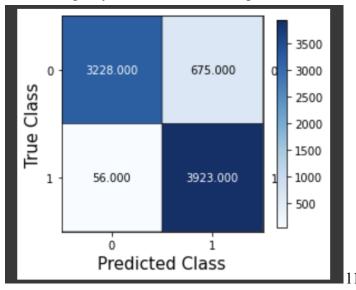
# Part 1

#### Question 1:

(7882, 23522)

## Question 2:

The contingency table of the clustering results is as following:



Contingency matrix doesn't have to be square-shaped. The number of predicted classes we got in the contingency table depends on the number of clusters we set in the k-means algorithm. Therefore, if we set the number of clusters to be smaller than the number of true classes, the number of True Class (rows) will be more than that of the Predicted Class (columns).

## Question 3:

```
Homogeneity: 0.582 ± 0.012

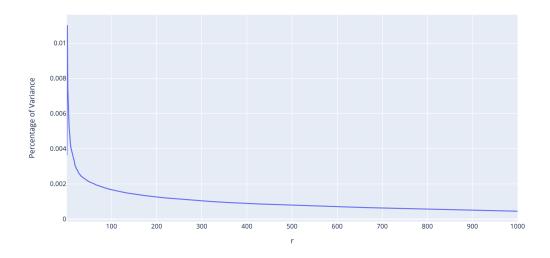
Completeness: 0.596 ± 0.011

V-measure: 0.589 ± 0.012

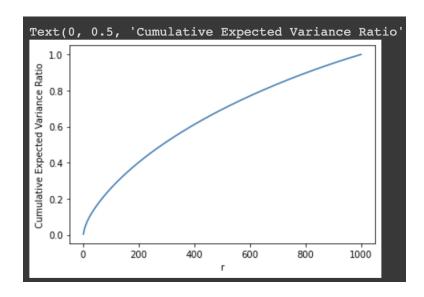
Adjusted Rand-Index: 0.644 ± 0.015

Adjusted mutual information score: 0.589 ± 0.012
```

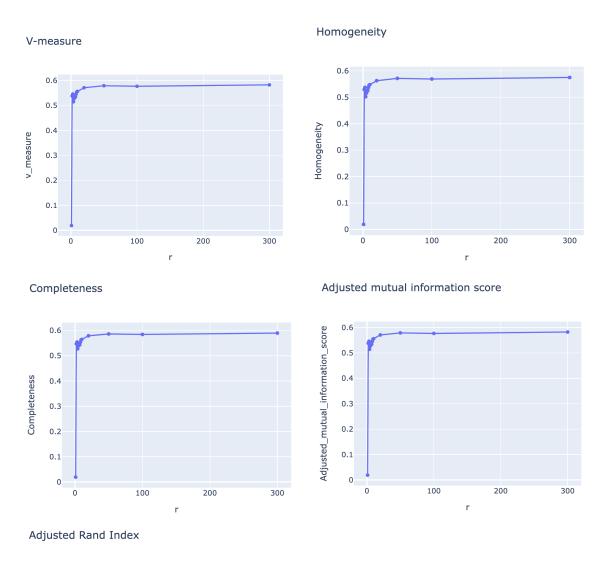
# Question 4: Explained Variance Ratio Per Components

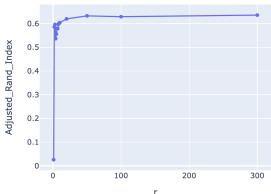


# Cumulative Expected Variance Ratio



# Question 5: SVD (the r with the best performance is 50)



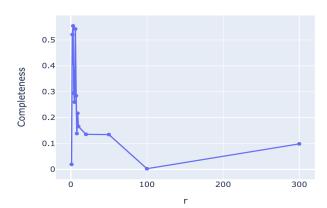


NMF (the r with the best performance is 3):

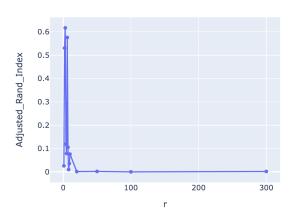
## Homogeneity

# 0.5 0.4 0.3 0.2 0.1 0 0 100 200 300

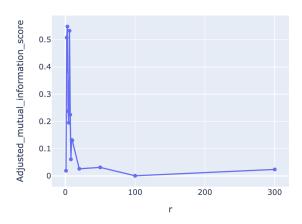
## Completeness



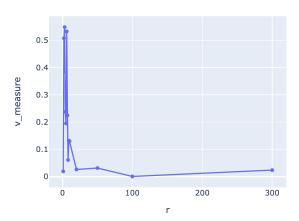
## Adjusted Rand Index



#### Adjusted mutual information score



#### V-measure



# Question 6:

In our example, the performance for SVD drastically improves as the number of components goes from 0 to 50. After this point, the model seems not improving as much as before. This might be due to the fact that more relatively irrelevant data come in as we increase the # of components, resulting in no further improvement in making predictions. For NMF, the average performance score reaches the peak at r = 3 and then decreases to nearly 0. This might be due to NMF being more noise sensitive compared to SVD.

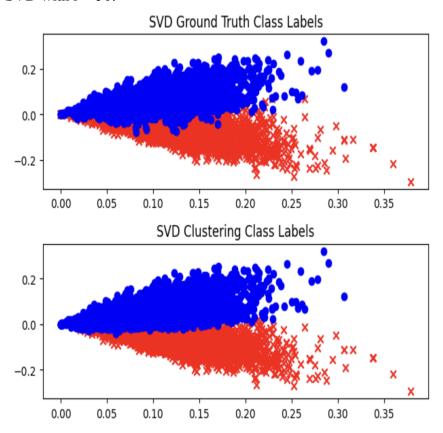
## Question 7:

No, the average in Question 3 is about 0.6, whereas here we only have a maximum average of 0.58 in Question 6.

