ECE 219 Project 2

Tanner Dulay, Loïc Maxwell, Zoë Tucker

February 3, 2021

1 Question 1

In this assignment, we explore various methods for clustering of text data. To begin, we use the"20 Newsgroups" dataset, taking the two well-separated classes previously defined ("computer technology" and "recreational activity"). As this project uses unsupervised learning, we do not distinguish between training and test data, and instead use the entire dataset throughout the project.

To start our process, we build the TF-IDF matrix. We use `min_df` = 3, exclude English stopwords/punctuation, and remove headers and footers, but we do not perform stemming or lemmatization. The resulting TF-IDF matrix is $\mathbf{7882} \times \mathbf{21909}$.

2 Question 2

Next, we apply k-means clustering directly to the TF-IDF data. As we are interested in two classes of documents, we set $k = 2$. The below contingency matrix (Figure 1) shows the results of this.

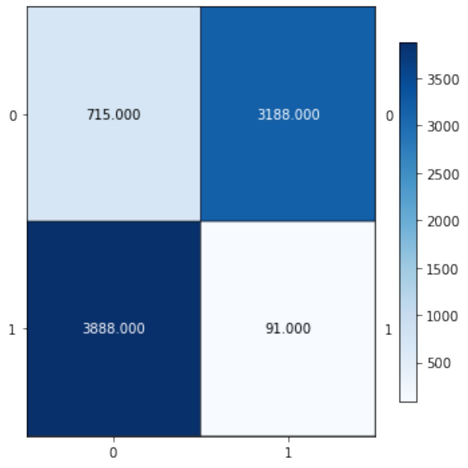


Figure 1: Contingency Matrix for K-Means on TF-IDF Matrix

3 Question 3

We then find the metrics we need to evaluate our classification results. We find the following values:

Homogeneity score	0.5600025412974119
Completeness score	0.5716557539454379
V_measure score	0.5657691483417876
Adjusted Rand Index score	0.6327475635535726
Adjusted Mutual Information Score	0.5657289846874574

4 Question 4

As the metrics found in Question 3 do not demonstrate a good clustering result, we now try dimensionality reduction to help our data better fit the assumptions made by k-means algorithms.

We begin by finding the percent of variance the top r principal components can retain for various values of r. Figure 2 shows our findings.

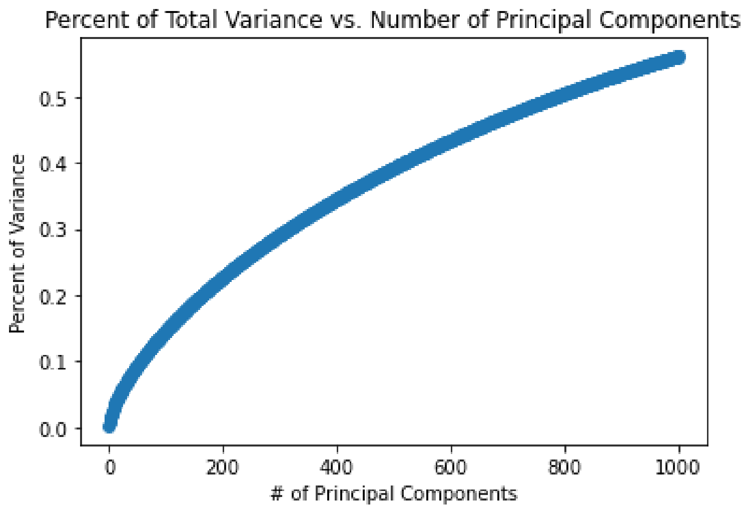


Figure 2: Principal Component Results

5 Question 5

Next, we will try to experimentally find the value of r (i.e. the number of dimensions) that we can reduce our data to in order to obtain the best clustering results. Testing $r \in \{1, 2, 3, 5, 10, 20, 50, 100, 300\}$ using both NMF and SVD, we plot the values of the five metrics found in Question 3 for each r. The results are shown in Figures 3 and 4.

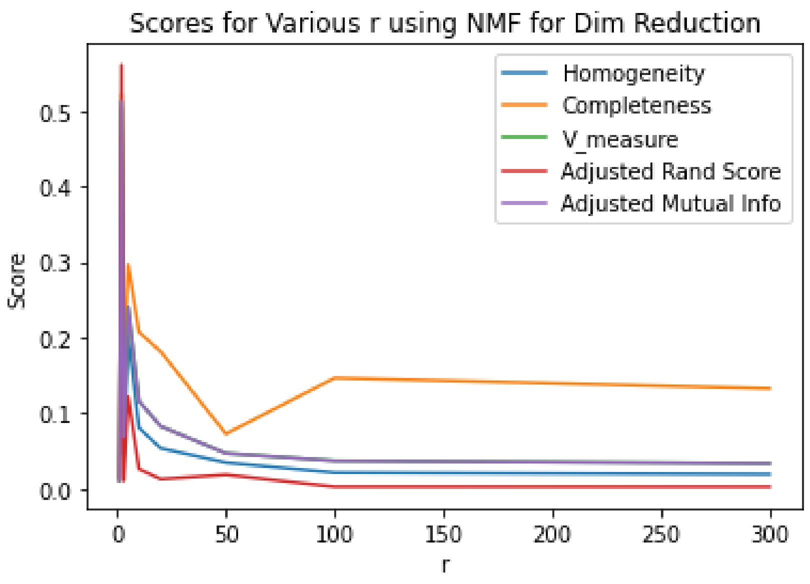


Figure 3: NMF Results

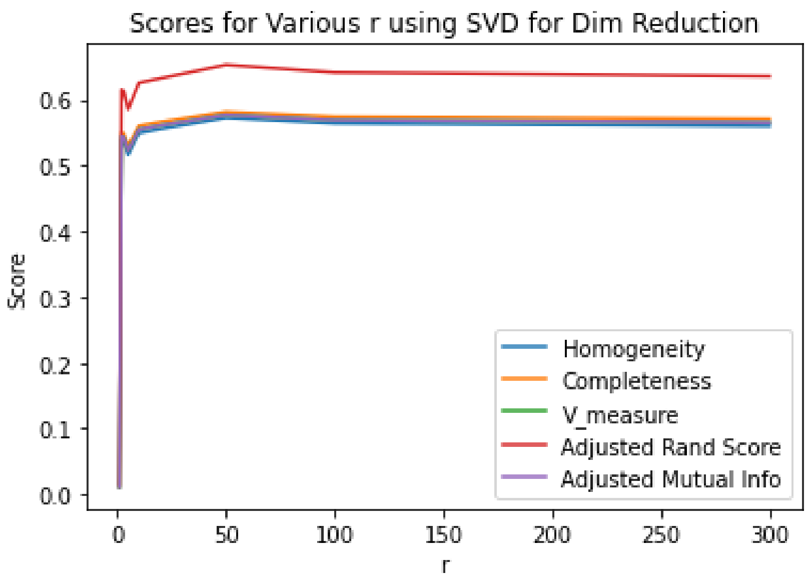


Figure 4: SVD Results

6 Question 6

As we can see, the behavior of the measures as r increases is non-monotonic.

This is due to a trade-off between a larger value of r giving a more faithful representation of the data, and a smaller value of r being better suited for k-means clustering. The higher the dimensionality of a dataset, the more difficult it is to meaningfully evaluate the "distance" between data points, which is what k-means relies on. Thus, we expect there to be an optimal value of r that balances these two concerns and maximizes our scores. In our case, we find that $r = 2$ for NMF, and $r = 50$ for SVD.

7 Question 7

Now, we want to visualize our clustering results. We do this by using k-means clustering with our dataset reduced by NMF and SVD with the optimal values of r found in Question 6, then projecting the data into 2-dimensional space with SVD and coloring the points according to their label. Figures 5 and 6 show the results for NMF with $r = 2$. Figures 7 and 8 show the results for SVD with $r = 50$.

