ECE 219 Project 2: Data Representations and Clustering

Sarah Wilen UCLA ECE Department

swilen1@g.ucla.edu
UID: 305070897

Marco Venegas UCLA ECE Department

venmarco310@g.ucla.edu
UID: 805461921

Henry Peters UCLA CS Department

hpeters@g.ucla.edu UID: 205704526

1. Question 1

The TD-IDF matrix shape: (4732, 17131)

2. Question 2

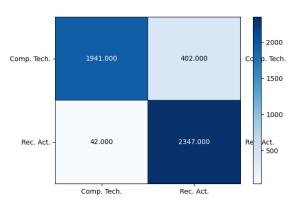


Figure 1. Contigency Matrix for means (k=2)

The contingency matrix does not have to be square. Let's say there are 2 underlying classes and we want to cluster them into 3 clusters. Therefore, contingency matrix A would have dimensions 2x3. In our example, we have a square contingency matrix because we have 2 underlying classes and chose to have 2 clusters.

3. Question 3

Homogeneity	0.589
Completeness	0.601
V-measure	0.595
Adjusted Rand Index	0.660
Adjusted Mutual Information Score	0.595

Table 1. Scores for Kmeans(k=2)

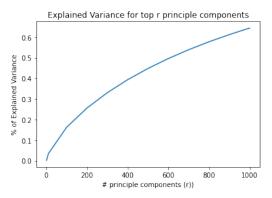


Figure 2. Explained Variance vs Principle Components

5. Question 5

Looking at the graphs below, a good choice for SVD would be r=20 and for NMF a good choice would be r=20

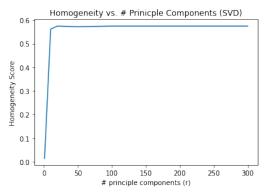


Figure 3. Homogeneity (SVD)

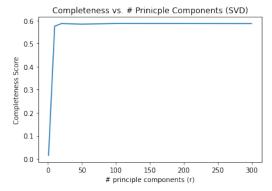


Figure 4. Completeness (SVD)

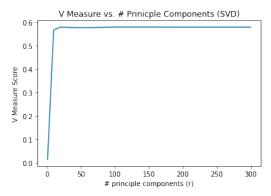


Figure 5. V-measure (SVD)

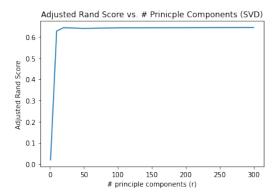


Figure 6. Adjusted Rand Index (SVD)

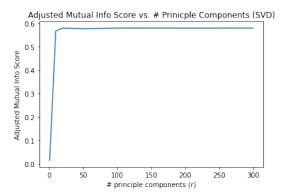


Figure 7. Adjusted Mutual Info Score (SVD)

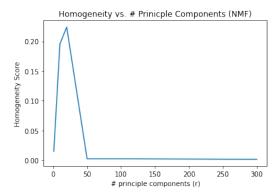


Figure 8. Homogeneity (NMF)

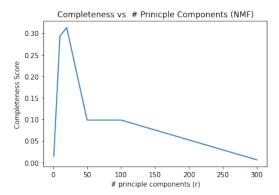


Figure 9. Completeness (NMF)

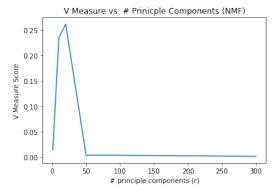


Figure 10. V-measure (NMF)

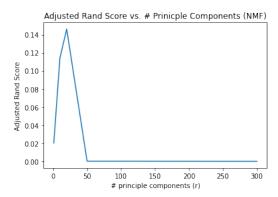


Figure 11. Adjusted Rand Index (NMF)

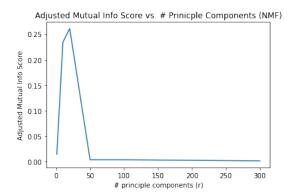


Figure 12. Adjusted Mutual Info Score (NMF)

As the number of principal components increases for SVD, the score increases and then converges to a steady score. As we increase the number of principal components, we think that the score should increase too as the number of features is larger and we have a more descriptive feature space. However, we also know that as we increase the number of principal components, there is more sparsity in the data, which leads to the curse of dimensionality, so the score does not increase proportionally as much as the increase in features from more principal components may suggest an improvement.

As the number of components increases for NMF, there is a sharp increase and then a sharp decrease in score. This makes sense due to the curse of dimensionality. The higher the dimension, the larger the distance all points are away from each other, hence, the score is a lot lower. Therefore, we found the number of components that not only retains a good amount of feature information but also keeps the dimension low enough so that we don't encounter the curse of dimensionality.

7. Question 7

On average, no we do not see a better performance, which goes to reinforces our notions in question 6 about non-monotonic behavior.

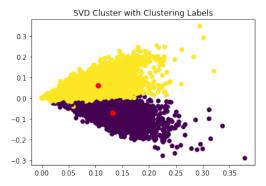


Figure 13. Data Visualization with SVD

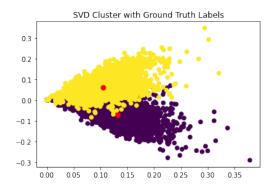


Figure 14. Data Visualization with SVD Ground Truth

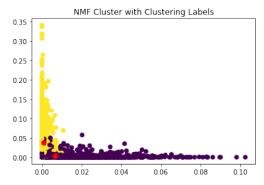


Figure 15. Data Visualization with NMF

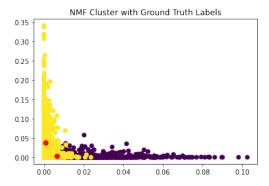


Figure 16. Data Visualization with NMF Ground Truth

One thing worth noting is that none of the clusterings look spherical, as expected. Both SVD and NMF clustering take on really odd shapes, which can seem like an uneven distribution. We also see that both clustering classes have really close boundaries when labeling the data, which can explain our poor measures shown in our earlier report regarding the curse of dimensionality, especially with NMF clustering. We can see the variance for both methods is not too good, which can be seen as evidence of the noise that is introduced in our models.

10. Question 10

We started by sweeping the number of components between 1-300 for SVD and NMF on the new 20 class data in order to determine the best number of principal components to use. We noticed as we changed the number of components closer to the optimal number, the contingency matrix started out scattered with values all over the table but became more defined on the diagonal from the top left to the bottom right as we approached the optimal value, which is what we are looking for. An example of the contingency matrices for the different number of components for SVD is shown. We did the same analysis for NMF but did not include the example because it is similar to SVD.

We ultimately determined that n_components = 300 bore best for SVD and n_components of 20 was best for NMF. With this said, the scores for SVD's n_components = 300, 100, 50, and 20 were close together, so we ended up using 20 since we don't want to use high dimensions in case of (1) cures of dimensionality and (2) long processing time.

