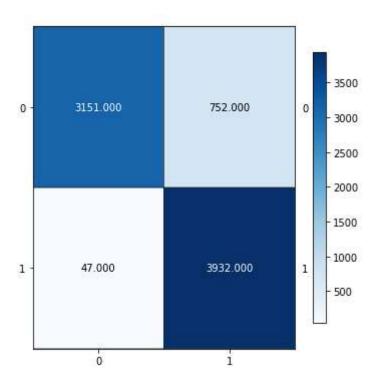
Project 2 Report

Haoting Ni (905545789), Yikai Wang (905522085), Yuanxuan Fang (005949389)

Question 1:

The shape of TF-IDF matrix is (7882, 21798)

Question 2:



The contingency matrix is shown above. The contingency matrix is squared most of the time, but it also exists that is not squared.

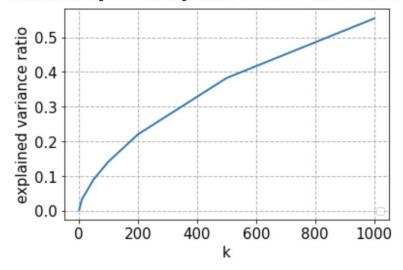
Question 3:

Homogeneity: 0.5774035409987348 Completeness: 0.5926514149929692 V-measure: 0.5849281246876363

Adjusted Rand Index: 0.6355770387145221

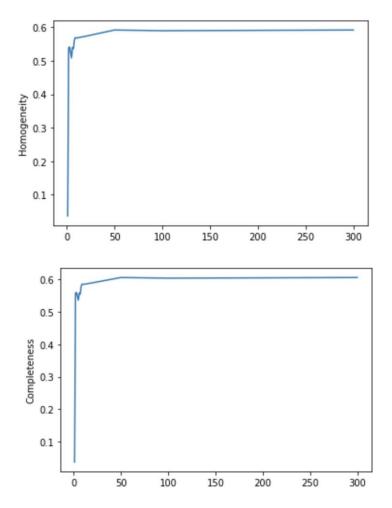
Question 4:

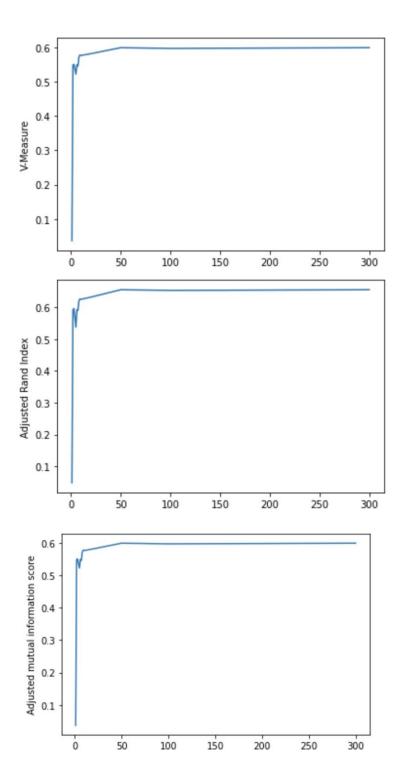
WARNING: matplotlib.legend: No handles with labels found to put in legend.



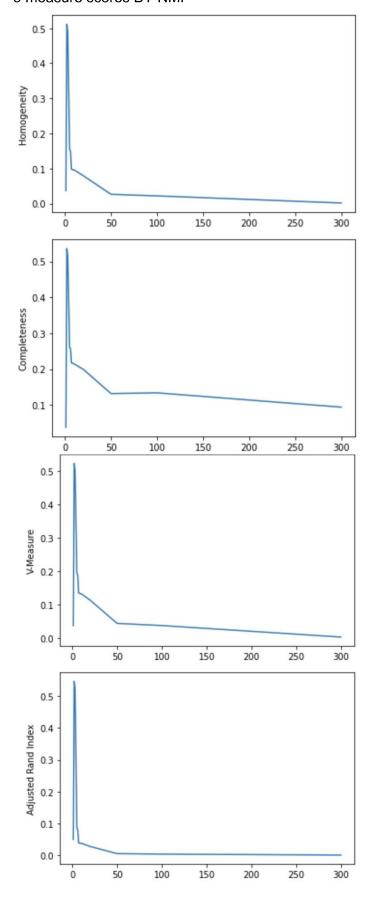
The plot is shown above as x-axis is k from 1 to 1000, y-axis is the explained variance ratio with respect to k.