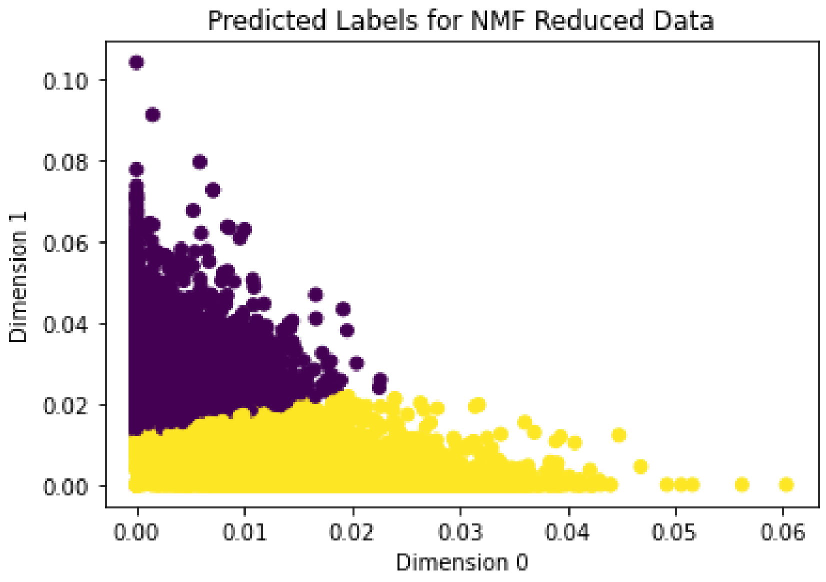


Figure 5: Predicted labels for NMF reduced data

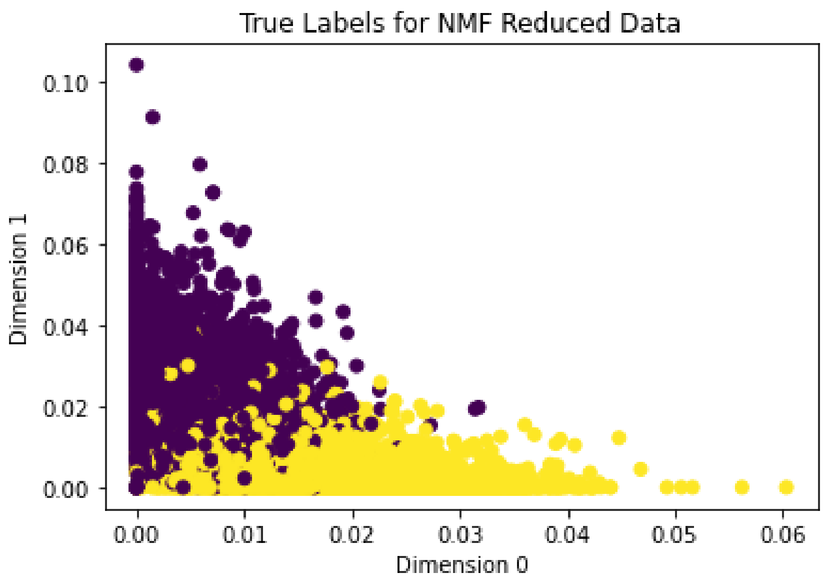


Figure 6: Ground truth labels for NMF reduced data

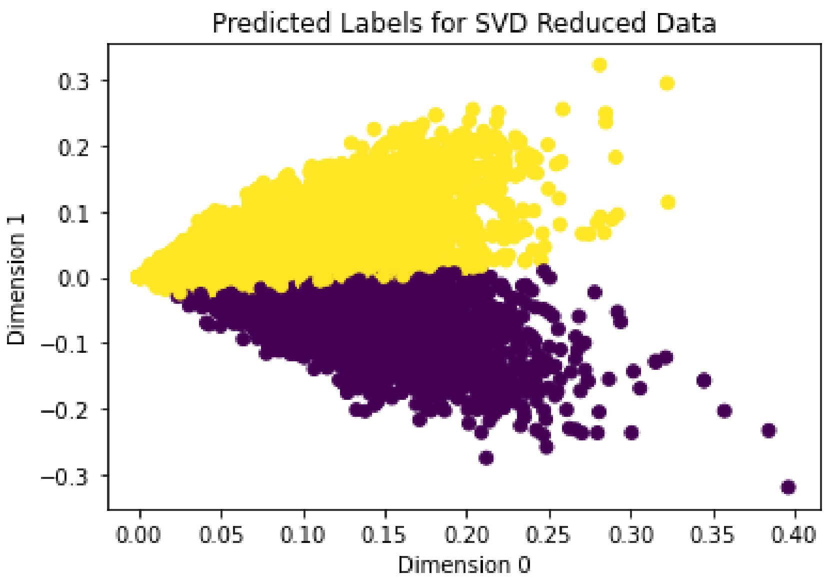


Figure 7: Predicted labels for SVD reduced data

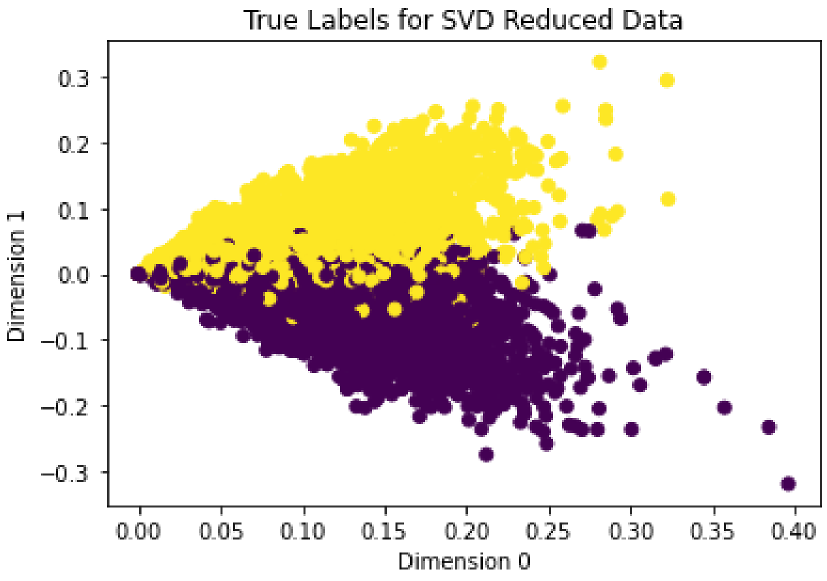


Figure 8: Ground truth labels for SVD reduced data

8 Question 8

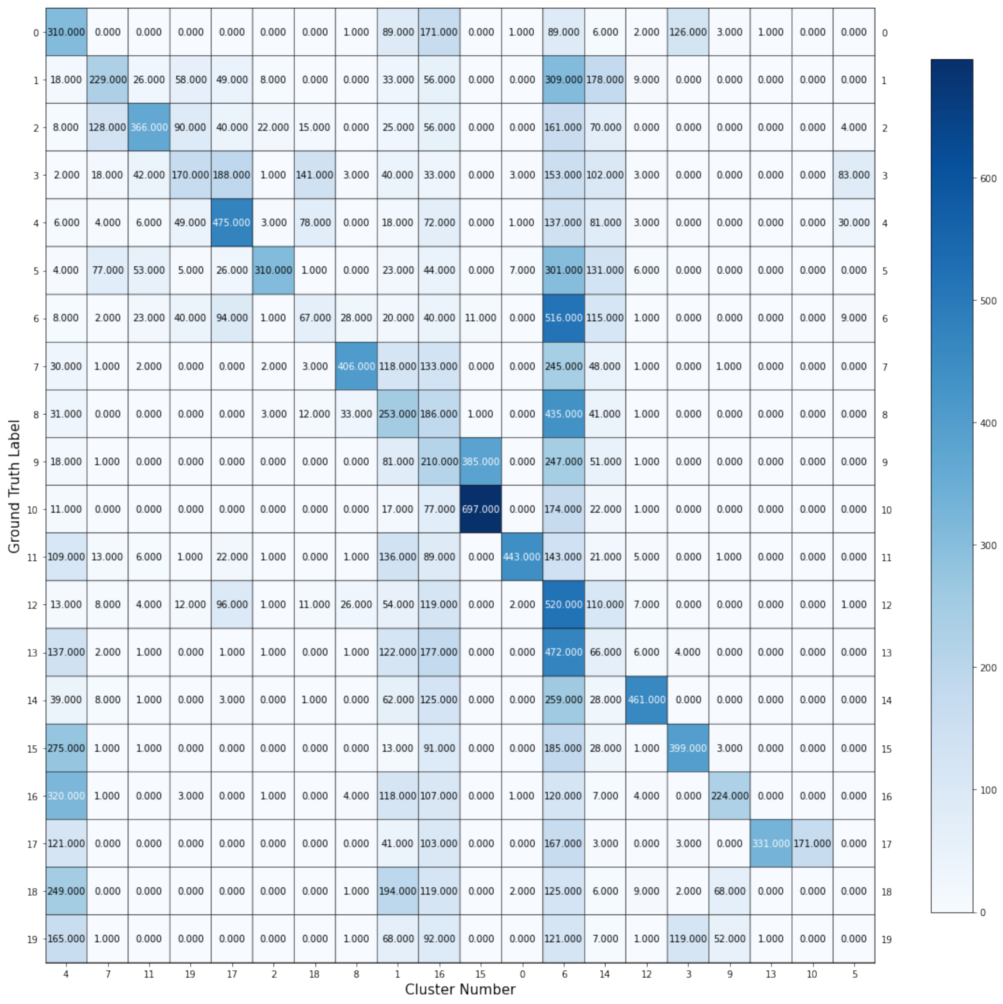
Because the NMF reduced data was already in 2D, we did not reduce it further with SVD. However, if we were to do SVD reduction, the resulting matrix would still be positive, as it is in Figures 5 and 6. SVD is an orthogonal projection that preserves the angles between feature vectors.