Optimal n_components: SVD - 20, NMF - 20

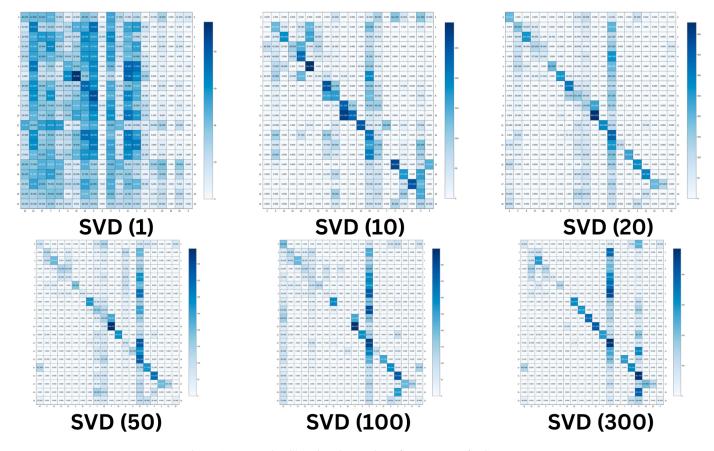


Figure 17. Example: Choosing the number of components for SVD

0 -	56.000	1.000	0.000	0.000	0.000	0.000	0.000	2.000	64.000	0.000	0.000	0.000	80.000	0.000	0.000	78.000	0.000	0.000	199.000	0.000	0
1 -	0.000	121.000	76.000	43.000	0.000	2.000	5.000	0.000	20.000	0.000	0.000	0.000	292.000	0.000	7.000	0.000	0.000	0.000	18.000	0.000	1
2 -	0.000	50.000	272.000	65.000	9.000	14.000	0.000	0.000	17.000	0.000	0.000	0.000	161.000	0.000	0.000	0.000	0.000	0.000	2.000	1.000	2
3 -	0.000	73.000	35.000	116.000	133.000	3.000	6.000	3.000	29.000	0.000	0.000	0.000	185.000	0.000	3.000	0.000	0.000	0.000	1.000	3.000	3
4 -	0.000	82.000	2.000	50.000	69.000	6.000	12.000	0.000	23.000	0.000	0.000	0.000	325.000	0.000	3.000	0.000	0.000	0.000	4.000	2.000	4
5 -	0.000	87.000	54.000	4.000	1.000	135.000	1.000	0.000	16.000	0.000	0.000	0.000	276.000	0.000	3.000	0.000	0.000	0.000	5.000	11.000	5
6 -	0.000	49.000	17.000	36.000	30.000	1.000	2.000	22.000	18.000	0.000	8.000	0.000	394.000	0.000	0.000	0.000	0.000	0.000	8.000	0.000	6
7 -	0.000	24.000	2.000	0.000	2.000	1.000	1.000	316.000	56.000	0.000	0.000	0.000	164.000	0.000	0.000	0.000	6.000	0.000	22.000	0.000	7
8 -	0.000	21.000	0.000	0.000	2.000	2.000	0.000	162.000	163.000	0.000	2.000	0.000	223.000	0.000	1.000	0.000	0.000	0.000	22.000	0.000	8
9 -	0.000	24.000	0.000	0.000	0.000	0.000	0.000	0.000	50.000	0.000	277.000	0.000	228.000	0.000	2.000	0.000	0.000	0.000	16.000	0.000	9
10 -	0.000	14.000	0.000	0.000	0.000	0.000	0.000	1.000	7.000	54.000	416.000	0.000	100.000	0.000	0.000	0.000	0.000	0.000	8.000	0.000	10
11 -	0.000	7.000	8.000	0.000	0.000	0.000	68.000	0.000	57.000	0.000	0.000	154.000	121.000	0.000	4.000	0.000	10.000	0.000	62.000	104.000	11
12 -	0.000	62.000	4.000	7.000	5.000	0.000	20.000	21.000	44.000	0.000	1.000	0.000	421.000	0.000	5.000	0.000	0.000	0.000	1.000	0.000	12
13 -	0.000	28.000	2.000	0.000	0.000	0.000	0.000	1.000	61.000	0.000	0.000	0.000	264.000	90.000	1.000	0.000	0.000	0.000	147.000	0.000	13
14 -	0.000	20.000	0.000	0.000	0.000	0.000	0.000	0.000	54.000	0.000	0.000	0.000	229.000	0.000	266.000	0.000	1.000	0.000	23.000	0.000	14
15 -	0.000	17.000	0.000	0.000	0.000	0.000	0.000	1.000	7.000	0.000	0.000	0.000	86.000	0.000	0.000	264.000	0.000	0.000	224.000	0.000	15
16 -	0.000	4.000	0.000	3.000	0.000	0.000	1.000	2.000	72.000	0.000	0.000	0.000	94.000	0.000	4.000	0.000	282.000	0.000	84.000	0.000	16
17 -	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	28.000	0.000	0.000	0.000	135.000	0.000	0.000	2.000	0.000	331.000	67.000	0.000	17
18 -	1.000	4.000	0.000	0.000	0.000	0.000	0.000	1.000	94.000	0.000	0.000	1.000	61.000	0.000	3.000	1.000	1.000	0.000	298.000	0.000	18
19 -	34.000	2.000	0.000	0.000	0.000	0.000	0.000	1.000	50.000	0.000	0.000	0.000	80.000	0.000	0.000	77.000	7.000	0.000	126.000	0.000	19
	4	10	Ó	11	7	5	19	i	8	16	9	15	6	3	18	12	13	2	14	17	I

- 250

Figure 18. 20 Cat - SVD Contingency Matrix

Homogeneity	0.342
Completeness	0.431
V-measure	0.381
Adjusted Rand Index	0.100
Adjusted Mutual Information Score	0.378

Table 2. Scores for SVD K-means(k=20)