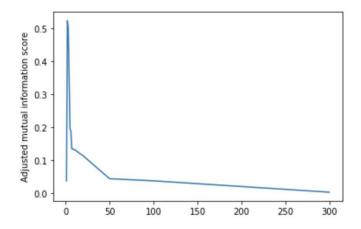
Question 5: 5 measure scores BY SVD:





5 measure scores BY NMF





These plots are different r as x-axis with respect to the y-axis of Homogeneity, Completeness, V-measure, Adjusted Rand Index, and Adjusted mutual information score. For SVD, the best r is 100, and for NMF, the best r is 2.

Question 6:

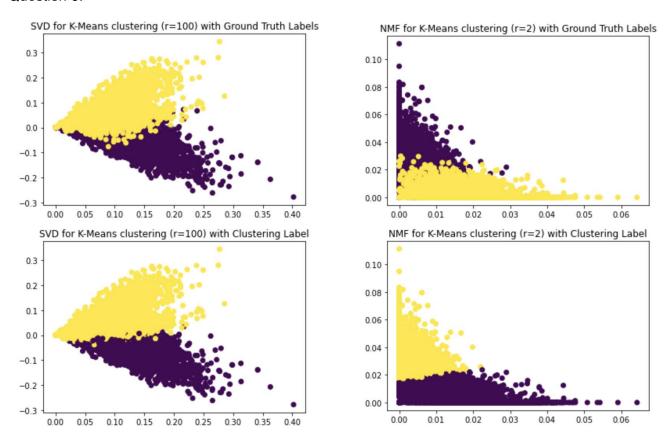
Non-monotonic behavior occurs because with a higher r, it carries more features and information. More information preserved should lead to a better clustering performance, but a large r also means more noise.

NMF shows a hard drop after r=2 because NMF only allows positive entries in reduced dimensional matrices.

Question 7:

These average measures are a little less than what we did in Question 3.

Question 8:



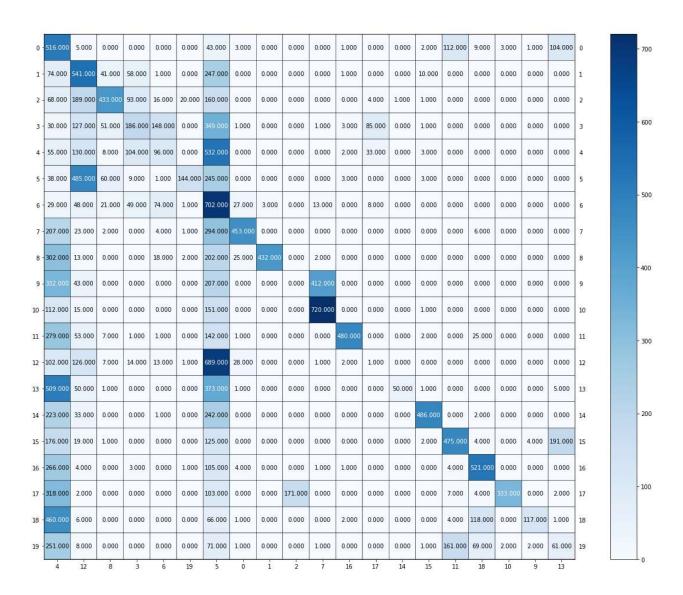
The visualization of the SVD and NMF for K-means clustering are shown above.

Question 9:

We observe that there are two clear clusters appearing on the graphs, and they separate by a minimal gap of euclidean distance. It barely draws a decision boundary between two clusters. However, it is not like K-means assumes that is centralized and clusters are far away from each other. So, the data is not ideal for K-means.