The 2D visualization from Figure 8 shows that the SVD reduced data looks somewhat linearly separable around a value of 0 for Dimension 1, which stems from the second principal component of the TF-IDF matrix. This suggests that this principal component holds important information pertaining to the class that the data belongs to. Although this analysis is restricted to two dimensions, this also suggests that this type of distribution would be good for KMeans clustering. A similar type of separation can be found at the diagonal between dimensions 0 and 1 for the NMF reduced data in Figure 6. This would also make it ideal for KMeans clusering.

However, for both types of reductions, because there is no way to perfectly linearly separate the data there are some better options out there for properly classifying this data. Perhaps a classifier that can learn non-linear behavior in the data, such as a neural network with ReLU/eLU, can perform better on this dataset.

9 Question 9

Using SVD with r=50, which gave us the best results on the two classes, we got the following contingency matrix for KMeans on the entire dataset of 20 classes:



Ground Truth Label \ Cluster Number	4	7	11	19	17	2	18	8	1	16	15	0	6	14	12	3	9	13	10	5	
0	310.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	89.000	171.000	0.000	1.000	89.000	6.000	2.000	126.000	3.000	1.000	0.000	0.000	0
1	18.000	229.000	26.000	58.000	49.000	8.000	0.000	0.000	33.000	56.000	0.000	0.000	309.000	178.000	9.000	0.000	0.000	0.000	0.000	0.000	1
2	8.000	128.000	366.000	90.000	40.000	22.000	15.000	0.000	25.000	56.000	0.000	0.000	161.000	70.000	0.000	0.000	0.000	0.000	0.000	4.000	2
3	2.000	18.000	42.000	170.000	188.000	1.000	141.000	3.000	40.000	33.000	0.000	3.000	153.000	102.000	3.000	0.000	0.000	0.000	0.000	83.000	3
4	6.000	4.000	6.000	49.000	475.000	3.000	78.000	0.000	18.000	72.000	0.000	1.000	137.000	81.000	3.000	0.000	0.000	0.000	0.000	30.000	4
5	4.000	77.000	53.000	5.000	26.000	310.000	1.000	0.000	23.000	44.000	0.000	7.000	301.000	131.000	6.000	0.000	0.000	0.000	0.000	0.000	5
6	8.000	2.000	23.000	40.000	94.000	1.000	67.000	28.000	20.000	40.000	11.000	0.000	516.000	115.000	1.000	0.000	0.000	0.000	0.000	9.000	6
7	30.000	1.000	2.000	0.000	0.000	2.000	3.000	406.000	118.000	133.000	0.000	0.000	245.000	48.000	1.000	0.000	1.000	0.000	0.000	0.000	7
8	31.000	0.000	0.000	0.000	0.000	3.000	12.000	33.000	253.000	186.000	1.000	0.000	435.000	41.000	1.000	0.000	0.000	0.000	0.000	0.000	8
9	18.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	81.000	210.000	385.000	0.000	247.000	51.000	1.000	0.000	0.000	0.000	0.000	0.000	9
10	11.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	17.000	77.000	697.000	0.000	174.000	22.000	1.000	0.000	0.000	0.000	0.000	0.000	10
11	109.000	13.000	6.000	1.000	22.000	1.000	0.000	1.000	136.000	89.000	0.000	443.000	143.000	21.000	5.000	0.000	1.000	0.000	0.000	0.000	11
12	13.000	8.000	4.000	12.000	96.000	1.000	11.000	26.000	54.000	119.000	0.000	2.000	520.000	110.000	7.000	0.000	0.000	0.000	0.000	1.000	12
13	137.000	2.000	1.000	0.000	1.000	1.000	0.000	1.000	122.000	177.000	0.000	0.000	472.000	66.000	6.000	4.000	0.000	0.000	0.000	0.000	13
14	39.000	8.000	1.000	0.000	3.000	0.000	1.000	0.000	62.000	125.000	0.000	0.000	259.000	28.000	461.000	0.000	0.000	0.000	0.000	0.000	14
15	275.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	13.000	91.000	0.000	0.000	185.000	28.000	1.000	399.000	3.000	0.000	0.000	0.000	15
16	320.000	1.000	0.000	3.000	0.000	1.000	0.000	4.000	118.000	107.000	0.000	1.000	120.000	7.000	4.000	0.000	224.000	0.000	0.000	0.000	16
17	121.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	41.000	103.000	0.000	0.000	167.000	3.000	0.000	3.000	0.000	331.000	171.000	0.000	17
18	249.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	194.000	119.000	0.000	2.000	125.000	6.000	9.000	2.000	68.000	0.000	0.000	0.000	18
19	165.000	1.000	0.000	0.000	0.000	0.000	0.000	1.000	68.000	92.000	0.000	0.000	121.000	7.000	1.000	119.000	52.000	1.000	0.000	0.000	19

Figure 9: Contingency Matrix for KMeans on 20 Class Data

The respective scores for this clustering are as follows:

Homogeneity score	0.31630458145750995
Completeness score	0.3682552023183464
V_measure score	0.3403086491478565
Adjusted Rand Index score	0.10379799750286293
Adjusted Mutual Information Score	0.3380068278889245

10 Question 10

Using the Kullback-Leibler Divergence loss metric for NMF and r=10, we were able to get the following scores:

Homogeneity score	0.37714629232162333
Completeness score	0.3900505160910544
V_measure score	0.3834898799077068
Adjusted Rand Index score	0.18833101982676798
Adjusted Mutual Information Score	0.38146535004967785

There is indeed a noticeable improvement over the performance using ordinary SVD with r=50 (from Q9).

11 Question 11

We compared the performance of KMeans on the dataset reduced with UMAP using various values for n_components. The plots are shown below:

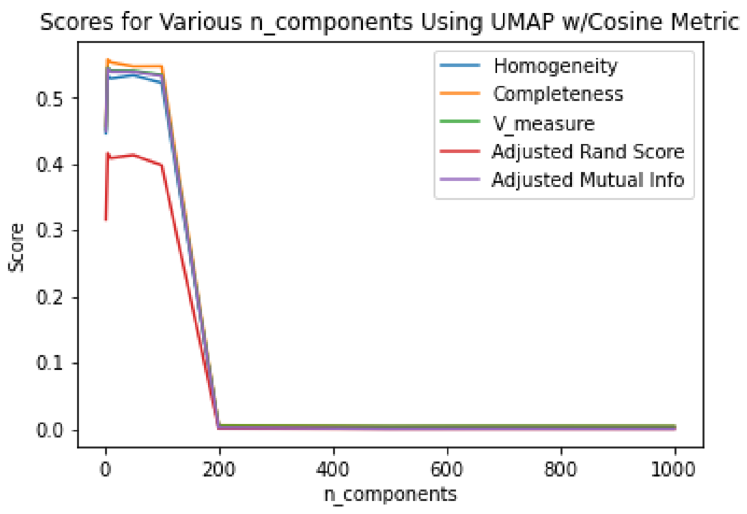


Figure 10: Metric scores for Cosine Metric vs. n_components

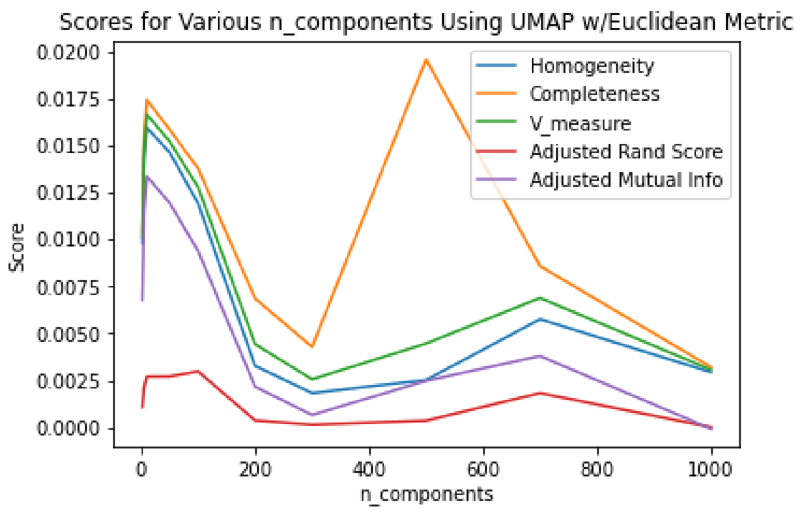


Figure 11: Metric scores for Euclidean Metric vs. n_components

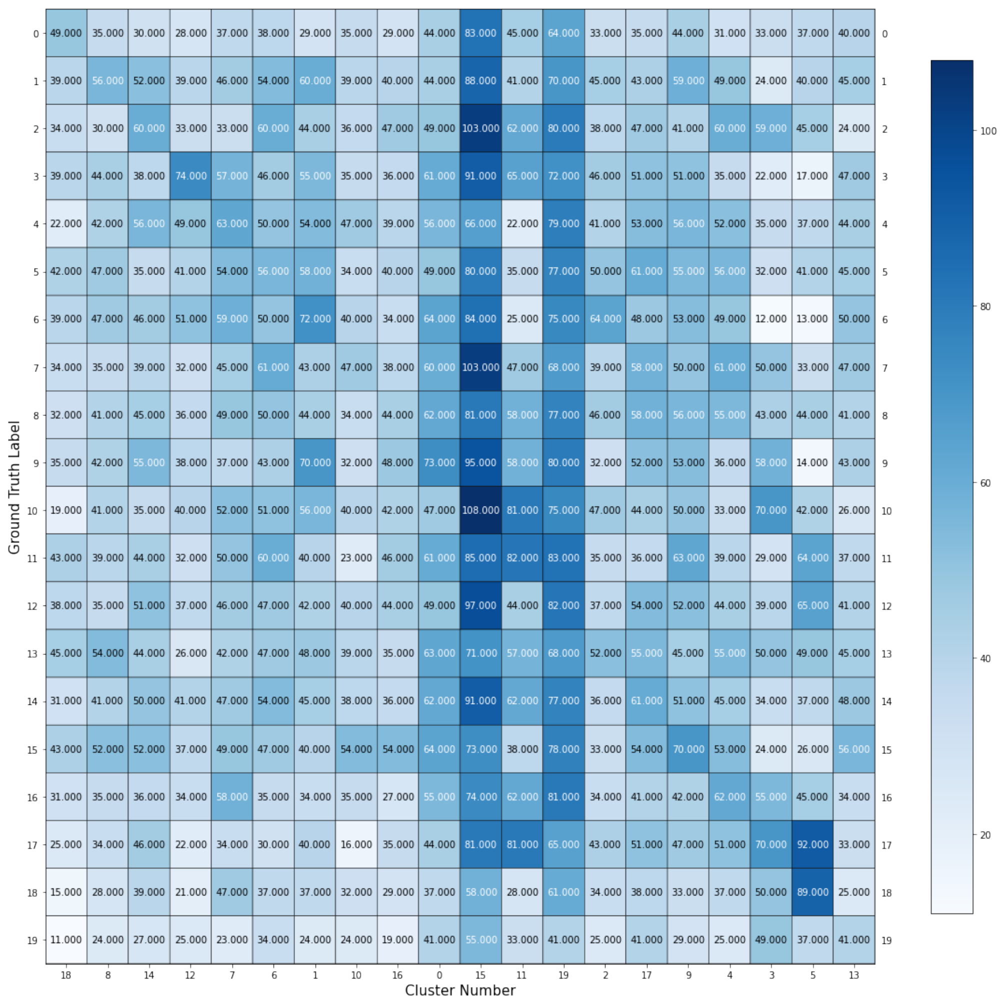
The best scores using both metrics came from **n_components=5**:

Cosine Metric	
Homogeneity score	0.5347926404764849
Completeness score	0.5549487009354954
V_measure score	0.5446842655666289
Adjusted Rand Index score	0.4167210932677909
Adjusted Mutual Information Score	0.5431756896373161

Euclidean metric	
Homogeneity score	0.008428056566878686
Completeness score	0.008490790660743977
V_measure score	0.008459307306639803
Adjusted Rand Index score	0.0014762057832842827
Adjusted Mutual Information Score	0.00524781170211466

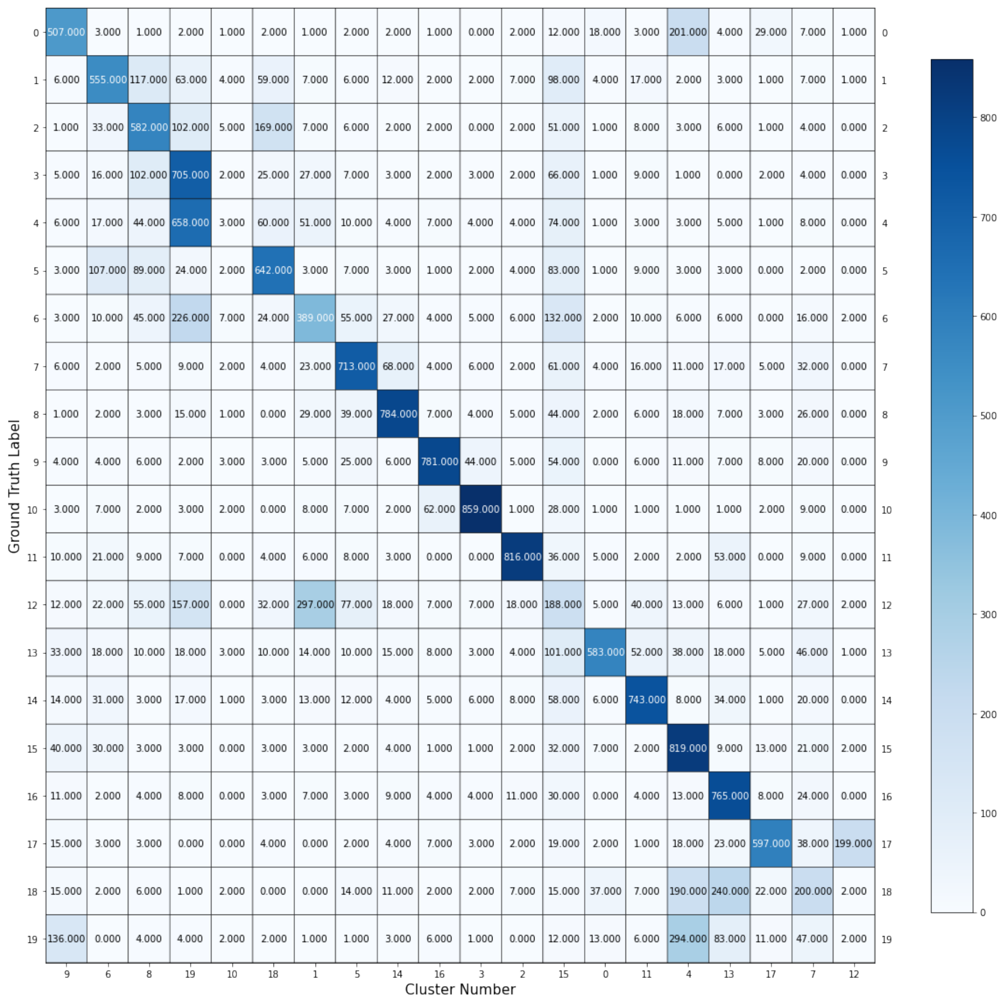
12 Question 12

Using n_components = 5 as determined above, we get the below contingency matrices – Figure 12 corresponds to the Euclidean metric, and Figure 13 corresponds to the cosine metric.