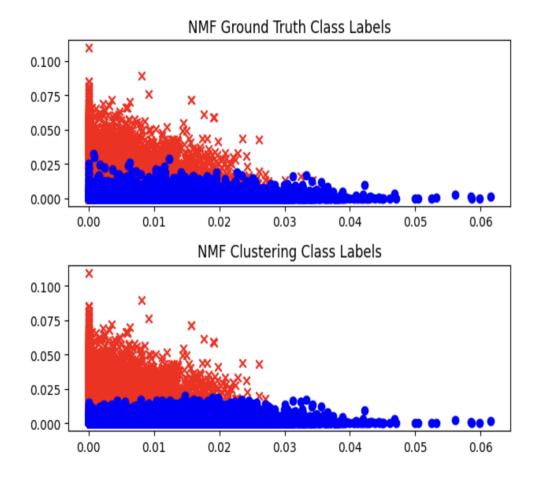
## Question 8:

Visualize the clustering results for:

#### SVD with r = 50:



#### NMF with r = 3:

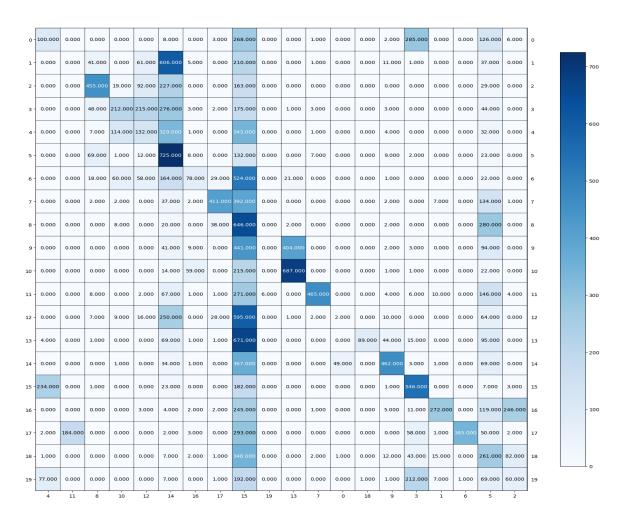


# Question 9:

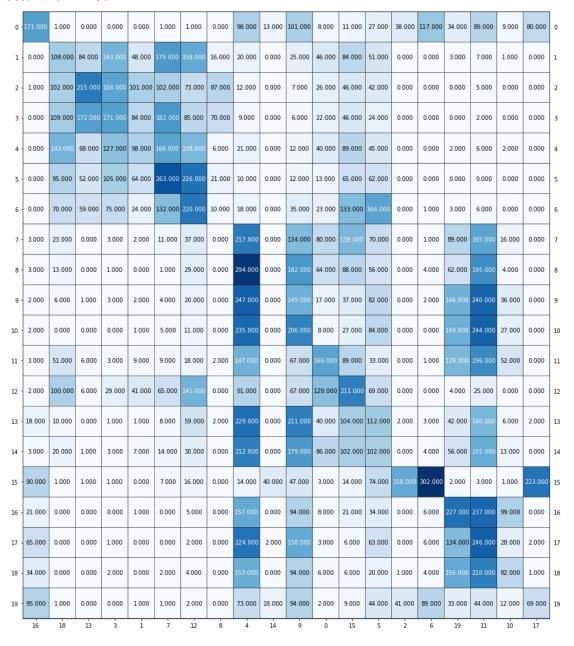
From these visualizations, we can see that the boundaries in the SVD/NMF Clustering Class Labels figures are more smooth and linear (some blue points are being classified as red points after the SVD/NMF clustering). The clustering data of the two classes are evenly distributed by the boundaries. Overall, we believe that the distribution is ideal for K-Means clustering, because 1. the boundaries are clear and well-defined, and 2. the clusters in SVD and NMF make sense, and 3. the data within the clusters are almost from the same class.

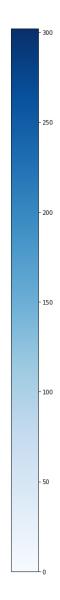
# Question 10:

For SVD, we set r = 1000, because we got better measure scores in Q3 (r = 1000) compared to the scores in Q5.



For NMF, we decided r to be 3, because the result of Question 8 indicates that the result is the best when r = 3.





# Question 11:

\*\*\*\*\*\*\* This is for components = 5 and metric = cosine \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Homogeneity score: 0.5787205426049561 Completeness score: 0.5887460516747586 V-measure score: 0.5836902505839211 Adjusted Rand score: 0.4595680059173617