10.2. NMF

0 -	62.000	2.000	0.000	0.000	0.000	0.000	133.000	2.000	62.000	120.000	0.000	0.000	0.000	0.000	0.000	88.000	11.000	0.000	0.000	0.000	0
1 -	0.000	149.000	55.000	56.000	1.000	9.000	221.000	0.000	24.000	57.000	0.000	1.000	1.000	0.000	10.000	0.000	0.000	0.000	0.000	0.000	1
2 -	0.000	59.000	246.000	66.000	10.000	18.000	119.000	0.000	21.000	50.000	0.000	0.000	2.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	2
3 -	0.000	70.000	25.000	116.000	77.000	1.000	156.000	2.000	32.000	26.000	0.000	2.000	56.000	0.000	3.000	0.000	0.000	0.000	24.000	0.000	3
4 -	0.000	82.000	2.000	86.000	54.000	0.000	232.000	0.000	24.000	64.000	0.000	2.000	29.000	0.000	3.000	0.000	0.000	0.000	0.000	0.000	4
5 -	0.000	95.000	30.000	6.000	1.000	196.000	209.000	0.000	19.000	31.000	0.000	4.000	0.000	0.000	2.000	0.000	0.000	0.000	0.000	0.000	5
6 -	0.000	48.000	16.000	35.000	32.000	1.000	367.000	23.000	20.000	29.000	7.000	0.000	7.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	6
7 -	0.000	30.000	2.000	0.000	3.000	2.000	146.000	278.000	68.000	60.000	0.000	0.000	0.000	0.000	0.000	0.000	5.000	0.000	0.000	0.000	7
8 -	0.000	25.000	0.000	0.000	4.000	3.000	223.000	65.000	166.000	110.000	1.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	8
9 -	0.000	25.000	0.000	0.000	0.000	0.000	157.000	0.000	48.000	139.000	227.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	9
10 -	0.000	15.000	0.000	0.000	0.000	0.000	90.000	0.000	15.000	52.000	428.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	10
11 -	0.000	11.000	6.000	1.000	0.000	0.000	135.000	1.000	78.000	56.000	0.000	281.000	0.000	0.000	4.000	0.000	22.000	0.000	0.000	0.000	11
12 -	0.000	74.000	3.000	11.000	4.000	0.000	345.000	20.000	53.000	73.000	0.000	1.000	2.000	0.000	4.000	0.000	1.000	0.000	0.000	0.000	12
13 -	0.000	37.000	1.000	0.000	0.000	0.000	324.000	1.000	70.000	68.000	0.000	0.000	0.000	91.000	2.000	0.000	0.000	0.000	0.000	0.000	13
14 -	0.000	18.000	0.000	0.000	0.000	0.000	188.000	0.000	57.000	89.000	0.000	0.000	0.000	0.000	240.000	0.000	1.000	0.000	0.000	0.000	14
15 -	0.000	20.000	0.000	0.000	0.000	0.000	204.000	0.000	8.000	53.000	0.000	0.000	0.000	0.000	0.000	314.000	0.000	0.000	0.000	0.000	15
16 -	0.000	4.000	0.000	3.000	0.000	0.000	95.000	2.000	82.000	83.000	0.000	0.000	0.000	0.000	3.000	0.000	274.000	0.000	0.000	0.000	16
17 -	0.000	4.000	0.000	0.000	0.000	0.000	148.000	0.000	27.000	65.000	0.000	0.000	0.000	0.000	0.000	2.000	1.000	193.000	0.000	124.000	17
18 -	1.000	4.000	0.000	0.000	0.000	0.000	174.000	1.000	129.000	108.000	0.000	1.000	0.000	0.000	4.000	1.000	42.000	0.000	0.000	0.000	18
19 -	35.000	4.000	0.000	0.000	0.000	0.000	126.000	1.000	55.000	59.000	0.000	0.000	0.000	0.000	0.000	90.000	7.000	0.000	0.000	0.000	19
	9	5	19	Ó	10	16	2	14	12	i	18	3	17	4	15	6	13	8	7	11	

Figure 19. 20 Cat - NMF Contingency Matrix

Homogeneity	0.301
Completeness	0.379
V-measure	0.335
Adjusted Rand Index	0.079
Adjusted Mutual Information Score	0.331

Table 3. Scores for NMF K-means(k=20)

11.0.1 5 Components Euclidean

0 -	30.000	28.000	5.000	31.000	24.000	35.000	24.000	25.000	30.000	26.000	11.000	28.000	28.000	17.000	27.000	32.000	25.000	32.000	22.000	0.000	0
1 -	39.000	37.000	15.000	44.000	28.000	32.000	26.000	24.000	37.000	32.000	6.000	37.000	31.000	37.000	33.000	44.000	29.000	30.000	23.000	0.000	1
2 -	22.000	27.000	35.000	37.000	37.000	41.000	20.000	32.000	29.000		15.000	37.000	30.000	37.000	26.000	42.000	29.000	25.000	30.000	10.000	2
3 -	30.000	29.000	22.000	47.000	33.000	31.000		36.000	34.000	37.000	5.000	37.000	23.000		38.000	37.000	36.000	27.000		0.000	3
4 -	28.000	22.000	5.000	49.000	41.000	40.000	18.000	28.000	34.000	27.000	30.000	40.000	24.000	34.000	33.000	38.000	35.000	31.000	21.000	0.000	4
5 -	33.000	32.000	16.000	37.000	23.000	52.000	25.000	31.000	28.000	34.000	20.000	33.000	32.000	31.000	30.000	43.000	36.000	38.000	19.000	0.000	5
6 -	19.000	37.000	11.000	39.000	26.000	35.000	38.000	30.000	40.000	41.000	3.000	34.000	25.000	35.000	40.000	37.000	28.000	38.000	29.000	0.000	6
7 -	27.000	37.000	12.000	39.000	20.000	44.000	24.000	36.000	45.000	38.000	2.000	39.000	25.000	27.000	28.000	42.000	38.000	37.000	34.000	0.000	7
8 -	35.000	34.000	4.000	40.000	20.000	40.000	25.000	29.000	42.000	28.000	18.000	38.000	39.000	29.000	32.000	41.000	29.000	41.000	34.000	0.000	8
9 -	25.000	30.000	19.000	30.000	38.000	31.000	30.000	27.000	29.000	56.000	26.000	37.000	29.000	23.000	28.000	34.000	34.000	36.000	35.000	0.000	9
10 -	20.000	24.000	18.000	35.000	22.000	47.000	25.000	25.000	34.000	39.000	42.000	28.000	37.000	31.000	44.000	37.000	33.000	38.000	21.000	0.000	10
11 -	36.000	22.000	4.000	33.000	28.000	44.000	28.000	22.000	38.000	30.000	32.000	42.000	21.000	38.000	32.000	45.000	38.000	39.000	23.000	0.000	11
12 -	23.000	33.000	19.000	35.000	18.000	30.000	29.000	28.000	37.000	39.000	22.000	36.000	41.000	39.000	33.000	37.000	30.000	36.000	26.000	0.000	12
13 -	32.000	23.000	4.000	34.000	24.000	44.000	27.000	24.000	41.000	47.000	15.000	35.000	27.000	42.000	43.000	34.000	28.000	39.000	31.000	0.000	13
14 -	38.000	28.000	1.000	39.000	36.000	25.000	25.000	22.000	38.000	43.000	15.000	38.000	29.000	41.000	46.000	35.000	32.000	41.000	21.000	0.000	14
15 -	27.000	25.000	17.000	32.000	38.000	49.000	34.000	28.000	40.000	37.000	0.000	43.000	23.000	31.000	29.000	48.000	33.000	36.000	29.000	0.000	15
16 -	22.000	21.000	31.000	38.000	21.000	40.000	24.000	31.000	28.000	42.000	32.000	35.000	15.000	29.000	26.000	30.000	35.000	31.000	15.000	0.000	16
17 -	25.000	23.000	28.000	27.000	24.000	31.000	20.000	33.000	43.000	29.000	24.000	42.000	37.000	37.000	26.000	24.000	24.000	43.000	24.000	0.000	17
18 -	20.000	23.000	10.000	29.000	16.000	18.000	21.000	12.000	32.000	29.000	28.000	29.000	33.000	23.000	26.000	33.000	20.000	32.000	31.000	0.000	18
19 -	16.000	17.000	10.000	24.000	21.000	22.000	14.000	22.000	26.000	23.000	14.000	24.000	22.000	23.000	21.000	26.000	16.000	24.000	12.000	0.000	19
	11	16	14	6	12	7	5	2	18	Ó	i	10	8	13	4	9	19	3	15	17	I