	18	8	14	12	7	6	1	10	16	0	15	11	19	2	17	9	4	3	5	13	
0	49.000	35.000	30.000	28.000	37.000	38.000	29.000	35.000	29.000	44.000	83.000	45.000	64.000	33.000	35.000	44.000	31.000	33.000	37.000	40.000	0
1	39.000	56.000	52.000	39.000	46.000	54.000	60.000	39.000	40.000	44.000	88.000	41.000	70.000	45.000	43.000	59.000	49.000	24.000	40.000	45.000	1
2	34.000	30.000	60.000	33.000	33.000	60.000	44.000	36.000	47.000	49.000	103.000	62.000	80.000	38.000	47.000	41.000	60.000	59.000	45.000	24.000	2
3	39.000	44.000	38.000	74.000	57.000	46.000	55.000	35.000	36.000	61.000	91.000	65.000	72.000	46.000	51.000	51.000	35.000	22.000	17.000	47.000	3
4	22.000	42.000	56.000	49.000	63.000	50.000	54.000	47.000	39.000	56.000	66.000	22.000	79.000	41.000	53.000	56.000	52.000	35.000	37.000	44.000	4
5	42.000	47.000	35.000	41.000	54.000	56.000	58.000	34.000	40.000	49.000	80.000	35.000	77.000	50.000	61.000	55.000	56.000	32.000	41.000	45.000	5
6	39.000	47.000	46.000	51.000	59.000	50.000	72.000	40.000	34.000	64.000	84.000	25.000	75.000	64.000	48.000	53.000	49.000	12.000	13.000	50.000	6
7	34.000	35.000	39.000	32.000	45.000	61.000	43.000	47.000	38.000	60.000	103.000	47.000	68.000	39.000	58.000	50.000	61.000	50.000	33.000	47.000	7
8	32.000	41.000	45.000	36.000	49.000	50.000	44.000	34.000	44.000	62.000	81.000	58.000	77.000	46.000	58.000	56.000	55.000	43.000	44.000	41.000	8
9	35.000	42.000	55.000	38.000	37.000	43.000	70.000	32.000	48.000	73.000	95.000	58.000	80.000	32.000	52.000	53.000	36.000	58.000	14.000	43.000	9
10	19.000	41.000	35.000	40.000	52.000	51.000	56.000	40.000	42.000	47.000	108.000	81.000	75.000	47.000	44.000	50.000	33.000	70.000	42.000	26.000	10
11	43.000	39.000	44.000	32.000	50.000	60.000	40.000	23.000	46.000	61.000	85.000	82.000	83.000	35.000	36.000	63.000	39.000	29.000	64.000	37.000	11
12	38.000	35.000	51.000	37.000	46.000	47.000	42.000	40.000	44.000	49.000	97.000	44.000	82.000	37.000	54.000	52.000	44.000	39.000	65.000	41.000	12
13	45.000	54.000	44.000	26.000	42.000	47.000	48.000	39.000	35.000	63.000	71.000	57.000	68.000	52.000	55.000	45.000	55.000	50.000	49.000	45.000	13
14	31.000	41.000	50.000	41.000	47.000	54.000	45.000	38.000	36.000	62.000	91.000	62.000	77.000	36.000	61.000	51.000	45.000	34.000	37.000	48.000	14
15	43.000	52.000	52.000	37.000	49.000	47.000	40.000	54.000	54.000	64.000	73.000	38.000	78.000	33.000	54.000	70.000	53.000	24.000	26.000	56.000	15
16	31.000	35.000	36.000	34.000	58.000	35.000	34.000	35.000	27.000	55.000	74.000	62.000	81.000	34.000	41.000	42.000	62.000	55.000	45.000	34.000	16
17	25.000	34.000	46.000	22.000	34.000	30.000	40.000	16.000	35.000	44.000	81.000	81.000	65.000	43.000	51.000	47.000	51.000	70.000	92.000	33.000	17
18	15.000	28.000	39.000	21.000	47.000	37.000	37.000	32.000	29.000	37.000	58.000	28.000	61.000	34.000	38.000	33.000	37.000	50.000	89.000	25.000	18
19	11.000	24.000	27.000	25.000	23.000	34.000	24.000	24.000	19.000	41.000	55.000	33.000	41.000	25.000	41.000	29.000	25.000	49.000	37.000	41.000	19

Figure 12: Confusion matrix for UMAP-reduced Kmeans with Euclidean metric



	9	6	8	19	10	18	1	5	14	16	3	2	15	0	11	4	13	17	7	12	
0	507.000	3.000	1.000	2.000	1.000	2.000	1.000	2.000	2.000	1.000	0.000	2.000	12.000	18.000	3.000	201.000	4.000	29.000	7.000	1.000	0
1	6.000	555.000	117.000	63.000	4.000	59.000	7.000	6.000	12.000	2.000	2.000	7.000	98.000	4.000	17.000	2.000	3.000	1.000	7.000	1.000	1
2	1.000	33.000	582.000	102.000	5.000	169.000	7.000	6.000	2.000	2.000	0.000	2.000	51.000	1.000	8.000	3.000	6.000	1.000	4.000	0.000	2
3	5.000	16.000	102.000	705.000	2.000	25.000	27.000	7.000	3.000	2.000	3.000	2.000	66.000	1.000	9.000	1.000	0.000	2.000	4.000	0.000	3
4	6.000	17.000	44.000	658.000	3.000	60.000	51.000	10.000	4.000	7.000	4.000	4.000	74.000	1.000	3.000	3.000	5.000	1.000	8.000	0.000	4
5	3.000	107.000	89.000	24.000	2.000	642.000	3.000	7.000	3.000	1.000	2.000	4.000	83.000	1.000	9.000	3.000	3.000	0.000	2.000	0.000	5
6	3.000	10.000	45.000	226.000	7.000	24.000	389.000	55.000	27.000	4.000	5.000	6.000	132.000	2.000	10.000	6.000	6.000	0.000	16.000	2.000	6
7	6.000	2.000	5.000	9.000	2.000	4.000	23.000	713.000	68.000	4.000	6.000	2.000	61.000	4.000	16.000	11.000	17.000	5.000	32.000	0.000	7
8	1.000	2.000	3.000	15.000	1.000	0.000	29.000	39.000	784.000	7.000	4.000	5.000	44.000	2.000	6.000	18.000	7.000	3.000	26.000	0.000	8
9	4.000	4.000	6.000	2.000	3.000	3.000	5.000	25.000	6.000	781.000	44.000	5.000	54.000	0.000	6.000	11.000	7.000	8.000	20.000	0.000	9
10	3.000	7.000	2.000	3.000	2.000	0.000	8.000	7.000	2.000	62.000	859.000	1.000	28.000	1.000	1.000	1.000	1.000	2.000	9.000	0.000	10
11	10.000	21.000	9.000	7.000	0.000	4.000	6.000	8.000	3.000	0.000	0.000	816.000	36.000	5.000	2.000	2.000	53.000	0.000	9.000	0.000	11
12	12.000	22.000	55.000	157.000	0.000	32.000	297.000	77.000	18.000	7.000	7.000	18.000	188.000	5.000	40.000	13.000	6.000	1.000	27.000	2.000	12
13	33.000	18.000	10.000	18.000	3.000	10.000	14.000	10.000	15.000	8.000	3.000	4.000	101.000	583.000	52.000	38.000	18.000	5.000	46.000	1.000	13
14	14.000	31.000	3.000	17.000	1.000	3.000	13.000	12.000	4.000	5.000	6.000	8.000	58.000	6.000	743.000	8.000	34.000	1.000	20.000	0.000	14
15	40.000	30.000	3.000	3.000	0.000	3.000	3.000	2.000	4.000	1.000	1.000	2.000	32.000	7.000	2.000	819.000	9.000	13.000	21.000	2.000	15
16	11.000	2.000	4.000	8.000	0.000	3.000	7.000	3.000	9.000	4.000	4.000	11.000	30.000	0.000	4.000	13.000	765.000	8.000	24.000	0.000	16
17	15.000	3.000	3.000	0.000	0.000	4.000	0.000	2.000	4.000	7.000	3.000	2.000	19.000	2.000	1.000	18.000	23.000	597.000	38.000	199.000	17
18	15.000	2.000	6.000	1.000	2.000	0.000	0.000	14.000	11.000	2.000	2.000	7.000	15.000	37.000	7.000	190.000	240.000	22.000	200.000	2.000	18
19	136.000	0.000	4.000	4.000	2.000	2.000	1.000	1.000	3.000	6.000	1.000	0.000	12.000	13.000	6.000	294.000	83.000	11.000	47.000	2.000	19

Figure 13: Confusion matrix for UMAP-reduced Kmeans with cosine metric

The cosine matrix is obviously much better, which corresponds with the greatly improved scores using this metric.