Adjusted mutual information score: 0.5823341943107677

0 -	604.000	3.000	0.000	0.000	1.000	1.000	1.000	0.000	2.000	0.000	2.000	1.000	14.000	8.000	6.000	125.000	8.000	23.000	0.000	0.000	0
1 -	1.000		84.000	1.000	90.000	28.000	13.000	5.000	3.000	2.000	3.000	2.000	90.000	3.000	11.000	0.000	3.000	0.000	1.000	3.000	1
2 -	1.000	80.000	573.000	28.000	121.000	70.000	9.000	14.000	3.000	2.000	3.000	11.000	48.000	4.000	9.000	2.000	6.000	1.000	0.000	0.000	2
3 -	4.000	13.000	98.000	316.000	432.000	11.000	46.000	4.000	2.000	3.000	3.000	3.000	34.000	2.000	8.000	0.000	1.000	0.000	0.000	2.000	3
4-	5.000	9.000	91.000	164.000	514.000	24.000	65.000	13.000	4.000	4.000	1.000	3.000	47.000	1.000	5.000	2.000	7.000	2.000	0.000	2.000	4
5 -	1.000	143.000	107.000	2.000	17.000	624.000	9.000	5.000	1.000	0.000	2.000	2.000	58.000	0.000	9.000	3.000	3.000	0.000	0.000	2.000	5
6 -	5.000	14.000	61.000	98.000	126.000	4.000	445.000	48.000	9.000	22.000	9.000	1.000	102.000	4.000	10.000	8.000	6.000	0.000	0.000	3.000	6
7 -	6.000	5.000	5.000	3.000	3.000	4.000	24.000	760.000	61.000	5.000	8.000	2.000	56.000	15.000	14.000	5.000	13.000	1.000	0.000	0.000	7
8 -	3.000	5.000	3.000	4.000	4.000	0.000	37.000	57.000	792.000	7.000	10.000	1.000	42.000	2.000	3.000	9.000	7.000	9.000	0.000	1.000	8
9 -	5.000	7.000	4.000	2.000	1.000	0.000	11.000	20.000	8.000	804.000	64.000	0.000	42.000	2.000	4.000	7.000	12.000	0.000	0.000	1.000	9
10 -	2.000	4.000	1.000	0.000	1.000	0.000	12.000	7.000	2.000	23.000	919.000	1.000	16.000	2.000	3.000	2.000	3.000	0.000	0.000	1.000	10
11 -	7.000	17.000	16.000	1.000	4.000	1.000	6.000	6.000	1.000	0.000	1.000	823.000	44.000	5.000	1.000	1.000	57.000	0.000	0.000	0.000	11
12 -	13.000	36.000	76.000	26.000	129.000	4.000	436.000	46.000	11.000	6.000	6.000	7.000	114.000	37.000	25.000	8.000	3.000	1.000	0.000	0.000	12
13 -	39.000	12.000	14.000	0.000	6.000	3.000	15.000	5.000	7.000	6.000	4.000	3.000	79.000	706.000	58.000	18.000	12.000	2.000	0.000	1.000	13
14 -	7.000	27.000	3.000	1.000	2.000	4.000	10.000	7.000	5.000	7.000	3.000	2.000	43.000	18.000	817.000	3.000	27.000	1.000	0.000	0.000	14
15 -	72.000	3.000	1.000	1.000	1.000	0.000	4.000	2.000	2.000	3.000	1.000	2.000	37.000	14.000	2.000	818.000	13.000	19.000	2.000	0.000	15
16 -	12.000	2.000	5.000	0.000	3.000	0.000	10.000	11.000	7.000	8.000	4.000	6.000	21.000	6.000	8.000	7.000	797.000	3.000	0.000	0.000	16
17 -	19.000	2.000	1.000	0.000	0.000	0.000	4.000	2.000	6.000	5.000	9.000	2.000	37.000	3.000	1.000	13.000	38.000	596.000	202.000	0.000	17
18 -	17.000	2.000	6.000	1.000	0.000	1.000	4.000	13.000	9.000	4.000	5.000	4.000	19.000	60.000	8.000	186.000	412.000	20.000	3.000	1.000	18
19 -	174.000	1.000	3.000	0.000	1.000	1.000	1.000	1.000	3.000	20.000	2.000	0.000	14.000	10.000	4.000	296.000	90.000	3.000	3.000	1.000	19
	14	3	13	11	17	12	19	16	6	15	4	8	í	7	9	2	5	0	18	10	'

200

Homogeneity score: 0.015028207643473433 Completeness score: 0.016265879653337943 V-measure score: 0.015622568865206654 Adjusted Rand score: 0.0025790051084612175

Adjusted mutual information score: 0.012336506360402432

0 -	1.000	37.000	0.000	0.000	56.000	61.000	43.000	55.000	55.000	41.000	79.000	36.000	75.000	68.000	30.000	53.000	67.000	42.000	0.000	0.000	0
1-	0.000	75.000	0.000	0.000	62.000	91.000	64.000	55.000	76.000	67.000	63.000	35.000	118.000	42.000	13.000	57.000	70.000	75.000	10.000	0.000	1
2 -	0.000	53.000	10.000	16.000	61.000		54.000	58.000	46.000	60.000		58.000	93.000	39.000	61.000	59.000	62.000	62.000	32.000	0.000	2
3 -	0.000	62.000	0.000	43.000	68.000	75.000	62.000	42.000		67.000	56.000	42.000	117.000	43.000	32.000	70.000	67.000	56.000	0.000	0.000	3
4 -	2.000	61.000	0.000	0.000	87.000	93.000	58.000	37.000	83.000	57.000	47.000	68.000	105.000	27.000	21.000	76.000	66.000	75.000	0.000	0.000	4
5 -	0.000	62.000	0.000	10.000	56.000	104.000	45.000	49.000	89.000	75.000	55.000	30.000	129.000	35.000	59.000	63.000	75.000	52.000	0.000	0.000	5
6 -	2.000	84.000	0.000	0.000	83.000	107.000	85.000	25.000	92.000	84.000	24.000	8.000	139.000	20.000	10.000	89.000	64.000	31.000	21.000	7.000	6
7 -	0.000	61.000	0.000	10.000	59.000	73.000	54.000	74.000	56.000	65.000	73.000	62.000	125.000	44.000	29.000	72.000	67.000	66.000	0.000	0.000	7
8 -	1.000	62.000	0.000	7.000	58.000	86.000	58.000	63.000	74.000	50.000	69.000	43.000	111.000	71.000	34.000	61.000	71.000	63.000	14.000	0.000	8
9 -	0.000	43.000	0.000	25.000	79.000	82.000	36.000	48.000	72.000	75.000	56.000	62.000	99.000	53.000	34.000	60.000	59.000	82.000	29.000	0.000	9
10 -	1.000	58.000	0.000	30.000	79.000	61.000	44.000	69.000	56.000	51.000	89.000	46.000	84.000	80.000	42.000	45.000	69.000	79.000	16.000	0.000	10
11 -	0.000	50.000	0.000	0.000	66.000	61.000	52.000	45.000	64.000	52.000	79.000	102.000	93.000	64.000	27.000	60.000	85.000	68.000	23.000	0.000	11
12 -	0.000	52.000	0.000	18.000	66.000	76.000	53.000	77.000	67.000	66.000	60.000	53.000	131.000	58.000	17.000	59.000	69.000	55.000	7.000	0.000	12
13 -	0.000	39.000	0.000	13.000	82.000	85.000	68.000	54.000	63.000	69.000	43.000	27.000	140.000	101.000	24.000	62.000	61.000	59.000	0.000	0.000	13
14 -	0.000	65.000	0.000	6.000	58.000	64.000	64.000	37.000	71.000	55.000	74.000	63.000	103.000	69.000	60.000	58.000	66.000	74.000	0.000	0.000	14
15 -	0.000	51.000	0.000	0.000	80.000	91.000	61.000	38.000	72.000	61.000	75.000	46.000	117.000	55.000	24.000	76.000	65.000	85.000	0.000	0.000	15
16 -	0.000	42.000	0.000	21.000	59.000	74.000	57.000	58.000	49.000	41.000	74.000	59.000	86.000	64.000	31.000	48.000	76.000	65.000	6.000	0.000	16
17 -	0.000	36.000	0.000	5.000	66.000	50.000	39.000	61.000	42.000	31.000	90.000	98.000	69.000	79.000	63.000	43.000	50.000	96.000	22.000	0.000	17
18 -	0.000	36.000	0.000	0.000	44.000	40.000	58.000	42.000	55.000	35.000	52.000	31.000	81.000	66.000	26.000	53.000	41.000	48.000	67.000	0.000	18
19 -	0.000	46.000	0.000	0.000	39.000	50.000	33.000	50.000	44.000	28.000	42.000	34.000	59.000	50.000	36.000	36.000	42.000	39.000	0.000	0.000	19
l	15	14	17	2	10	8	12	9	7	Ó	1	3	4	5	6	18	19	11	16	13	l