Figure 20. UMAP (5 components, euclidean) Contingency Matrix

Homogeneity	0.010
Completeness	0.010
V-measure	0.010
Adjusted Rand Index	0.001
Adjusted Mutual Information Score	0.005

Table 4. Scores for UMAP (5 components, euclidean)

11.0.2 20 Components Euclidean

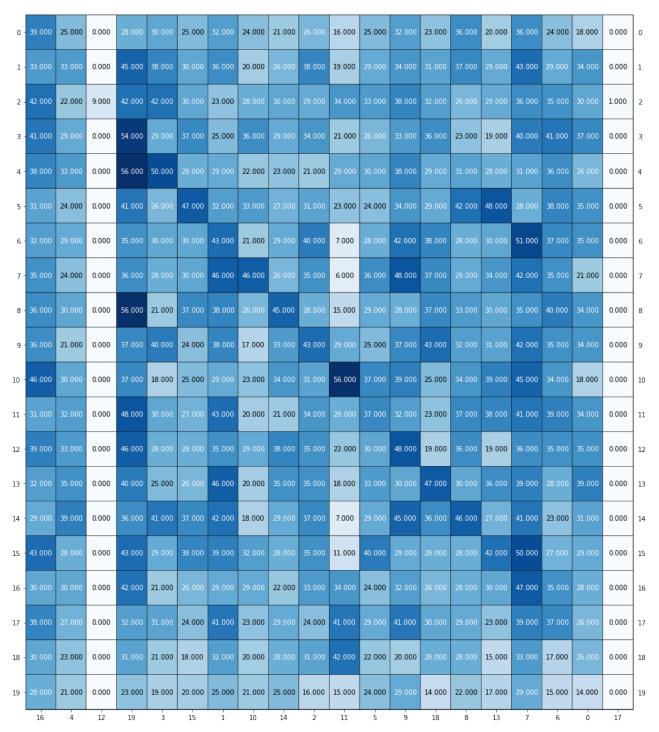
0 -	28.000	23.000	0.000	28.000	23.000	27.000	18.000	26.000	34.000	28.000	0.000	28.000	27.000	32.000	24.000	38.000	19.000	33.000	21.000	23.000	0
1 -	31.000	33.000	0.000	25.000	35.000	32.000	36.000	32.000	29.000	30.000	5.000	36.000	38.000	36.000	32.000	44.000	18.000	36.000	30.000	26.000	1
2 -	30.000	27.000	10.000	31.000	36.000	21.000	31.000	34.000	33.000	24.000	24.000	39.000	33.000	36.000	19.000	45.000	25.000	38.000	28.000	27.000	2
3 -	27.000	30.000	0.000	49.000	48.000	21.000	36.000	25.000	33.000	32.000	0.000	36.000	40.000	32.000	28.000	42.000	25.000	32.000	30.000	24.000	3
4 -	40.000	27.000	0.000	27.000	50.000	33.000	33.000	26.000	38.000	31.000	13.000	35.000	23.000	39.000	28.000	37.000	15.000	35.000	28.000	20.000	4
5 -	20.000	25.000	0.000	32.000	32.000	36.000	32.000	27.000	33.000	37.000	15.000	35.000	31.000	52.000	17.000	47.000	13.000	50.000	33.000	26.000	5
6 -	23.000	26.000	0.000	43.000	33.000	18.000	40.000	32.000	34.000	35.000	11.000	24.000	41.000	40.000	24.000	41.000	7.000	44.000	33.000	36.000	6
7 -	15.000	39.000	0.000	50.000	34.000	25.000	20.000	42.000	39.000	26.000	16.000	43.000	36.000	38.000	24.000	47.000	5.000	49.000	25.000	21.000	7
8 -	20.000	33.000	0.000	38.000	32.000	36.000	35.000	24.000	41.000	33.000	13.000	37.000	39.000	36.000	23.000	48.000	9.000	41.000	33.000	27.000	8
9 -	30.000	28.000	0.000	31.000	40.000	26.000	36.000	29.000	25.000	39.000	16.000	32.000	36.000	40.000	26.000	32.000	28.000	45.000	26.000	32.000	9
10	22.000	19.000	0.000	25.000	39.000	25.000	29.000	30.000	37.000	40.000	42.000	34.000	30.000	38.000	32.000	33.000	31.000	47.000	31.000	16.000	10
11 -	28.000	28.000	0.000	21.000	36.000	36.000	38.000	15.000	41.000	32.000	4.000	43.000	35.000	37.000	20.000	40.000	26.000	58.000	32.000	25.000	11
12	19.000	32.000	0.000	37.000	38.000	19.000	37.000	25.000	34.000	38.000	8.000	37.000	47.000	24.000	35.000	35.000	34.000	46.000	21.000	25.000	12
13	22.000	25.000	0.000	33.000	28.000	22.000	39.000	26.000	34.000	40.000	3.000	30.000	44.000	49.000	24.000	40.000	16.000	46.000	42.000	31.000	13
14	36.000	31.000	0.000	29.000	40.000	24.000	32.000	25.000	37.000	40.000	4.000	25.000	42.000	26.000	44.000	33.000	18.000	49.000	29.000	29.000	14
15	30.000	34.000	0.000	39.000	33.000	21.000	36.000	21.000	39.000	36.000	12.000	32.000	32.000	49.000	23.000	53.000	6.000	45.000	27.000	31.000	15
16 -	20.000	33.000	0.000	34.000	34.000	18.000	25.000	24.000	31.000	27.000	16.000	37.000	28.000	40.000	23.000	31.000	41.000	35.000	16.000	33.000	16
17 -	18.000	37.000	0.000	28.000	25.000	28.000	29.000	25.000	34.000	32.000	33.000	30.000	35.000	37.000	21.000	28.000	18.000	60.000	25.000	21.000	17
18 -	25.000	19.000	0.000	17.000	22.000	24.000	32.000	14.000	29.000	34.000	18.000	20.000	26.000	21.000	19.000	23.000	30.000	28.000	36.000	28.000	18
19	16.000	13.000	0.000	21.000	28.000	13.000	16.000	20.000	19.000	27.000	2.000	15.000	24.000	29.000	19.000	25.000	13.000	33.000	18.000	26.000	19
	8	2	13	4	6	18	11	Ó	10	3	7	16	17	19	14	15	i	5	9	12	•