Examining the cosine metric more closely, we see that most categories were well identified, but some were more likely to be confused with each other. Tracing the connections, we can see that points with ground truth label 19 were likely to be identified with points labeled 0 or 15, and points from label 12 were misidentified as 1 or 6. In the original dataset, this means that points from `talk.religion.misc` were identified as `alt.atheism` or `soc.religion.christian`, and points from `sci.electronics` were identified as `comp.graphics` or `misc.forsale`. The religion groups form an intuitive grouping, and the confusion between the latter three groups perhaps indicates that selling hardware and electronics was common on these newsgroups.

13 Question 13

Using the best hyperparameter combination from Q11/12, n_components=5 and the cosine metric, we reduced the Tfidf matrix with UMAP. The scores for agglomerative clustering using 'ward' and 'single' linkage on this data are shown below:

Ward linkage	
Homogeneity score	0.524473300498607
Completeness score	0.5479912976714623
V_measure score	0.53597443687123
Adjusted Rand Index score	0.3940501791969448
Adjusted Mutual Information Score	0.5344316766896055

Single linkage	
Homogeneity score	0.022917457355135236
Completeness score	0.3464156710750703
V_measure score	0.04299081646294147
Adjusted Rand Index score	0.00046761208355549543
Adjusted Mutual Information Score	0.03748616963750802

It's clear that ward linkage performs significantly better on the dataset. Single linkage uses the minimum distance between all points in each cluster to determine whether to merge them or not, so it's a lot more sensitive to outliers and tends to result in uneven cluster sizes. Ward linkage on the other hand, minimizes the variance between pairs of clusters, which can produce more regular cluster sizes and hence better results for our particular dataset.

14 Question 14

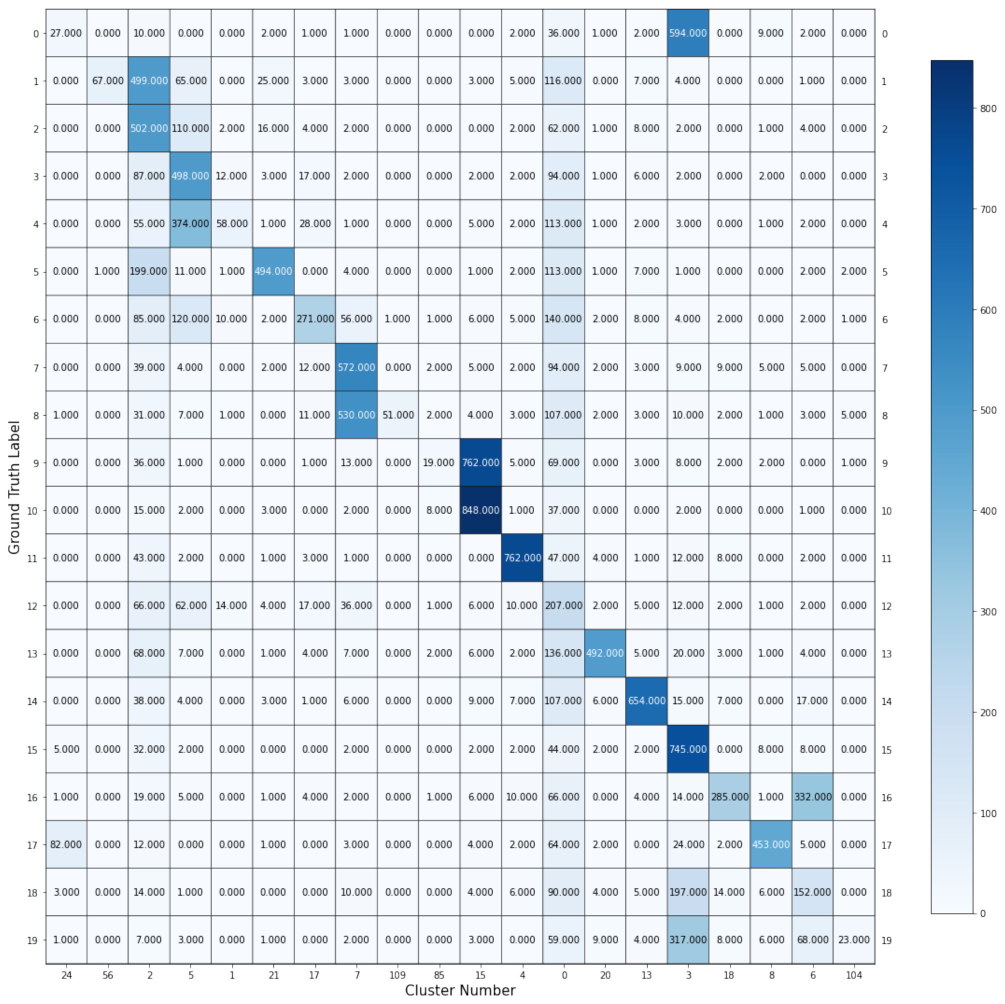
In this section we investigate the DBSCAN and HDBSCAN clustering algorithms. We kept UMAP with n_components=5 and the cosine metric as the dimensionality reduction algorithm. The HDBSCAN min_cluster_size was set to 100, and other hyperparameters were tuned to improve the performance of each of these algorithms. We performed hyperparameter tuning on DBSCAN and HDBSCAN for epsilon values of [0.01,0.05,0.1,0.3,0.5,0.7,0.9,1], metric types of [Euclidean, cosine, Manhattan] (if applicable), and minimum sample values of [10, 20, 50, 70, 100]. The following table depicts the scores for each clustering algorithm for the hyperparameter combination that resulted in the best average score across all 5 clustering metrics:

DBSCAN: Epsilon = 0.5, Metric = Manhattan, Minimum samples = 10	
Homogeneity score	0.5407607595568981
Completeness score	0.45471096739764455
V_measure score	0.49401673890037173
Adjusted Rand Index score	0.2967260971944622
Adjusted Mutual Information Score	0.47920988791262803

HDBSCAN: Epsilon = 0.3, Metric = Euclidean, Minimum samples = 10	
Homogeneity score	0.43889528158878366
Completeness score	0.5288819318072416
V_measure score	0.47970500064413385
Adjusted Rand Index score	0.2565568616115164
Adjusted Mutual Information Score	0.4780526456871693

15 Question 15

DBSCAN was the better performing clustering algorithm. The associated contingency matrix is shown below:



	24	56	2	5	1	21	17	7	109	85	15	4	0	20	13	3	18	8	6	104
0	27.000	0.000	10.000	0.000	0.000	2.000	1.000	1.000	0.000	0.000	0.000	2.000	36.000	1.000	2.000	594.000	0.000	9.000	2.000	0.000
1	0.000	67.000	499.000	65.000	0.000	25.000	3.000	3.000	0.000	0.000	3.000	5.000	116.000	0.000	7.000	4.000	0.000	0.000	1.000	0.000
2	0.000	0.000	502.000	110.000	2.000	16.000	4.000	2.000	0.000	0.000	0.000	2.000	62.000	1.000	8.000	2.000	0.000	1.000	4.000	0.000
3	0.000	0.000	87.000	498.000	12.000	3.000	17.000	2.000	0.000	0.000	2.000	2.000	94.000	1.000	6.000	2.000	0.000	2.000	0.000	0.000
4	0.000	0.000	55.000	374.000	58.000	1.000	28.000	1.000	0.000	0.000	5.000	2.000	113.000	1.000	2.000	3.000	0.000	1.000	2.000	0.000
5	0.000	1.000	199.000	11.000	1.000	494.000	0.000	4.000	0.000	0.000	1.000	2.000	113.000	1.000	7.000	1.000	0.000	0.000	2.000	2.000
6	0.000	0.000	85.000	120.000	10.000	2.000	271.000	56.000	1.000	1.000	6.000	5.000	140.000	2.000	8.000	4.000	2.000	0.000	2.000	1.000
7	0.000	0.000	39.000	4.000	0.000	2.000	12.000	572.000	0.000	2.000	5.000	2.000	94.000	2.000	3.000	9.000	9.000	5.000	5.000	0.000
8	1.000	0.000	31.000	7.000	1.000	0.000	11.000	530.000	51.000	2.000	4.000	3.000	107.000	2.000	3.000	10.000	2.000	1.000	3.000	5.000
9	0.000	0.000	36.000	1.000	0.000	0.000	1.000	13.000	0.000	19.000	762.000	5.000	69.000	0.000	3.000	8.000	2.000	2.000	0.000	1.000
10	0.000	0.000	15.000	2.000	0.000	3.000	0.000	2.000	0.000	8.000	848.000	1.000	37.000	0.000	0.000	2.000	0.000	0.000	1.000	0.000
11	0.000	0.000	43.000	2.000	0.000	1.000	3.000	1.000	0.000	0.000	0.000	762.000	47.000	4.000	1.000	12.000	8.000	0.000	2.000	0.000
12	0.000	0.000	66.000	62.000	14.000	4.000	17.000	36.000	0.000	1.000	6.000	10.000	207.000	2.000	5.000	12.000	2.000	1.000	2.000	0.000
13	0.000	0.000	68.000	7.000	0.000	1.000	4.000	7.000	0.000	2.000	6.000	2.000	136.000	492.000	5.000	20.000	3.000	1.000	4.000	0.000
14	0.000	0.000	38.000	4.000	0.000	3.000	1.000	6.000	0.000	0.000	9.000	7.000	107.000	6.000	654.000	15.000	7.000	0.000	17.000	0.000
15	5.000	0.000	32.000	2.000	0.000	0.000	0.000	2.000	0.000	0.000	2.000	2.000	44.000	2.000	2.000	745.000	0.000	8.000	8.000	0.000
16	1.000	0.000	19.000	5.000	0.000	1.000	4.000	2.000	0.000	1.000	6.000	10.000	66.000	0.000	4.000	14.000	285.000	1.000	332.000	0.000
17	82.000	0.000	12.000	0.000	0.000	1.000	0.000	3.000	0.000	0.000	4.000	2.000	64.000	2.000	0.000	24.000	2.000	453.000	5.000	0.000
18	3.000	0.000	14.000	1.000	0.000	0.000	0.000	10.000	0.000	0.000	4.000	6.000	90.000	4.000	5.000	197.000	14.000	6.000	152.000	0.000
19	1.000	0.000	7.000	3.000	0.000	1.000	0.000	2.000	0.000	0.000	3.000	0.000	59.000	9.000	4.000	317.000	8.000	6.000	68.000	23.000

