value
Homogeneity	0.3801432134472916
completeness	0.3947415434878439
V-measure	0.38730486689693383
adjusted Rand Index	0.22525406954213492
adjusted mutual information score	0.37814122660925575

For SVD, the following combination gave the best results:
Logarithm transformation on feature vectors with ‘r’ as 100.

NMF:
The following dimensions for r were tried out for NMF:

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

The metrics used to compare the performance of the classifiers were:

1. Homogeneity Score
2. Completeness Score
3. V Measure Score
4. Adjusted Rand Score
5. Adjusted Mutual Info Score

The following table shows the value of r which gave the best results for each metric:

Table 19: Best value of R for each clustering performance metric

Metric	Best Value of R
Homogeneity Score	10
Completeness Score	10
V Measure Score	10

Adjusted Rand Score	10
Adjusted Mutual Info Score	10

The classifier created using r = 10 gave the best results on <u>all</u> the 5 metrics.

Values of r close to 10, i.e. [5, 10, 20] were tried out in the next step, along with the 4 combinations of transformations.

The 4 transformations tried out with the values of r = [5, 10, 20] were

a. Unit Variance
b. Logarithmic Transformation
c. Unit Variance followed by Logarithmic Transformation
d. Logarithmic Transformation followed by Unit Variance

Table 20: Best Combination of R and Transformation for each Clustering Metric

Metric	**Best R**	**Best Transformation**	**Value**
Homogeneity Score	10	Unit	0.301469781212141
Completeness Score	20	Unit	0.3657162906588827
V Measure Score	20	Unit	0.3201180377046113
Adjusted Rand Score	10	Unit	0.1115350853744247
Adjusted Mutual Info Score	10	Unit	0.2992055446662649

Since the value of r = 10 is the best parameter over a majority of the performance metrics, this was chosen for the final model. The transformation of only Unit Variance was the best over all metrics and combinations with r, and was chosen.

Finally a model with **NMF with r = 10, Unit Variance transformation** was trained with the following parameters:

max_iter=5000

n_init=100

The results of this are:

Table 21: Results of Best NMF, with R = 10, and Unit Variance Transformation

Metric	Results
Homogeneity Score	0.3011653592438494
Completeness Score	0.3375465519954479
V Measure Score	0.3183198146281171
Adjusted Rand Score	0.11110893551643714
Adjusted Mutual Info Score	0.29890009843193877

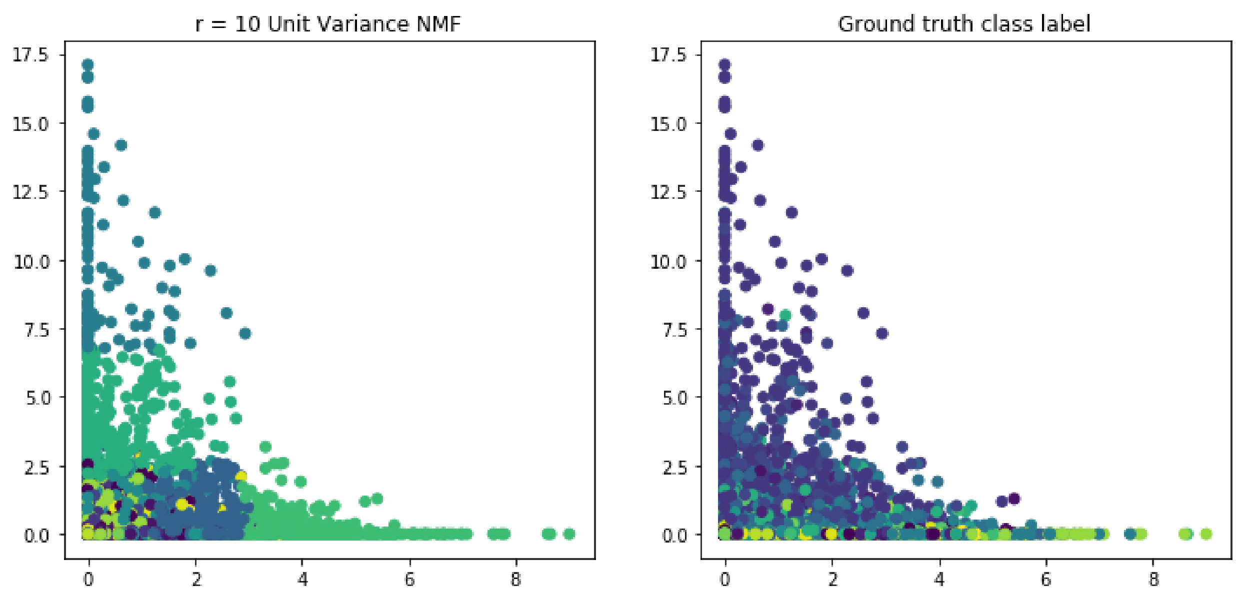


Fig 21 : Plot of Clustering results vs Ground Truth for best NMF (with r=10 and Unit Variance)

Best Results: SVD v.s NMF

In the table below, the best results using SVD and NMF are presented, with the best choice for each score highlighted in green:

Table 22: Comparison of best clustering results of SVD, NMF and TF-IDF for 20 clusters

Metric	Best SVD	Best NMF	TF-IDF

Homogeneity Score	0.3801432134472916	0.3011653592438494	0.35942082651801804
Completeness Score	0.3947415434878439	0.3375465519954479	0.45111242050273204
V Measure Score	0.38730486689693383	0.3183198146281171	0.4000803165708632
Adjusted Rand Score	0.22525406954213492	0.11110893551643714	0.13663613501490818
Adjusted Mutual Info Score	0.37814122660925575	0.29890009843193877	0.35731878968094594

Final Model:
Majority of the scores have highest values associated with SVD with log transformation and r as 100.
The combination that gave the best results is:
Applying SVD on feature vectors with n_components as 100 and further applying logarithm transformations on them.
This particular combination is found to be better than other combinations. On an average, the scores for this combination are higher than other scores by around 0.5.

Table 23: Scores for the best combination: r=100 and log transformation

Metric	Value
Homogeneity	0.3801432134472916
completeness	0.3947415434878439
V-measure	0.38730486689693383
adjusted Rand Index	0.22525406954213492
adjusted mutual information score	0.37814122660925575

Combinations that seemed undesirable:
We observed relatively low scores for r=5, r=50.
We observed relatively low scores for unit variance..
Combinations that seemed desirable:
We observed relatively high scores for r=100,r=10,r=300.
We observed relatively high scores when we applied logarithm transformation.