Figure 21. UMAP (20 components, euclidean) Contingency Matrix

Homogeneity	0.031
Completeness	0.023
V-measure	0.027
Adjusted Rand Index	0.005
Adjusted Mutual Information Score	0.021

Table 5. Scores for UMAP (20 components, euclidean)

11.0.3 200 Components Euclidean



- 40

30

20

Figure 22. UMAP (200 components, euclidean) Contingency Matrix

Homogeneity	0.033
Completeness	0.024
V-measure	0.026
Adjusted Rand Index	0.004
Adjusted Mutual Information Score	0.022

Table 6. Scores for UMAP (200 components, euclidean)

11.0.4 5 Components Cosine

0 -		8.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	9.000	2.000	159.000	4.000	5.000	12.000	0.000	0
1 -	1.000	145.000		101.000	3.000	39.000	10.000	6.000	6.000	0.000	2.000	2.000	0.000	3.000	11.000	0.000	1.000	0.000	4.000	5.000	1
2 -	0.000	36.000	277.000	127.000	42.000	57.000	10.000	1.000	8.000	0.000	18.000	1.000	0.000	1.000	1.000	2.000	5.000	1.000	1.000	3.000	2
3 -	1.000	33.000	38.000	138.000	318.000	8.000	30.000	4.000	0.000	0.000	1.000	3.000	0.000	2.000	8.000	0.000	3.000	1.000	0.000	2.000	3
4 -	4.000	21.000	20.000	107.000	348.000	10.000	30.000	12.000	1.000	5.000	2.000	3.000	0.000	2.000	2.000	2.000	2.000	1.000	4.000	2.000	4
5 -	0.000	47.000	70.000	22.000	5.000	423.000	3.000	4.000	0.000	5.000	0.000	6.000	1.000	0.000	6.000	0.000	0.000	0.000	1.000	0.000	5
6 -	1.000	78.000	30.000	43.000	77.000	3.000	263.000	27.000	12.000	13.000	11.000	1.000	0.000	3.000	6.000	4.000	4.000	0.000	7.000	2.000	6
7 -	0.000	38.000	3.000	0.000	2.000	7.000	8.000	422.000	60.000	1.000	3.000	2.000	0.000	2.000	2.000	2.000	4.000	2.000	34.000	2.000	7
8 -	1.000	19.000	5.000	2.000	5.000	1.000	9.000	50.000	473.000	1.000	3.000	1.000	0.000	3.000	3.000	5.000	7.000	0.000	9.000	1.000	8
9 -	1.000	31.000	3.000	0.000	1.000	1000	9.000	4.000	3.000	496.000	30.000	0.000	1000	3.000	2.000	4.000	2.000	0.000	4.000	2.000	9
10 -	2.000	20.000	0.000	0.000	1.000	1.000	5.000	0.000	0.000	24.000	528.000	0.000	0.000	0.000	3.000	1.000	3.000	3.000	7.000	2.000	10
11 -	1.000	24.000	11.000	1.000	2.000	1.000	4.000	0.000	2.000	0.000	2.000	477.000	1.000	3.000	2.000	3.000	48.000	4.000	9.000	0.000	11
12 -	0.000	95.000	22.000	29.000	65.000	4.000	234.000	93.000	4.000	7.000	5.000	3.000	2.000	4.000	14.000	5.000	4.000	0.000	1.000	0.000	12
13 -	2.000	44.000	3.000	1.000	0.000	6.000	6.000	11.000	7.000	3.000	1.000	1.000	0.000	421.000	18.000	46.000	7.000	0.000	16.000	1.000	13
14 -	0.000	47.000	5.000	3.000	0.000	0.000	7.000	13.000	1.000	2.000	2.000	1.000	1.000	5.000	478.000	7.000	14.000	1.000	6.000	0.000	14
15 -	6.000	22.000	3.000	0.000	1.000	1.000	0.000	2.000	1.000	2.000	1.000	1.000	3.000	1.000	0.000	535.000	5.000	3.000	12.000	0.000	15
16 -	2.000	10.000	2.000	0.000	1.000	0.000	5.000	3.000	7.000	1.000	3.000	6.000	0.000	0.000	12.000	12.000	449.000	3.000	30.000	0.000	16
17 -	3.000	16.000	1.000	1.000	0.000	0.000	4.000	3.000	6.000	0.000	2.000	4.000	139.000	6.000	0.000	12.000	11.000	336.000	20.000	0.000	17
18 -	7.000	12.000	2.000	0.000	0.000	0.000	3.000	1.000	11.000	2.000	0.000	4.000	1.000	2.000	6.000	11.000	173.000	9.000	220.000	1.000	18
19 -	61.000	10.000	1.000	0.000	1.000	0.000	2.000	2.000	3.000	1.000	1.000	0.000	3.000	8.000	1.000	186.000	37.000	2.000	57.000	1.000	19
	i	14	5	17	0	11	16	2	15	19	3	4	10	8	7	12	9	13	18	6	

300

200

Figure 23. UMAP (5 components, cosine) Contingency Matrix

Homogeneity	0.567
Completeness	0.585
V-measure	0.576
Adjusted Rand Index	0.436
Adjusted Mutual Information Score	0.573

Table 7. Scores for UMAP (5 components, cosine)