- 100

Homogeneity score: 0.5653669230886433 Completeness score: 0.5873123138431823 V-measure score: 0.5761307137854349 Adjusted Rand score: 0.4405249114723275

Adjusted mutual information score: 0.5747234858808318

0 -	603.000	1.000	0.000	1.000	0.000	0.000	1.000	1.000	2.000	1.000	0.000	1.000	14.000	8.000	5.000	129.000	6.000	21.000	5.000	0.000	0
1 -	1.000	624.000	50.000	85.000	3.000	27.000	9.000	6.000	2.000	4.000	2.000	2.000	137.000	1.000	5.000	1.000	2.000	0.000	11.000	1.000	1
2 -	1.000	180.000	476.000	155.000	5.000	71.000	13.000	3.000	3.000	0.000	12.000	1.000	42.000	0.000	9.000	1.000	2.000	1.000	10.000	0.000	2
3 -	2.000	20.000	91.000	719.000	3.000	13.000	64.000	5.000	2.000	0.000	2.000	1.000	40.000	1.000	10.000	1.000	1.000	0.000	7.000	0.000	3
4 -	2.000	46.000	59.000	649.000	5.000	20.000	76.000	11.000	5.000	5.000	1.000	0.000	52.000	2.000	5.000	2.000	6.000	2.000	15.000	0.000	4
5 -	2.000	153.000	71.000	16.000	1.000	633.000	9.000	4.000	2.000	0.000	3.000	1.000	66.000	0.000	8.000	2.000	3.000	0.000	13.000	1.000	5
6 -	4.000	24.000	37.000	205.000	6.000	4.000	457.000	44.000	5.000	3.000	10.000	2.000	78.000	4.000	4.000	5.000	5.000	0.000	78.000	0.000	6
7 -	3.000	5.000	2.000	10.000	1.000	4.000	22.000	737.000	47.000	2.000	6.000	0.000	50.000	6.000	4.000	7.000	14.000	1.000	69.000	0.000	7
8 -	4.000	3.000	5.000	11.000	0.000	1.000	22.000	82.000	758.000	4.000	8.000	3.000	33.000	3.000	5.000	3.000	9.000	9.000	33.000	0.000	8
9 -	9.000	5.000	1.000	2.000	3.000	0.000	4.000	20.000	7.000	795.000	57.000	2.000	28.000	2.000	12.000	4.000	8.000	0.000	35.000	0.000	9
10 -	2.000	3.000	0.000	4.000	1.000	0.000	4.000	2.000	2.000	18.000	907.000	0.000	19.000	1.000	1.000	1.000	3.000	4.000	27.000	0.000	10
11 -	5.000	18.000	9.000	4.000	0.000	1.000	7.000	0.000	2.000	0.000	2.000	824.000	37.000	5.000	1.000	1.000	58.000	2.000	15.000	0.000	1:
12 -	12.000	45.000	50.000	135.000	0.000	4.000	370.000	113.000	13.000	2.000	4.000	3.000	167.000	5.000	23.000	7.000	4.000	1.000	24.000	2.000	1:
13 -	40.000	13.000	7.000	4.000	2.000	3.000	15.000	6.000	12.000	3.000	0.000	0.000	76.000	662.000	55.000	19.000	14.000	0.000	58.000	1.000	13
14 -	7.000	22.000	3.000	3.000	0.000	2.000	10.000	13.000	3.000	2.000	5.000	2.000	36.000	9.000	823.000	3.000	23.000	3.000	17.000	1.000	14
15 -	63.000	2.000	1.000	3.000	0.000	1.000	1.000	2.000	2.000	0.000	2.000	2.000	37.000	11.000	10.000	817.000	10.000	18.000	13.000	2.000	15
16 -	7.000	2.000	0.000	5.000	0.000	0.000	7.000	9.000	7.000	3.000	3.000	9.000	18.000	3.000	7.000	13.000	781.000	6.000	30.000	0.000	16
17 -	12.000	3.000	0.000	0.000	0.000	0.000	4.000	1.000	8.000	2.000	1.000	2.000	22.000	3.000	0.000	12.000	39.000	603.000	27.000	201.000	11
18 -	21.000	1.000	6.000	1.000	1.000	1.000	4.000	1.000	12.000	2.000	2.000	3.000	13.000	35.000	9.000	198.000	382.000	27.000	54.000	2.000	18
19 -	176.000	3.000	2.000	1.000	1.000	3.000	1.000	1.000	2.000	8.000	0.000	0.000	19.000	10.000	5.000	268.000	86.000	3.000	38.000	1.000	19
	11	12	6	13	8	19	1	18	5	17	9	14	10	15	o	2	7	4	3	16	