Question 10:

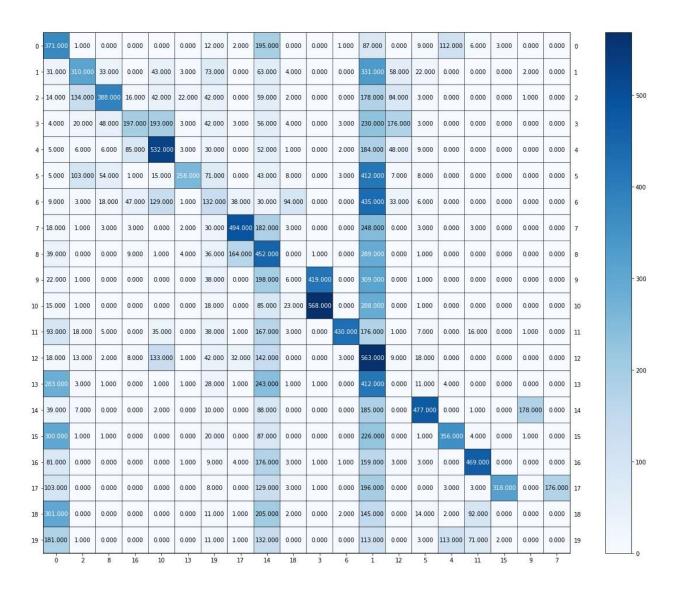
As the whole 20 classes loaded, the TF-IDF matrix has the shape of (18846, 42170). For SVD, we choose the components of 100, For NMF, we choose the components of 20, and we both get it from Question 8.



The contingency matrix of SVD for k-means is shown above and 5 metrics:

Homogeneity: 0.36344527780811314 Completeness: 0.4636394506312786 V-measure: 0.40747353479839793

Adjusted Rand Index: 0.12180153342657082



The contingency matrix of NMF for k-means is shown above and 5 metrics:

Homogeneity: 0.33395741408685753 Completeness: 0.39777564656828923 V-measure: 0.36308357092882243

Adjusted Rand Index: 0.10406823198250942

Question 11:

		-1 7/		0	45			5	comp	onent	s + eu	clidea	n						e y	2	_	
0 -		32.000	37.000	31.000	45.000		36.000	35.000		65.000	22.000							70.000	14.000	1.000	0	
1 -	49.000	64.000	51.000	62.000	55.000	41.000	40.000	65.000	40.000	76.000	16.000	61.000		56.000	59.000	66.000	51.000	45.000	20.000	0.000	1	
2 -	44.000	45.000	67.000	42 000	60,000	53:000	50.000	49 000	52.000	65:000	55.000	53.000	48.000	54.000	36.000	63.000	58.000	44.000	47.000	0.000	2	- 80
3 -	31.000	51.000	61.000	59.000	59.000	62.000	50.000	48.000	42 000	65:000	61.000	44.000	51.000	53.000	50.000	73.000	49.000	68.000	5.000	0.000	3	
4 -	50.000	43:000	57:000	54.000	67.000	58.000	31.000	46.000	54.000	72.000	34.000	48.000	58.000	55:000	47.000	52.000	54.000	65.000	16.000	2.000	4	
5 -	44.000	52.000	61.000	61.000	54.000	73.000	36.000	38.000	49.000	73.000	34.000	43.000	55.000	59.000	50.000	62.000	51.000	67:000	26.000	0.000	5	
6 -	43.000	55.000	59.000	64.000	58.000	54.000	71.000	56.000	58.000	72.000	9.000	34.000	56.000	59.000	38.000	55.000	42.000	57.000	33.000	2.000	6	100
7 -	34.000	59,000	52.000	50.000	49.000	57.000	57.000	63.000	44.000	71.000	23.000	57.000	40.000	57.000	49.000	58.000	55.000	65.000	50.000	0.000	7	- 60
8 -	42 000	51.000	64.000	51.000	61.000	53.000	56.000	48 000	57.000	72.000	26.000	56.000	53.000	52.000	40.000	54.000	56.000	61.000	42.000	1.000	8	
9 -	34.000	61.000	59:000	46.000	66.000	68.000	39.000	47.000	53.000	91.000	52.000	40.000	59.000	61.000	34.000	59.000	43.000	53.000	29.000	0.000	9	
10 -	31.000	53.000	48.000	41.000	45:000	60.000	41.000	52.000	47.000	82.000	64.000	47.000	61.000	60.000	52.000	67.000	44 000	53.000	50 000	1.000	10	
11	41 000	55.000	53.000	54.000	62.000	50.000	53.000	39.000	48.000	73.000	45.000	73.000	44.000	55:000	54.000	59.000	53.000	60.000	20.000	0.000	11	- 40
12 -	36.000	41.000	63.000	56.000	47.000	56.000	36.000	48 000	53.000	81.000	59.000	66.000	56.000	48,000	45 000	69.000	48.000	49.000	26.000	1.000	12	
13 -	46.000	45.000		57.000	59.000	48.000		44 000	52.000		41 000	51.000				59.000	51.000			0.000	13	
750	Entrare Con		No.	\$500000000	Terror of	estimate in	ACCOUNTS	49.000		0000000		Name of the last		NOTICE OF		CONTRACTOR OF THE PARTY OF THE	NAME OF THE OWNER, OF T			(2004-20 (4004-20		
14 -	44.000	54.000	60.000	43.000	52.000	65.000			43.000	65.000	25.000	46.000	49.000	61,000	61.000		55:000		26.000	0.000	14	
15 -	36.000	48.000	63.000	62.000	49.000	60,000	43.000	58.000	53.000	81.000	12.000	46.000	59.000	52.000	50.000	80.000	47 000	73.000	25.000	0.000	15	- 20
16 -	44.000	45.000	35.000	44.000	52,000	42.000	48.000	42.000	40.000	61.000	30.000	55.000	45.000	53.000	54.000	65.000	55.000	52 000	48.000	0.000	16	
17	27.000	46,000	42.000	39.000	42.000	51.000	46.000	60.000	36.000	65.000	56.000	42.000	46.000	47.000	46.000	54.000	43 000	74.000	78.000	0.000	17	
18 -	34.000	39.000	28.000	47.000	51.000		45.000	41.000	36.000	50.000	12.000	29.000	36.000	45.000	27.000	40.000	43.000	45.000	80.000	0.000	18	
19 -	37.000	34.000	36.000	29.000	41.000	25.000	19.000	36.000	22.000	67.000	16.000	29.000	47.000	31.000	34.000	31.000	38.000	43.000	13.000	0.000	19	-ano
	10	14	ó	18	ıi	2	16	9	i	13	19	12	4	7	15	8	3	5	6	17		- 0

Nomogeneity: 0.006963310345264541 Completeness: 0.007098643880587983 V-measure: 0.007030325881757074 Adjusted Rand Index: 0.000860714947023135 Adjusted mutual information score: 0.0038033666578713276