11.0.5 20 Components Cosine

0 -	269.000	8.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	2.000	0.000	4.000	8.000	1.000	166.000	2.000	2.000	14.000	0.000	0
1 -	0.000	153.000	275.000	88.000	5.000	22.000	10.000	2.000	0.000	1.000	2.000	0.000	13.000	2.000	8.000	0.000	0.000	1.000	2.000	0.000	1
2 -	0.000	34.000	331.000	83.000	50.000	45.000	9.000	0.000	4.000	0.000	17.000	1.000	6.000	1.000	3.000	1.000	5.000	1.000	0.000	0.000	2
3 -	1.000	27.000	43.000	126.000	327.000	8.000	32.000	3.000	1.000	0.000	1.000	3.000	3.000	2.000	10.000	1.000	1.000	1.000	0.000	0.000	3
4 -	4.000	19.000	40.000	69.000	373.000	11.000	34.000	5.000	3.000	3.000	0.000	2.000	7.000	1.000	2.000	1.000	1.000	1.000	2.000	0.000	4
5 -	0.000	51.000	74.000	12.000	7.000	426.000	3.000	0.000	0.000	4.000	0.000	4.000	5.000	1.000	5.000	1.000	0.000	0.000	0.000	0.000	5
6 -	1.000	64.000	34.000	32.000	81.000	5.000	274.000	27.000	3.000	13.000	7.000	1000	23.000	2.000	4.000	3.000	4.000	0.000	7.000	0.000	6
7 -	1.000	41.000	2.000	0.000	2.000	5.000	13.000	386.000	43.000	2.000	4.000	2.000	78.000	3.000	2.000	1.000	3.000	3.000	3.000	0.000	7
8 -	1.000	19.000	2.000	2.000	3.000	1.000	10.000	34.000	459.000	3.000	3.000	3.000	44.000	1.000	4.000	3.000	3.000	0.000	3.000	0.000	8
9 -	0.000	23.000	5.000	0.000	2.000	1.000	10.000	2.000	0.000	493.000	26.000	1.000	16.000	1.000	8.000	2.000	3.000	1.000	3.000	0.000	9
10 -	1.000	16.000	0.000	2.000	0.000	0.000	6.000	1.000	2.000	27.000	525.000	0.000	6.000	2.000	4.000	1.000	2.000	3.000	2.000	0.000	10
11 -	1.000	26.000	13.000	1.000	1.000	2.000	4.000	0.000	2.000	0.000	2.000	476.000	10.000	4.000	2.000	1.000	24.000	4.000	22.000	0.000	11
12	0.000	111.000	24.000	20.000	64.000	4.000	237.000	70.000	4.000	4.000	2.000	0.000	21.000	2.000	17.000	7.000	1.000	0.000	3.000	0.000	12
13 -	1.000	45.000	6.000	1.000	1.000	3.000	6.000	6.000	6.000	2.000	0.000	1.000	18.000	418.000	23.000	41.000	6.000	1.000	9.000	0.000	13
14 -	0.000	40.000	6.000	0.000	4.000	0.000	8.000	10.000	1.000	2.000	3.000	0.000	9.000	11.000	475.000	9.000	7.000	1.000	7.000	0.000	14
15	5.000	22.000	1.000	0.000	1.000	0.000	1.000	1.000	0.000	3.000	0.000	1.000	7.000	1.000	0.000	538.000	3.000	1.000	11.000	3.000	15
16 -	1.000	6.000	2.000	0.000	0.000	0.000	5.000	0.000	1.000	1.000	2.000	9.000	24.000	4.000	5.000	8.000	436.000	2.000	40.000	0.000	16
17 -	2.000	11.000	1.000	0.000	0.000	0.000	2.000	2.000	2.000	1.000	0.000	2.000	13.000	7.000	0.000	14.000	3.000	320.000	45.000	139.000	17
18 -	7.000	8.000	2.000	0.000	0.000	0.000	5.000	1.000	3.000	3.000	0.000	5.000	43.000	2.000	9.000	8.000	17.000	5.000	347.000	0.000	18
19 -	51.000	5.000	1.000	1.000	0.000	1.000	2.000	0.000	3.000	9.000	1.000	0.000	8.000	17.000	3.000	193.000	31.000	1.000	49.000	1.000	19
	12	14	9	19	6	Ó	4	17	5	13	2	11	7	15	10	3	i	8	16	18	ı

400

300

Figure 24. UMAP (20 components, cosine) Contingency Matrix

Homogeneity	0.580
Completeness	0.591
V-measure	0.585
Adjusted Rand Index	0.448
Adjusted Mutual Information Score	0.583

Table 8. Scores for UMAP (20 components, cosine)

11.0.6 200 Components Cosine

0 -	301.000	0.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000	0.000	1.000	0.000	9.000	7.000	1.000	134.000	2.000	5.000	16.000	1.000	0
1 -	2.000	85.000	293.000	3.000	8.000	25.000	5.000	4.000	5.000	0.000	1.000	1.000	138.000	3.000	7.000	0.000	0.000	0.000	4.000	0.000	1
2 -	0.000	79.000	327.000	10.000	31.000	64.000	11.000	0.000	5.000	0.000	18.000	1.000	39.000	2.000	1.000	2.000	0.000	1.000	0.000	0.000	2
3 -	1.000	120.000	52.000	225.000	102.000	10.000	25.000	3.000	1.000	0.000	3.000	2.000	33.000	3.000	8.000	0.000	0.000	1.000	1.000	0.000	3
4 -	5.000	66.000	31.000	98.000	283.000	16.000	24.000	11.000	3.000	6.000	3.000	1.000	22.000	1.000	2.000	1.000	0.000	1.000	4.000	0.000	4
5 -	0.000	10.000	79.000	4.000	2.000	426.000	4.000	4.000	2.000	4.000	0.000	0.000	50.000	1.000	6.000	0.000	0.000	0.000	1.000	0.000	5
6 -	2.000	33.000	42.000	48.000	43.000	6.000	256.000	27.000	12.000	10.000	11.000	2.000	71.000	7.000	5.000	4.000	1.000	0.000	5.000	0.000	6
7 -	2.000	2.000	4.000	4.000	0.000	6.000	15.000	427.000	48.000	4.000	3.000	3.000	35.000	2.000	1.000	3.000	5.000	1.000	29.000	0.000	7
8 -	1.000	1.000	3.000	2.000	0.000	1.000	13.000	37.000	482.000	4.000	5.000	1.000	22.000	2.000	4.000	4.000	8.000	0.000	8.000	0.000	8
9 -	2.000	1.000	4.000	0.000	0.000	2.000	9.000	2.000	3.000	491.000	29.000	0.000	30.000	5.000	5.000	7.000	1.000	0.000	6.000	0.000	9
10 -	2.000	1.000	0.000	1.000	2.000	0.000	7.000	1.000	0.000	21.000	537.000	0.000	16.000	1.000	2.000	1.000	1.000	3.000	4.000	0.000	10
11 -	2.000	1.000	13.000	0.000	1.000	1.000	5.000	0.000	4.000	0.000	3.000	486.000	24.000	5.000	3.000	2.000	24.000	3.000	18.000	0.000	11
12 -	1.000	27.000	22.000	14.000	67.000	4.000	234.000	91.000	4.000	6.000	3.000	0.000	87.000	9.000	14.000	5.000	1.000	1.000	1.000	0.000	12
13 -	34.000	1.000	5.000	0.000	2.000	5.000	6.000	5.000	7.000	0.000	1.000	1.000	55.000	424.000	17.000	8.000	4.000	0.000	19.000	0.000	13
14 -	7.000	0.000	7.000	1.000	0.000	1.000	8.000	11.000	4.000	2.000	1.000	0.000	46.000	10.000	479.000	3.000	6.000	1.000	5.000	1.000	14
15 -	10.000	0.000	3.000	1.000	0.000	1.000	0.000	3.000	1.000	2.000	0.000	1.000	19.000	1.000	0.000	535.000	3.000	3.000	14.000	2.000	15
16 -	4.000	0.000	2.000	0.000	0.000	0.000	5.000	2.000	15.000	2.000	2.000	10.000	8.000	5.000	5.000	9.000	429.000	5.000	43.000	0.000	16
17 -	6.000	0.000	3.000	0.000	0.000	0.000	2.000	2.000	5.000	1.000	1.000	2.000	15.000	11.000	1.000	10.000	1.000	317.000	48.000	139.000	17
18 -	10.000	0.000	4.000	0.000	0.000	0.000	4.000	2.000	17.000	3.000	1.000	4.000	15.000	4.000	6.000	10.000	14.000	7.000	364.000	0.000	18
19 -	81.000	0.000	0.000	1.000	2.000	3.000	1.000	3.000	5.000	5.000	1.000	0.000	9.000	17.000	0.000	190.000	25.000	2.000	30.000	2.000	19
	19	Ó	14	8	17	11	7	18	i	6	16	9	4	13	12	3	5	10	2	15	1