- 800 - 600 200 \*\*\*\*\*\*\* This is for components = 20 and metric = euclidean\*\*\*\*\*\*\*\*\*\*\*\*\*\*

- 100

20

Homogeneity score: 0.01604421948605504 Completeness score: 0.01686563660129975 V-measure score: 0.01644467691897745 Adjusted Rand score: 0.00329081513922261

Adjusted mutual information score: 0.01320192276043358

0 -	81.000	51.000	0.000	0.000	17.000	41.000	58.000	0.000	49.000	52.000	36.000	81.000	45.000	46.000	86.000	45.000	47.000	62.000	1.000	1.000
1 -	47.000	104.000	0.000	10.000	34.000	69.000	94.000	0.000		77.000	56.000		73.000	60.000	60.000	58.000	57.000	21.000	11.000	0.000
2 -		55.000	10.000	19.000	31.000	58.000	68.000	0.000	57.000	69.000	64.000	61.000	60.000	46.000	70.000	51.000	62.000	90.000	39.000	0.000
3 -		77.000	0.000	39.000	27.000	55.000	93.000	0.000		99.000	45.000		68.000	53.000	57.000	65.000	53.000	23.000	11.000	0.000
4-	62.000	75.000	0.000	0.000	39.000	62.000	85.000	0.000		77.000	52.000	55.000	62.000	62.000	74.000	72.000	55.000	31.000	30.000	2.000
5 -	78.000	88.000	0.000	0.000	21.000	76.000	101.000	0.000		83.000	56.000		73.000	52.000	51.000	61.000	35.000	30.000	41.000	0.000
6 -	29.000	106.000	0.000	20.000	35.000	84.000	110.000	7.000	98.000	105.000	22.000	42.000	85.000	67.000	28.000	87.000	25.000	13.000	11.000	1.000
7 -	74.000	84.000	0.000	8.000	24.000	57.000	79.000	0.000	69.000	69.000	55.000		65.000	52.000	64.000	58.000	57.000	54.000	41.000	0.000
8 -	83.000	70.000	0.000	0.000	27.000	50.000	83.000	0.000		74.000	64.000		71.000	44.000	74.000	52.000	53.000	65.000	32.000	1.000
9 -	87.000	51.000	0.000	22.000	20.000	44.000		0.000		92.000	52.000		63.000	50.000	87.000	53.000		60.000	33.000	0.000
10 -	86.000	58.000	0.000	24.000	22.000	52.000	64.000	0.000		51.000	90.000	85.000	46.000	47.000		40.000		74.000	59.000	1.000
11 -	107.000	54.000	0.000	9.000	20.000			0.000		54.000		116.000	56.000	47.000	87.000	62.000	46.000	83.000	4.000	0.000
12 -	81.000	62.000	0.000	9.000	27.000	45.000		0.000		72.000	65.000		75.000	60.000	89.000	56.000	37.000	48.000	39.000	1.000
13 -	70.000	67.000	0.000	7.000	23.000	59.000	75.000	0.000		86.000	69.000		68.000	57.000	73.000	66.000	54.000	36.000	31.000	0.000
14 -		69.000	0.000	6.000	23.000	41.000	69.000	0.000	58.000	72.000	52.000	94.000	64.000	65.000	122.000	56.000		46.000	0.000	0.000
15 -	71.000	80.000	0.000	0.000	25.000	45.000	89.000	0.000	87.000	77.000	72.000	62.000	65.000	65.000	72.000	85.000	53.000	33.000	16.000	0.000
16 -	50.000	59.000	0.000	22.000	22.000	54.000	47.000	0.000	44.000	52.000	67.000	84.000	49.000	59.000	78.000	51.000	98.000	58.000	16.000	0.000
17 -	86.000	53.000	0.000	15.000	14.000	27.000	52.000	0.000	49.000	39.000	71.000	119.000	40.000	34.000	70.000	33.000	93.000	111.000	34.000	0.000
18 -	61.000	42.000	0.000	27.000	18.000	32.000	43.000	0.000	53.000	45.000	47.000	66.000	53.000	31.000	58.000	51.000	56.000	45.000	47.000	0.000
19 -	55.000	41.000	0.000	0.000	19.000	33.000	45.000	0.000	36.000	46.000	35.000	56.000	45.000	27.000	59.000	27.000	55.000	43.000	6.000	0.000
	4	11	9	17	19	7	Ó	16	18	8	14	13	2	12	15	1	3	10	6	5

Homogeneity score: 0.5483641463877107 Completeness score: 0.5751842516757032 V-measure score: 0.5614540890796574 Adjusted Rand score: 0.4210116316148985