5 components + cosine

- 800

- 600

400

200

19	16	3	15	18	2	6	Ó	4	8	11	10	i	9	17	13	12	14	5	7	
147.000	14.000	2.000	0.000	0.000	8.000	3.000	2.000	3.000	43.000	2.000	1.000	0.000	3.000	5.000	279.000	16.000	19.000	13.000	68.000	19
29.000	15.000	5.000	2.000	1.000	0.000	4.000	7.000	4.000	171.000	7.000	2.000	2.000	35.000	21.000	8.000	132.000	27.000	186.000	117.000	18
4.000	25.000	0.000	0.000	0.000	2.000	3.000	3.000	8.000	28.000	5.000	2.000	208.000	3.000	2.000	15.000	17.000	613.000	1.000	1.000	17
9.000	23.000	0.000	0.000	0.000	3.000	14.000	2.000	1.000	33.000	9.000	8.000	2.000	5.000	2.000	7.000	399.000	12.000	2.000	379.000	16
62.000	39.000	1.000	1.000	0.000	2.000	3.000	2.000	2.000	19.000	9.000	1.000	2.000	14.000	9.000	756.000	0.000	14.000	56.000	5.000	15
14.000	62.000	7.000	2.000	1.000	19.000	9.000	18.000	1.000	15.000	6.000	4.000	1.000	7.000	798.000	6.000	10.000	1.000	1.000	5.000	14
28.000	81.000	4.000	5.000	2.000	11.000	14.000	11.000	17.000	24.000	6.000	2.000	1.000	707.000	52.000	15.000	0.000	2.000	4.000	4.000	13
8.000	187.000	52.000	133.000	0.000	38.000	368.000	125.000	6.000	3.000	8.000	8.000	2.000	24.000	20.000	0.000	1.000	0.000	0.000	1000	12
3.000	62.000	10.000	3.000	0.000	17.000	12.000	0.000	1.000	12.000	3.000	793.000	1.000	1.000	2.000	2.000	58.000	8.000	0.000	3.000	11
0.000	39.000	0.000	3.000	0.000	0.000	5.000	3.000	1.000	2.000	940.000	0.000	0.000	2.000	0.000	1.000	1.000	0.000	1.000	1.000	10
4.000	59.000	1.000	3.000	0.000	3.000	8.000	14.000	3.000	7.000	877.000	2.000	0.000	2.000	1.000	3.000	4.000	1.000	1.000	1.000	9
2.000	39.000	4.000	11.000	1.000	3,000	15.000	62.000	814.000	15.000	10.000	1.000	0.000	2.000	3.000	4.000	4.000	2.000	2.000	2.000	8
4.000	57.000	3.000	7.000	1.000	7.000	25.000	780.000	54.000	13.000	6.000	1.000	0.000	12.000	4.000	5.000	4.000	1.000	3.000	3.000	7
3.000	100.000	39.000	190.000	3.000	28.000	467.000	60.000	10.000	22.000	29.000	4.000	0.000	2.000	9.000	4.000	2.000	0.000	3.000	0.000	6
3.000	86.000	84.000	13.000	2.000	771.000	11.000	5.000	1.000	3.000	1.000	0.000	1.000	0.000	4.000	1.000	2.000	0.000	0.000	0.000	5
1.000	44.000	80.000	637.000	3.000	51.000	121.000	12.000	3.000	3.000	0.000	1.000	0.000	1.000	4.000	0.000	0.000	1.000	0.000	1.000	4
1.000	39.000	90.000	731.000	0.000	27.000	77.000	4.000	2.000	0.000	2.000	2.000	0.000	1.000	5.000	1.000	0.000	0.000	0.000	0.000	3
2.000	46.000	472.000	149.000	14.000	252.000	17.000	6.000	0.000	4.000	9.000	2.000	0.000	0.000	8.000	2.000	0.000	1.000	0.000	1000	2
1.000	198.000	42.000	94.000	5.000	588.000	13.000	3.000	4.000	10.000	4.000	1.000	2.000	0.000	6.000	1.000	0.000	1.000	0.000	0.000	1
528.000	19.000	0.000	0.000	1.000	1.000	2.000	2.000	0.000	6.000	1.000	1.000	0.000	3.000	6.000	157.000	4.000	45.000	17.000	6.000	0