Figure 25. UMAP (200 components, cosine) Contingency Matrix

Homogeneity	0.578
Completeness	0.587
V-measure	0.582
Adjusted Rand Index	0.454
Adjusted Mutual Information Score	0.580

Table 9. Scores for UMAP (200 components, cosine)

For both COS and EUC models, we got that 20 components were the best number of components. However, we can see that using the cosine method is far superior compared to the euclidean distance metric. The reason for this is that cosine is not affected by the magnitude of vectors, rather it uses the angles between sample points. So the documents can have varying lengths which lead to varying distances, but this will be factored in when using the cosine, which makes it better than euclidean. Equally important, in higher dimensions, we have mentioned that data become sparse, so in euclidean measure, the distance can converge when a document becomes too large between sample points. Furthermore, when features start to become equidistant, UMAP will not be able to find significant trends with respect to the distance of the data.

13. Question 13

Based on all 5 metrics, the best clustering method is UMAP reduced data with the cosine distance metric and using twenty principal components. SVD reduction also performs fairly well, but its scores are noticeably worse than the best method of UMAP. NMF performs slightly worse than SVD reduction, and UMAP with euclidean distance performs worst of all by a large margin.

14. Question 14

Ward linkage criteria performed better than single linkage. We may guess that this is because ward is a variance-minimizing approach, which minimizes the sum of squared distances within all clusters. Whereas, the single linkage minimizes the distance between the closest observation of pairs of clusters. In this way, Single linkage is not robust for noisy data, so did not perform the best.

14.1. WARD

Homogeneity	0.563
Completeness	0.583
V-measure	0.573
Adjusted Rand Index	0.422
Adjusted Mutual Information Score	0.570

Table 10. Scores for WARD

14.2. SINGLE

Homogeneity	0.108
Completeness	0.642
V-measure	0.185
Adjusted Rand Index	0.020
Adjusted Mutual Information Score	0.179

Table 11. Scores for SINGLE

Homogeneity	0.453
Completeness	0.491
V-measure	0.471
Adjusted Rand Index	0.113
Adjusted Mutual Information Score	0.459

Table 12. Scores for HDBSCAN (min_cluster_size=20)

Homogeneity	0.382
Completeness	0.619
V-measure	0.473
Adjusted Rand Index	0.172
Adjusted Mutual Information Score	0.472

Table 13. Scores for HDBSCAN (min_cluster_size=100)

Homogeneity	0.351
Completeness	0.593
V-measure	0.441
Adjusted Rand Index	0.151
Adjusted Mutual Information Score	0.439

Table 14. Scores for HDBSCAN (min_cluster_size=200)

										D	BSCA	N											
0 -	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0	
1 -	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	410.000	0.000	0.000	0.000	0.000	0.000	1.000	68.000	0.000	1.000	0.000	0.000	1	
2 -	0.000	0.000	0.000	0.000	0.000	0.000	7.000	0.000	0.000	575.000	0.000	0.000	0.000	0.000	0.000	2.000	0.000	0.000	0.000	0.000	0.000	2	
3 -	0.000	0.000	0.000	0.000	8.000	0.000	9.000	0.000	0.000	572.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000	3	- 500
4 -	0.000	0.000	0.000	0.000	148.000	0.000	2.000	0.000	1.000	433.000	1.000	0.000	1.000	0.000	1.000	3.000	0.000	0.000	0.000	0.000	0.000	4	
5 -	0.000	0.000	0.000	0.000	65.000	0.000	2.000	0.000	0.000	505.000	2.000	1.000	2.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	5	
6 -	0.000	0.000	0.000	0.000	0.000	0.000	301.000	0.000	0.000	291.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	6	- 400
7 -	0.000	0.000	0.000	0.000	28.000	0.000	0.000	0.000	17.000	536.000	1.000	0.000	0.000	0.000	0.000	2.000	1.000	0.000	0.000	0.000	0.000	7	
8 -	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	224.000	361.000	2.000	1.000	1.000	0.000	1.000	0.000	2.000	0.000	1.000	0.000	0.000	8	
9 -	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	8.000	584.000	0.000	1.000	1.000	0.000	0.000	0.000	3.000	0.000	0.000	0.000	0.000	9	
10 -	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	186.000	395.000	10.000	0.000	0.000	0.000	3.000	2.000	0.000	0.000	0.000	0.000	10	- 300
11 -	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	210.000	4.000	384.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000	0.000	0.000	11	
12 -	0.000	0.000	0.000	0.000	0.000	0.000	2.000	0.000	0.000	165.000	1.000	0.000	423.000	0.000	2.000	1.000	1.000	0.000	0.000	0.000	0.000	12	
13 -	0.000	0.000	0.000	0.000	6.000	0.000	1.000	0.000	8.000	566.000	3.000	1.000	1.000	0.000	0.000	4.000	1.000	0.000	0.000	0.000	0.000	13	- 200
14 -	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	315.000	0.000	0.000	1.000	0.000	275.000	3.000	0.000	0.000	0.000	0.000	0.000	14	
15 -	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	226.000	0.000	0.000	0.000	0.000	0.000	365.000	1.000	0.000	0.000	0.000	0.000	15	
16 -	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	215.000	0.000	0.000	1.000	0.000	0.000	0.000	381.000	0.000	1.000	0.000	0.000	16	
17 -	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	536.000	1.000	0.000	2.000	0.000	0.000	3.000	4.000	0.000	0.000	0.000	0.000	17	- 100
18 -	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	270.000	0.000	0.000	3.000	0.000	3.000	0.000	7.000	0.000	281.000	0.000	0.000	18	
19 -	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	455.000	1.000	0.000	2.000	0.000	0.000	1.000	5.000	0.000	0.000	0.000	0.000	19	
20 -	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	266.000	1.000	0.000	0.000	0.000	0.000	2.000	107.000	0.000	1.000	0.000	0.000	20	
	11	16	13	18	10	12	7	14	8	0	2	6	3	15	9	4	i	17	5	19	20	1	٥ -

Figure 26. 20-Cat Contingency Matrix (HDBSCAN)

Print out the unique labels, we see there is [0, 10] meaning there are 11 clusters. Looking at the contingency matrix, we have 11 major clusters. We can see this in the deep blue within the contingency matrix, which can represent our clusters. Noisy samples are given the label -1, which means it is an outlier or noise that was clustered by the algorithm. We can see there are a lot of labelings that did not have many clusters associated with them. This is due to when running the code with the sample size. This is a hyperparameter than can affect our clustering. What does not help is the noise when going to higher categories.