Adjusted mutual information score: 0.5599906979952898

0 -	543.000	3.000	0.000	1.000	0.000	0.000	15.000	4.000	3.000	0.000	0.000	1.000	1.000	7.000	7.000	172.000	6.000	23.000	12.000	1.000	0
1 -	1.000	671.000	0.000	88.000	4.000	84.000	64.000	5.000	4.000	3.000	2.000	2.000	13.000	8.000	19.000	0.000	3.000	1.000	0.000	1.000	1
2 -	1.000	141.000	10.000	188.000	3.000	564.000	39.000	4.000	2.000	1.000	3.000	1.000	11.000	2.000	10.000	2.000	2.000	1.000	0.000	0.000	2
3 -	3.000	39.000	0.000	737.000	1.000	83.000	31.000	3.000	1.000	0.000	2.000	1.000	65.000	4.000	9.000	1.000	2.000	0.000	0.000	0.000	3
4 -	4.000	30.000	0.000	665.000	5.000	92.000	40.000	10.000	3.000	4.000	2.000	3.000	84.000	4.000	6.000	1.000	5.000	2.000	3.000	0.000	4
5 -	1.000	144.000	0.000	20.000	2.000	735.000	54.000	3.000	2.000	1.000	1.000	1.000	9.000	0.000	6.000	3.000	3.000	1.000	1.000	1.000	5
6 -	4.000	33.000	0.000	218.000	8.000	42.000	116.000	48.000	7.000	21.000	10.000	1.000	436.000	5.000	9.000	6.000	7.000	0.000	4.000	0.000	6
7 -	4.000	5.000	0.000	8.000	2.000	6.000	58.000	780.000	48.000	2.000	6.000	2.000	20.000	19.000	6.000	3.000	14.000	2.000	5.000	0.000	7
8 -	2.000	6.000	0.000	13.000	1.000	1.000	37.000	69.000	780.000	4.000	3.000	2.000	30.000	10.000	3.000	10.000	14.000	9.000	2.000	0.000	8
9 -	3.000	6.000	0.000	6.000	2.000	1.000	48.000	22.000	7.000	795.000	61.000	0.000	11.000	7.000	7.000	6.000	9.000	2.000	1.000	0.000	9
10 -	1.000	5.000	0.000	3.000	0.000	0.000	30.000	8.000	3.000	14.000	910.000	0.000	11.000	5.000	1.000	1.000	2.000	4.000	1.000	0.000	10
11 -	5.000	23.000	1.000	4.000	0.000	7.000	43.000	6.000	2.000	1.000	0.000	826.000	5.000	6.000	1.000	1.000	57.000	3.000	0.000	0.000	11
12 -	10.000	91.000	0.000	146.000	0.000	26.000	108.000	56.000	10.000	3.000	3.000	4.000	455.000	38.000	21.000	10.000	1.000	1.000	1.000	0.000	12
13 -	32.000	19.000	0.000	4.000	2.000	13.000	95.000	4.000	5.000	1.000	3.000	0.000	17.000	700.000	52.000	14.000	18.000	0.000	10.000	1.000	13
14 -	13.000	34.000	0.000	5.000	0.000	4.000	36.000	10.000	5.000	2.000	3.000	3.000	11.000	16.000	820.000	4.000	21.000	0.000	0.000	0.000	14
15 -	27.000	6.000	0.000	1.000	0.000	2.000	39.000	2.000	2.000	0.000	2.000	3.000	2.000	16.000	6.000	823.000	10.000	15.000	39.000	2.000	15
16 -	4.000	3.000	0.000	1.000	0.000	2.000	26.000	8.000	11.000	4.000	1.000	11.000	12.000	11.000	5.000	10.000	788.000	4.000	9.000	0.000	16
17 -	7.000	2.000	0.000	0.000	0.000	2.000	34.000	2.000	6.000	1.000	3.000	3.000	5.000	9.000	5.000	9.000	45.000	600.000	4.000	203.000	17
18 -	15.000	3.000	0.000	0.000	1.000	4.000	20.000	7.000	10.000	2.000	1.000	3.000	7.000	71.000	16.000	8.000	371.000	25.000	209.000	2.000	18
19 -	144.000	3.000	0.000	1.000	2.000	1.000	17.000	7.000	3.000	8.000	0.000	0.000	1.000	58.000	6.000	264.000	84.000	6.000	22.000	1.000	19
	i	18	11	16	2	5	ó	3	15	17	4	13	6	10	8	19	7	14	12	9	

Homogeneity score: 0.015418617033985728
Completeness score: 0.016614834994086638
V-measure score: 0.015994390965556018
Adjusted Rand score: 0.003162182536014381