19 16 3 15 18 2 6 0 4

Homogeneity: 0.5579646383397878

Completeness: 0.5897530610088147

V-measure: 0.5734186265182245

Adjusted Rand Index: 0.42012638848880346

Adjusted mutual information score: 0.5719905310550398

								20	com	onent	s + e	uclidea	an								
0	75.000	35.000	15.000	42.000	34.000	39.000	35.000	47.000	58.000	41.000	28.000		44.000	58.000	36.000	38.000	47.000	18.000	1.000	57.000	0
1	53.000		25.000				59.000	62.000		43.000					56.000	58.000	63.000	20.000	0.000	66.000	1
2	62.000	45.000		43.000	50.000	52.000	37.000	33.000	42.000	55.000	40.000	61.000	42.000	66.000	56.000		59.000	44 000	9.000	69.000	2
3	75.000	40.000	50.000	61.000	40.000	59.000	54.000	44.000				67.000	41.000	55.000		68.000		14.000	0.000	62.000	3
4	61.000	53.000	8.000	51.000	60.000	54.000		46.000	59.000	53.000		64.000	34.000	61.000		51.000		41.000	2.000	69.000	4
5	62.000	41.000	17.000	57.000	55.000	67.000		57.000	61.000	50.000		44.000	39.000	59.000	65:000	51.000	55.000	24.000	0.000	84.000	5
6	60.000	40.000	31.000	54.000	44.000	55.000	62.000	43.000		50.000		51.000	46.000	70.000	59.000	42.000	65.000	8.000	2.000	80.000	6
7	69.000	53.000	40.000	58.000	50.000	57.000	45.000	60.000	49.000	48.000	36.000	61.000	38.000	54.000	56.000	54.000	69.000	22.000	0.000	71.000	7
8	63.000	33.000	35.000	54.000	44.000	61.000	51.000	46.000	69.000	56.000	48.000	75.000	35.000	51.000	64.000	46.000	55.000	43 000	1.000	66.000	8
9	59.000	41.000	49:000	68.000	43.000	52.000	43.000	30.000	51.000	70.000	45.000	48.000	42.000	55.000	58.000	55.000	55.000	47.000	0.000	83.000	9
10	66.000	35.000	45.000	64.000	41.000	55.000	46.000	48.000		39.000	61.000	55:000	54.000	56.000		56.000	50.000		1.000	64.000	10
11	65.000	29.000	31.000	56,000	48.000	51.000	52.000	60.000	60.000	53.000	40.000	76.000	30.000	64.000		55.000	55.000	46.000	0.000	61.000	11
12	56.000	35.000	22.000	49.000	48.000	49.000		44:000		51.000	44.000	68.000	49.000	58.000	57.000	54.000	40.000	63.000	1.000	80.000	12
13	61.000	41 000	28.000	48.000	57.000	47.000		60.000		37.000		71.000	41.000	75.000		62.000	57.000	45.000	0.000	60.000	13
14	77.000	36.000	17.000	60.000	54.000	50.000		58.000	47.000	62.000		58.000	44.000	68.000	61.000	41 000	68.000	22.000	0.000	66.000	14
15	80.000	42.000	17.000	39.000	473000	52.000	60.000	56.000	60:000	44.000	56.000		46.000	56.000		71.000	51.000	32.000	0.000	78.000	15
16	67.000	24.000	27.000	40.000	46.000	45.000	48.000	52.000	47.000	48.000	41.000	61.000	38.000	65.000	34.000	59.000	62.000		0.000	53.000	16
17	74.000	32.000	69.000	52.000	29.000	55.000	40.000	45.000	45.000	38.000	40.000	53.000	42.000	40.000			48.000	72.000	0.000	61.000	17
18	48.000	33.000	50.000	37.000	35.000	45.000	44.000	36.000	41.000	39.000	35.000	45.000	22.000	47.000	26.000	46.000	47.000	48.000	0.000	51.000	18
19	34.000	34.000	14.000	36.000	33.000	21.000	24.000	33.000	26.000	39.000	36.000	30.000	24.000	36.000	46,000	39.000	34.000	16.000	0.000	73.000	19
	8	12	ó	3	13	18	6	7	5	14	10	19	i	11	17	4	16	9	15	2	

- 60

- 50

- 40

- 30

- 20

- 10

Homogeneity: 0.006727234238865325 Completeness: 0.006856836499122388 V-measure: 0.006791414174403483 Adjusted Rand Index: 0.0007462319543346768 Adjusted mutual information score: 0.0035326420178268107