Figure 14: Contingency Matrix for DBSCAN on 20 Class Data

The contingency matrix indicates that there must be at least 96 clusters since that's the largest number shown on the x-axis ticks. By looking at the number of unique labels in the trained DBSCAN model, we are able to determine that there are actually 168 clusters formed by the algorithm (excluding the -1 labels). This number is high because of our low value for the mimimum samples hyperparameter, so most of these cluters contain few data points. However, most of the data points were captured by the larger clusters shown in the contingency matrix, which resulted in acceptable clustering metric scores.

A label of -1 indicates that the data point was classified as noise, and was not assigned to any cluster. This cluster label is represented as cluster number 0 in the contingency matrix. We can tell because this cluster contains many data points from each of the 20 classes.

16 Question 16

We used the “read_csv” function of the Pandas library to acquire the BBC dataset as a dataframe. After converting the written class labels to numerical target values for each document, we then used our established pipeline to vectorize (excluding stopwords/punctuation) then normalize the data and analyze the dimensions of the resulting TF-IDF matrix. As with the 20 Newsgroups dataset, we then reduced dimensionality with UMAP and searched for the optimal parameter configuration for multiple clustering algorithms. Starting with KMeans, we tested how our five aforementioned accuracy metrics vary with the number of principle components used in clustering for both Euclidean (Figure 15) and Cosine distance metrics (Figure 16). The best combination was n_components = 10 with the Euclidean distance metric.

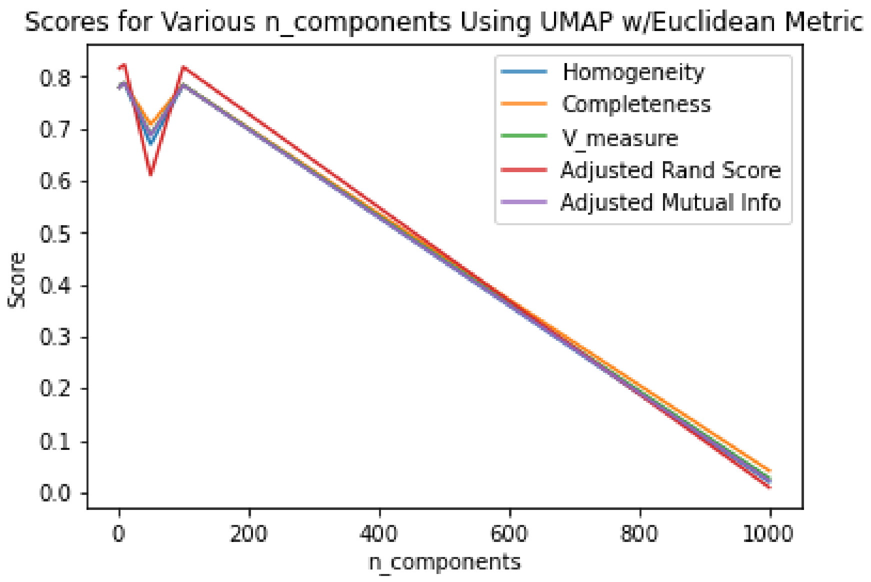


Figure 15: Variation in accuracy metrics by number of components for Kmeans Clustering and Euclidean distance

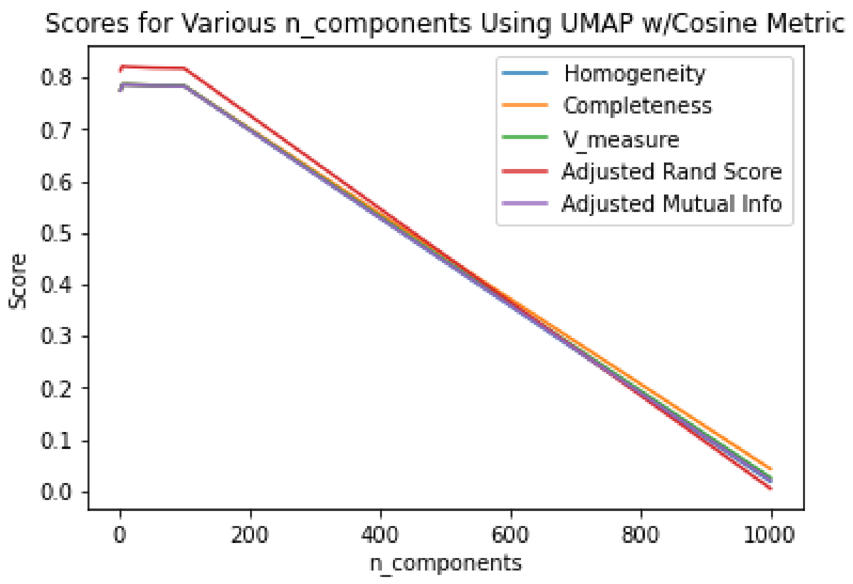


Figure 16: Variation in accuracy metrics by number of components for Kmeans Clustering and Cosine distance

We then tested the accuracy of the DBSCAN clustering algorithm for both distance metrics. We cycled through values from .01 to 1 for epsillon and from 1 to 50 for the minimum number of points per cluster to find the best combination of hyperparameters for our analysis. We used 3-dimensional surface plots to visualize how these parameters in concert affected our five accuracy metrics and the estimated number of clusters generated by DBSCAN (Figure 17).