Adjusted mutual information score: 0.012725610671530494

0 -	91.000	51.000	0.000	35.000	44.000	57.000	35.000	63.000	58.000	45.000	36.000	49.000	1.000	52.000	75.000	40.000	44.000	23.000	0.000	0.000	0	
1 -	79.000	93.000	0.000	63.000	81.000	83.000	61.000	49.000	67.000	77.000	49.000	44.000	0.000	58.000	58.000	51.000	39.000	7.000	14.000	0.000	1	
2 -	84.000	69.000	9.000	53.000	63.000	61.000	50.000	73.000	45.000	53.000	53.000	58.000	0.000	62.000	95.000	41.000	47.000	50.000	19.000	0.000	2	- 100
3 -	65.000	75.000	0.000	64.000	88.000	69.000	60.000	46.000	73.000	69.000	35.000	78.000	0.000	58.000	61.000	54.000	39.000	7.000	41.000	0.000	3	
4 -	71.000	67.000	0.000	49.000	88.000	73.000	55.000	53.000	60.000	56.000	63.000	60.000	2.000	65.000	75.000	60.000	31.000	35.000	0.000	0.000	4	
5 -	45.000	92.000	0.000	62.000	66.000	86.000	63.000	51.000	70.000	74.000	45.000	59.000	0.000	78.000	66.000	48.000	40.000	43.000	0.000	0.000	5	
6 -	40.000	110.000	0.000	72.000	101.000	94.000	85.000	25.000	95.000	84.000	23.000	30.000	2.000	74.000	21.000	73.000	16.000	3.000	20.000	7.000	6	- 80
7 -	71.000	81.000	0.000	64.000	78.000	69.000	42.000	96.000	65.000	58.000	32.000	63.000	0.000	63.000	55.000	45.000	58.000	40.000	10.000	0.000	7	
8 -	71.000	80.000	0.000	44.000	80.000	73.000	59.000	52.000	74.000	53.000	59.000	81.000	1.000	44.000	82.000	49.000	50.000	44.000	0.000	0.000	8	
9 -	81.000	65.000	0.000	53.000	57.000	66.000	56.000	74.000	70.000	73.000	57.000	75.000	0.000	40.000	51.000	37.000	77.000	40.000	22.000	0.000	9	- 60
10 -	88.000	47.000	0.000	45.000	48.000	56.000	50.000	74.000	55.000	51.000	76.000	83.000	1.000	41.000	78.000	44.000	70.000	61.000	31.000	0.000	10	
11 -	86.000	58.000	0.000	45.000	54.000	56.000	50.000	72.000	63.000	53.000	58.000	85.000	0.000	59.000	96.000	48.000	80.000	16.000	12.000	0.000	11	
12 -	74.000	82.000	0.000	57.000	60.000	74.000	37.000	74.000	65.000	66.000	36.000	64.000	1.000	63.000	68.000	58.000	56.000	31.000	18.000	0.000	12	- 40
13 -	59.000	76.000	0.000	50.000	60.000	74.000	54.000	64.000	71.000	63.000	58.000	67.000	0.000	76.000	86.000	60.000	39.000	20.000	13.000	0.000	13	
14 -	70.000	73.000	0.000	59.000	62.000	70.000	55.000	81.000	45.000	68.000	41.000	75.000	0.000	56.000	92.000	53.000	73.000	8.000	6.000	0.000	14	
15 -	70.000	64.000	0.000	63.000	79.000	82.000	64.000	51.000	66.000	67.000	57.000	56.000	0.000	62.000	91.000	72.000	42.000	11.000	0.000	0.000	15	- 20
16 -	94.000	60.000	0.000	43.000	51.000	39.000	46.000	57.000	57.000	42.000	54.000	66.000	0.000	49.000	95.000	44.000	85.000	6.000	22.000	0.000	16	
17 -	92.000	45.000	0.000	33.000	47.000	40.000	40.000	94.000	49.000	25.000	71.000	93.000	0.000	28.000	84.000	41.000	78.000	70.000	10.000	0.000	17	
18 -	43.000	52.000	0.000	27.000	40.000	49.000	33.000	61.000	47.000	40.000	39.000	51.000	0.000	40.000	74.000	44.000	62.000	27.000	46.000	0.000	18	
19 -	62.000	39.000	0.000	31.000	48.000	38.000	34.000	27.000	30.000	36.000	48.000	61.000	0.000	36.000	61.000	20.000	50.000	7.000	0.000	0.000	19	- 0
	15	8	7	2	13	19	6	10	17	Ó	14	3	5	16	9	4	i	18	12	11	ı	

# Question 12:

Based on the matrix and the evaluation metrics, we can see that when n\_components = 5, it has a higher average performance score than the other. In terms of the metric choice, 'cosine' generally works much better than 'euclidean'. cosine similarity works better in the high-dimensional and

sparse data, and the goal is to preserve this similarity structure in the lower-dimensional representation. Cosine metric doesn't measure the magnitude. Instead, it measures the direction and orientation of the normalized vector. So, it's more suitable for high dimensional data.

#### Question 13:

Homogeneity score: 0.3488058772382193 Completeness score: 0.3977815186570097 V-measure score: 0.37168731303837976 Adjusted Rand score: 0.12268896281424234 Adjusted mutual information score: 0.369515271261787 Homogeneity score: 0.32313809864466086 Completeness score: 0.35118683066780604 V-measure score: 0.3365791172713842 Adjusted Rand score: 0.12666159829052642 Adjusted mutual information score: 0.33434061518575153 Homogeneity score: 0.2984695950441029 Completeness score: 0.347013570400772 V-measure score: 0.3209161923253039 Adjusted Rand score: 0.10409342440782972 Adjusted mutual information score: 0.31852083220776384 Homogeneity score: 0.5626567911160126 Completeness score: 0.5986812998984997 V-measure score: 0.5801103084594229 Adjusted Rand score: 0.4407262326533106 Adjusted mutual information score: 0.5787001873527008

Based on the above graph, we can clearly see that UMAP works much better on average than other models.

#### Question 14:

Homogeneity score: 0.016507812112058495 Completeness score: 0.3484280424185708 V-measure score: 0.031522168005201384

Adjusted Rand score: 0.00047947250505774106

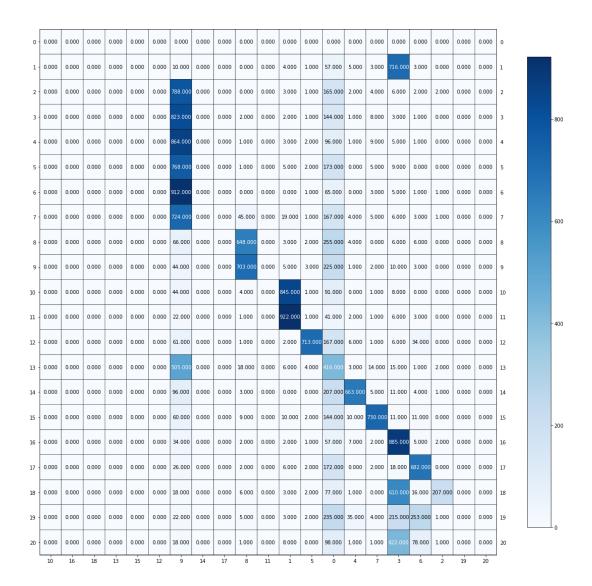
#### Question 15:

HDBSCAN clustering using n = 200,

```
Homogeneity score: 0.00037348001795512565
Completeness score: 0.08251480356407795
V-measure score: 0.0007435943654491489
Adjusted Rand score: -2.495451403736458e-06
Adjusted mutual information score: 0.00035476952203728103
Homogeneity score: 0.4015384506528541
Completeness score: 0.6125333385609805
V-measure score: 0.4850853565893493
Adjusted Rand score: 0.20812476293885468
Adjusted mutual information score: 0.48402734491274263
Homogeneity score: 0.41338483157974676
Completeness score: 0.6136479744870404
V-measure score: 0.4939915513586462
Adjusted Rand score: 0.20838513655272836
Adjusted mutual information score: 0.4929646448093636
```

# Question 16:

The model (with min\_cluster\_size = 100) gives 9 clusters. And we have found around 3300 items that have a label of '-1', and it refers to that data that does not belong to any cluster (they may belong to some clusters that contain less than 100 data points). They are classified as cluster 0 in the contingency matrix. (From the column with label '0', we can find how many data points in each class are classified as '-1', being to none clusters).