20		
	components	

12	18	11	i	2	8	10	13	5	7	9	19	4	17	16	14	3	6	15	Ó
158.000	19.000	0.000	0.000	0.000	5.000	0.000	1.000	0.000	3.000	1.000	0.000	1.000	3.000	6.000	281.000	88.000	13.000	13.000	36.000
24.000	11.000	5.000	0.000	0.000	0.000	5.000	6.000	2.000	1.000	5.000	3.000	3.000	34.000	19.000	9.000	319.000	25.000	186.000	118.000
6.000	12.000	0.000	0.000	0.000	2.000	3.000	2.000	8.000	207.000	4.000	3.000	2.000	3.000	0.000	14.000	33.000	623.000	1.000	17.000
12.000	23.000	0.000	1.000	1.000	4.000	7.000	2.000	1.000	1.000	6.000	11.000	0.000	5.000	2.000	7.000	799.000	9.000	1.000	18.000
48.000	30.000	2.000	1.000	0.000	3.000	1.000	2.000	1.000	2.000	1.000	2.000	2.000	13.000	3.000	779.000	10.000	17.000	54.000	26.000
26.000	53.000	3.000	1.000	3.000	31.000	1.000	11.000	0.000	0.000	7.000	1.000	2.000	7.000	806.000	4.000	18.000	1.000	1.000	11.000
35.000	82.000	6.000	1.000	4.000	13.000	4.000	6.000	10.000	0.000	4.000	2.000	14.000	688.000	47.000	15.000	9.000	3.000	2.000	45.000
11.000	146.000	51.000	23.000	106.000	47.000	40.000	80.000	21.000	0.000	11.000	10.000	397.000	2.000	5.000	7.000	6.000	1.000	0.000	20.000
2.000	39.000	10.000	3.000	8.000	19.000	1.000	1.000	2.000	0.000	4.000	809.000	18.000	2.000	2.000	1.000	59.000	3.000	0.000	8.000
1000	33.000	0.000	2.000	3.000	2.000	5.000	2.000	2.000	0.000	936.000	0.000	0.000	2.000	0.000	1.000	3.000	0.000	0.000	7.000
2.000	57.000	1.000	0.000	1.000	2.000	3.000	15.000	3.000	0.000	884.000	3.000	0.000	0.000	0.000	4.000	5.000	1.000	1.000	12.000
3,000	37.000	2.000	4.000	6.000	3.000	8.000	66.000	822.000	0.000	9.000	1.000	5.000	1.000	2.000	4.000	5.000	9.000	1.000	8.000
4.000	50.000	5.000	4.000	4.000	6.000	8.000	791.000	51.000	0.000	3.000	1.000	9.000	6.000	8.000	3.000	10.000	0.000	4.000	23.000
6.000	116.000	45.000	77.000	128.000	21.000	413.000	64.000	9.000	0.000	18.000	2.000	39.000	4.000	10.000	4.000	6.000	0.000	3.000	10.000
5.000	71.000	81.000	1.000	15.000	778.000	7.000	4.000	2.000	0.000	2.000	1.000	4.000	0.000	5.000	1.000	2.000	0.000	0.000	9.000
0.000	51.000	82.000	152.000	519.000	41.000	36.000	12.000	2.000	0.000	2.000	2.000	48.000	0.000	5.000	2.000	4.000	2.000	0.000	3.000
1.000	41.000	91.000	306.000	456.000	20.000	34.000	3.000	1.000	0.000	3.000	1.000	18.000	1.000	2.000	1.000	1.000	0.000	0.000	2.000
4.000	53.000	493.000	29.000	126.000	238.000	6.000	3.000	1.000	0.000	9.000	2.000	5.000	2.000	5.000	1.000	2.000	0.000	0.000	6.000
1.000	132.000	62.000	1.000	106.000	606.000	7.000	5.000	4.000	1.000	5.000	3.000	2.000	1.000	6.000	2.000	2.000	0.000	1.000	26.000
544.000	12.000	1.000	0.000	0.000	1.000	1.000	1.000	0.000	0.000	1.000	1.000	0.000	2.000	4.000	162.000	12.000	36.000	10.000	11.000