After running all possibilities of dimensionality reduction techniques and clustering methods provided in the table, we noticed that kmeans with 50 clusters using the top 20 features of NMF reduction was actually the best model. This is reinforced by the evidence from previous questions on this project. However, we knew that UMAP would actually perform the best when given the right conditions. Hence, take a look at our next question where we found a creative way to further enhance our performance.

	Dim Red	Dim Red Param	Clustering	Clustering Param	Homogeneity	Completeness	V-measure	Adjusted Rand-Index	Adjusted Mutual Information
89	UMAP	200	Kmeans	20	0.580000	0.604759	0.592121	0.457428	5.898660e-01
77	UMAP	5	Kmeans	20	0.565415	0.585196	0.575135	0.444008	5.727952e-01
91	UMAP	200	Agglomerative	20	0.567220	0.581735	0.574386	0.438778	5.720547e-01
83	UMAP	20	Kmeans	20	0.562184	0.585492	0.573601	0.421849	5.712455e-01
79	UMAP	5	Agglomerative	20	0.549878	0.573298	0.561343	0.412517	5.589183e-01

Figure 27. Top few dimensionality reduction and clustering parameters ranked

18. Question 18

Following the evidence in Question 17 and the previous UMAP question on this project, we knew that a UMAP with cosine distance metric would actually perform very well. Since we ran Question 17 with the euclidean method, UMAP did not get a chance to prove its power. Therefore, we used cosine as a distance metric in this extra credit, and to our expectation, it outperformed all of our other models. The best model combination ended up being UMAP dimensionality reduction with 200 dimensions with kmeans clustering of 20 groups.

19. Question 19

We take advantage of the transfer learning concept. We use a previously loaded neural network (VGG) and modify it to fit our scenario changing the dataset's input feature size and expected output size. VGG is trained on a task that has a larger scope than our classification problem. The idea is we know more initially and we apply it to learn something new. We take the sensitivity of VGG and apply it to different tasks.

VGG will take in an input dataset and extract important features and it will break it down to lower-level features, such as edges. The more varying different data sets fed into the neural network the lower level features it will extract, and it begins to train its weights within the neural network. When a VGG network is done learning, it can keep its trained weights. So now this network has information on a smaller data set, and it can be used as a backbone for another that will have some discriminative powers to extract lower-level features from the custom dataset.

20. Question 20

The helper code first imports the flower images from tensor flow. A FeatureExtractor object is initialized.

FeatureExtractor().eval() is to set dropout and batch normalization layers to evaluation mode before running inference. These layers prevent overfitting and stabilize training.

Then, the dataset is resized, cropped to its center, converted to a tensor, and normalized. A data loader is created from the dataset. The data loader wraps an iterable around the Dataset to enable easy access to the samples to allow us to pass samples in mini-batches (set size to 64), shuffle (benefit for training), and multiprocess.

Afterward, the features are extracted. Two arrays are created: one feature 2d and one target array. For each minibatch in the data loader, the following feature extraction occurs. For the target array, the original targets are just concatenated. For the feature array, the feature extractor is run.

From VGG-16, the feature and pooling layers are taken. The feature layer extracted from VGG-16 is 16 convolutional layers. These layers and trained weights can detect generic features from our pictures, much more sensitive than our flower classification requires (transfer learning principle). Essentially the convolution layer is a kernel filter that is a layer on top of the image. This filter is then used to extract features, where it outputs a feature map which is a reduced and summarized version of the input data.

The pooling layer is the process of merging for the purpose of reducing the size of the data. This process removes some of the noise in the data and extracts significant ones. This reduces overfitting and speeds up computation. Pooling methods

depend on the user, which is meant to reduce the spatial dimensions of the feature map while trying to retain as much information as possible. The convolution layer occurs 16 times, and the pooling layer occurs every 2 - 3 a convolution layer occurs.

Then, the image is flattened, which means it is converted into a 1D vector. This vector is fed into a fully connected network, meaning the output flattened feature is connected to every neuron, where the connection between features and neurons has weights (the weights can be trained). Lastly, we extract the first part of the full-connected network to get out classification.

21. Question 21

There are 224x224 pixels in the original image. There are 4096 features extracted. The dimension for each feature vector for an image sample is (1x4096)

22. Question 22

About 71 percent of values in the feature map are close to zero, so the features are sparse.

23. Question 23

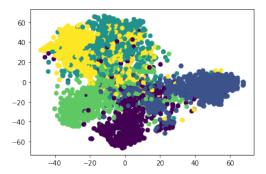


Figure 28. Plot of features projected onto two dimensions using t-SNE

There are four distinct clusters in the t-SNE projection and the fifth class's projections are intermingled with the other four classes. This makes sense since there are five ground truth classes in the source dataset. This likely means this class is more difficult to classify than the other four, and we would expect this to manifest itself in quantitative measures of classifier performance as well as the qualitative view we get from the t-SNE plot.

24. Question 24

Our best result: UMAP(cosine, n_components=50) and agglomerative clustering (n_cluster=5, linkage='ward') with adjusted rand index of 0.4998

Our conservative HDBSCAN grid was over min_cluster_size=[1, 25, 50, 100, 200] and min_samples=[5, 15, 30, 60, 100, 200, 500]. We found that the best results were min_cluster_size of 50 and min_samples of 60. Although, our resulting best parameters used agglomerative clustering.

25. Question 25

We trained the classifier on a random train test split with 80% of the data points as the training set and the remaining 20% as the validation/test set. Our classifier was able to achieve a final accuracy of 89.78% on the testing set. The classifier was trained for 100 epochs with negative log-likelihood loss and a learning rate of 10^{-3} . The performance suffers a little for UMAP and Autoencoder reduced representations (79% and 86% accuracy, respectively), but the classifier achieves similar accuracy with no dimensionality reduction or with SVD reduction. These results are a little surprising in the context of our clustering results, as this accuracy is better than the homogeneity of our clustering methods would have us believe.