## Question 17:

Based on our experiments, UMAP with n\_components = 5 and metric = 'cosine' and K-Means(k = 20) have the best performance score (because almost all the measure scores are the highest), reaching v-measures about 0.582. This combination works the best because SVD and NMF are both linear methods, while UMAP is a non-linear technique that could preserve the local structure of the data. So, UMAP is able to capture the non-linear relationships between points in the data, and the results would be better by using 'cosine' distance metric. While the choice of hyperparameters in HDBSCAN highly affect the performance of the model, and given the data we have is well-separated, KMeans is a good choice here.

#### Question 18:

We first decided to normalize the data before we pass it into the K-Means (n\_clusters = 20) algorithm (for dimensionality reduction, we used UMAP with n\_components = ) And the result has been improved by a little. The scalar we used is the StandardScaler(with\_mean=False). The result we got is:

Homogeneity score: 0.5993390576636822 Completeness score: 0.6205075032379176 V-measure score: 0.6097396085438602 Adjusted Rand score: 0.4944043000970835

Adjusted mutual information score: 0.6084367375884226

Then, we tried to use Glove, but failed. Because we are not able to generate a keywords list for each data sample, while the original data is too large (contains too many words).

# Part 2:

#### Question 19:

The pretrained VGG neural network have discriminative power for a custom dataset because the earlier layers of VGG neural networks(VGGNN) are trained to find low level and general features such as shapes, edges, and textures that can help to distinguish between different objects and classes in an image. Therefore, when VGGNN is used on custom dataset, it will act as a important feature extraction tool to extract meaningful features in custom dataset. Those extracted features will allow users to classify their custom dataset accurately and perform their own task.

#### Question 20:

The helper code created its own feature extraction neural network called FeatureExtractor. The first few layers are extracted from VGG16 feature layers. In other words, the helper code uses pretrained VGG16 feature extraction layers as its own feature extraction layer. In VGG -16 feature extraction layers, Conv2D layers are used to extract features and each Conv2D layer is followed by a ReLU layer to ensure no-linearlity. Conv2D layers apply multiple filters to the input image, and those filters with pretrained weights allow features such as edges, shapes to be extracted.

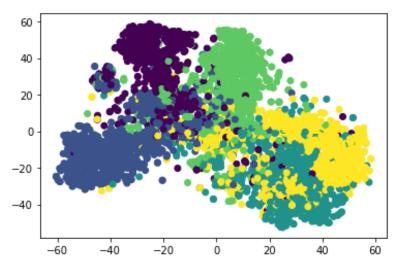
## Question 21:

The size of the input image is (320\*263 = 84160 pixels). The VGG network extracts 4096 features per image because each feature vector has 4096 dimensions.

## Question 22:

The extracted features are pretty dense compared to the TF-IDF features. In the TF-IDF feature matrix, most of the cells/elements are 0. However, the cells in the extracted feature matrix we get from the neural network are all non-zero. Therefore, the extracted features from the VGG16 network are dense compared to that of TF-IDF.

## Question 23:



After we map the features we extracted onto 2 dimensions with t-SNE, we observed clear signs of clustering from the plot.

## Question 24:

In this question, we first tried to find an appropriate value for the hyperparameter min\_cluster\_size and min\_samples in HDBSCAN (we directly used the 'f\_all' as our fitting data, and did not apply any dimensionality reduction methods here). For min\_cluster\_size, we considered the following values (100, 200, 300), and for min\_samples, we used (50, 100, 200). The combination of min\_cluster\_size & min\_samples with the best rand\_score is (300, 200). So, we used this combination in HDBSCAN for the following report.

This figure shows the rand\_score report, the first number represents the dimensionality reduction being used (0: None, 1: SVD, 2: UMAP, 3. Autoencoder), the second number represents the Clustering model we used (0: K-Means, 1: Agglomerative Clustering, 2: HDBSCAN).

```
comb_scores_dic

{0.7079360694173067: ['0; 0'],
  0.6862000871875192: ['0; 1'],
  0.20358404572368982: ['0; 2', '1; 2', '3; 2'],
  0.7078806674672471: ['1; 0'],
  0.68676286999925: ['1; 1'],
  0.7989080023141083: ['2; 0'],
  0.7797151626819594: ['2; 1'],
  0.6171422248264604: ['2; 2'],
  0.7116075997216534: ['3; 0'],
  0.6986368595263505: ['3; 1']}
```

The best result (for rand score) is about 0.7989. The Clustering model used here is the K-Means with k = 5, and the dimensionality reduction method used here is the UMAP with n\_components = 50.

```
l=list(comb_scores_dic.items())
1.sort(reverse=True)
1[0]
(0.7989080023141083, ['2; 0'])
```

#### Question 25:

The test accuracy of the MLP classifier on the original VGG features is: 0.9168937329700273. For the accuracy, we just simply used # of correct predictions / total # of data.

Then, we used PCA (num\_features = 50) for dimensionality reduction, and got a test accuracy: 0.885558583106267.

Yes, the performance of the model suffers with the reduced-dimension representations, might because the number of features has been reduced. Not significant.

Yes, HDBSCAN and other clustering models are sensitive to hyperparameter choices, i.e. the num\_cluster\_size and num\_samples we chose might not be good enough for the data. While MLP is insensitive to the number of features (num\_features = 50 in the case with PCA, and 4096 on the original VGG features), and it can be used for supervised learning tasks, and with the help of cuda, it is capable of learning complex non-linear relationships between features and labels. So, it makes sense that MLP's performance is better than that of methods in Q24.