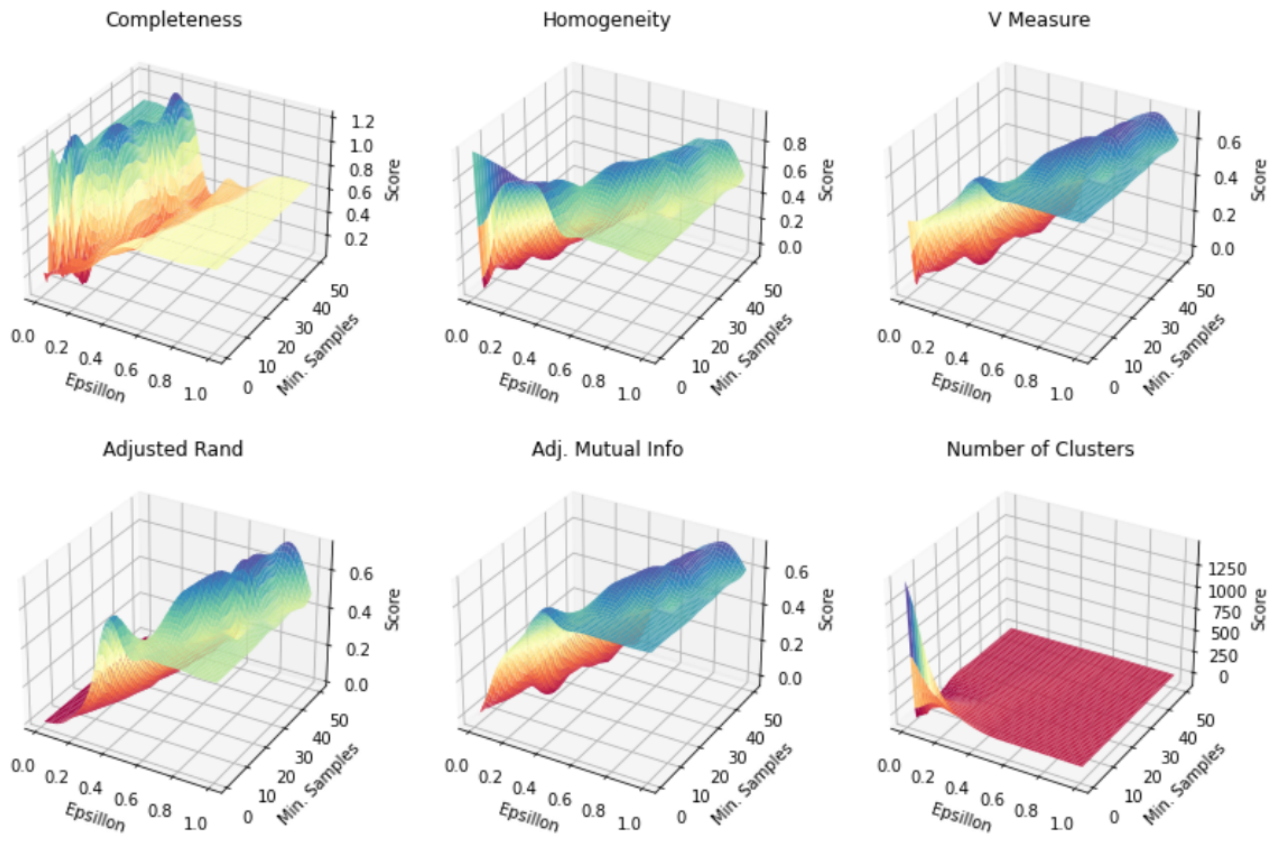


Figure 17: Variation in accuracy metrics corresponding to changes in epsilon and minimum cluster metrics for DBSCAN clustering of UMAP reduced data with Cosine distance metric (Euclidean metric results visually similar)

Avoiding the low values of epsilon and minimum samples that resulted in excessively high numbers of estimated clusters, we found the combination of parameters that maximized the average of the accuracy metrics to be 0.8 for epsilon and 35 for minimum samples. We then compared the results of each of our protocols (Kmeans with Euclidean/Cosine, DBSCAN with Euclidean/Cosine) to find the best reduction and clustering technique for our data. We report the five accuracy metrics for each protocol below:

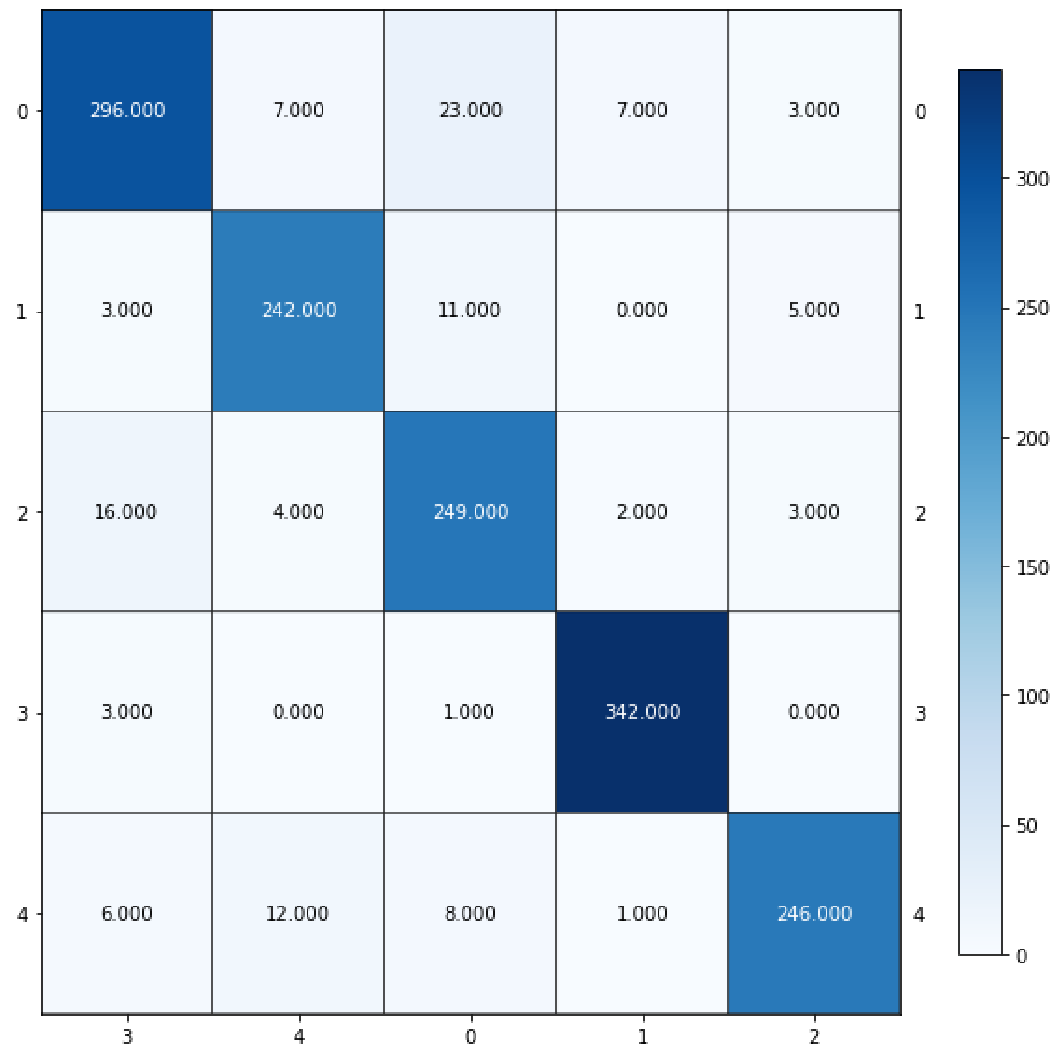
Accuracy Metrics for DBSCAN, Metric = Euclidean	
Homogeneity score	0.7896799513244787
Completeness score	0.667782538810624
V_measure score	0.7236336939202158
Adjusted Rand Index score	0.7108732847743878
Adjusted Mutual Information Score	0.7221232251101021

Accuracy Metrics for DBSCAN, Metric = Cosine	
Homogeneity score	0.7896799513244787
Completeness score	0.667782538810624
V_measure score	0.7236336939202158
Adjusted Rand Index score	0.7108732847743878
Adjusted Mutual Information Score	0.7221232251101021

Accuracy Metrics for KMeans, Metric = Euclidean	
Homogeneity score	0.7874360944663693
Completeness score	0.7874929063651126
V_measure score	0.7874644993910623
Adjusted Rand Index score	0.8229763601775667
Adjusted Mutual Information Score	0.786746935945491

Accuracy Metrics for KMeans, Metric = Cosine	
Homogeneity score	0.7863911146778318
Completeness score	0.7863053396249526
V_measure score	0.7863482248123015
Adjusted Rand Index score	0.8206748429226303
Adjusted Mutual Information Score	0.785626959948192

The best scoring protocol was Kmeans with Euclidean distance metric and n_components set to 10. The contingency matrix for this clustering is shown in Figure 18 below



	3	4	0	1	2	
0	296.000	7.000	23.000	7.000	3.000	0
1	3.000	242.000	11.000	0.000	5.000	1
2	16.000	4.000	249.000	2.000	3.000	2
3	3.000	0.000	1.000	342.000	0.000	3
4	6.000	12.000	8.000	1.000	246.000	4

Figure 18: Contingency Matrix for optimal clustering of BBC News data

Our pipeline and parameter selections produce a fairly good clustering through both DBSCAN and KMeans. One factor that may affect the accuracy of our analysis is the relatively small size of the BBC dataset when compared to the 20 Newsgroups dataset. With more data, our clustering would likely become more accurate.