- 400

- 200

12 18 11 1 2 8 10 13 5

Homogeneity: 0.5727266672477204

Completeness: 0.5950927361971373

V-measure: 0.5836955243253182

Adjusted Rand Index: 0.44247696596196257

Adjusted mutual information score: 0.5823243993906418

200 components + euclidean 39.000 29.000 38.000 37.000 37.000 43.000 40.000 41.000 43.000 31.000 42.000 18.000 41.000 1.000 0 0 43.000 24.000 26.000 0.000 38.000 32.000 40.000 39.000 10.000 2 43.000 42,000 18.000 40,000 37.000 0.000 38.000 28.000 41.000 29.000 2.000 41.000 14.000 43.000 0.000 5 79.000 70.000 43.000 35.000 39.000 76.000 73.000 43.000 19.000 31.000 2.000 6 86.000 73.000 40.000 41.000 23.000 37.000 40.000 0.000 41.000 32.000 38.000 38.000 39.000 1.000 8 36.000 41.000 84.000 78.000 70.000 36.000 43.000 35.000 39.000 0.000 9 10 42.000 41.000 37.000 33.000 35.000 41.000 1.000 10 11 27.000 72.000 68.000 43.000 24.000 0.000 11 12 40 000 39.000 86.000 79.000 16.000 1 000 12 57.000 66.000 41.000 13 42 000 39 000 28.000 0.000 13 84.000 41.000 22.000 22.000 41.000 0.000 14 14 41.000 40.000 13.000 0.000 35.000 15 15 39.000 31.000 37.000 39.000 37.000 0.000 16 16 38.000 17 32.000 39.000 39.000 39.000 35.000 32.000 38.000 0.000 17 37.000 35.000 35.000 42.000 36.000 42.000 40.000 41.000 29.000 31.000 13.000 32.000 82 000 0.000 18 19 19.000 33.000 26.000 42.000 27.000 41.000 41.000 23.000 33.000 26.000 7.000 34.000 36.000 33.000 26.000 23.000 0.000 19

15

16 14 19 13 11 60

40

20

12

3

0 Homogeneity: 0.007043579932128244 Completeness: 0.007179839386213435 V-measure: 0.007111056980711039

2 18

Adjusted Rand Index: 0.0009614342903204385

Adjusted mutual information score: 0.0038519780440671642

10 17

								2	00 co	mpone	ents +	cosin	e									
0 -	1.000	0.000	1.000	0.000	1.000	2.000	1.000	1.000	1.000	8.000	1.000	1.000	13.000	1.000	9.000	675.000	5.000	56.000	16.000	6.000	0	
1-	2.000	5.000	57.000	86.000	43.000		11.000	2.000	7.000	2.000	1.000	2.000	151.000	0.000	4.000	3.000	0.000	1.000	0.000	0.000	1	
2 -	0.000	5.000	490.000	116.000	88.000	178.000	7.000	4.000	0.000	7.000	9.000	10.000	60.000	2.000	5.000	3.000	0.000	0.000	0.000	1.000	2	
3 -	0.000	3.000	71.000	632.000	142.000	18.000	63.000	2.000	1.000	1.000	3.000	3.000	37.000	1.000	3.000	1.000	0.000	0.000	1.000	0.000	3	- 80
4 -	0.000	5.000	21.000	589.000	173.000	14.000	94.000	7.000	3.000	3.000	2.000	7.000	35.000	0.000	4.000	1.000	0.000	2.000	0.000	3.000	4	
5 -	1.000	2.000	79.000	13.000	23.000	760.000	11.000	3.000	1.000	3.000	4.000	0.000	74.000	0.000	7.000	6.000	1.000	0.000	0.000	0.000	5	
6 -	0.000	4.000	34.000	170.000	73.000	9.000	452.000	64.000	8.000	18.000	20.000	2.000	102.000	1.000	9.000	6.000	1.000	0.000	2.000	0.000	6	
7 -	0.000	1.000	3.000	5.000	0.000	8.000	25.000	791.000	54.000	12.000	7.000	0.000	49.000	11.000	4.000	5.000	8.000	1.000	5.000	1.000	7	- 60
8 -	0.000	1.000	1.000	8.000	2.000	4.000	19.000	56.000	824.000	11.000	9.000	1.000	37.000	1.000	2.000	5.000	1.000	9.000	1.000	4.000	8	
9 -	0.000	3.000	0.000	0.000	2.000	6.000	10.000	15.000	4.000	14.000	883.000	0.000	40.000	2.000	2.000	4.000	7.000	1.000	1.000	0.000	9	
10 -	0.000	2.000	1.000	2.000	0.000	0.000	3.000	2.000	0.000	5.000	941.000	0.000	34.000	2.000	1.000	2.000	2.000	0.000	1.000	1.000	10	
11 -	0.000	0.000	9.000	4.000	4.000	19.000	10.000	1.000	1.000	13.000	1.000	792.000	41.000	1.000	1.000	1.000	84.000	7.000	0.000	2.000	11	- 40
12 -	1.000	0.000	46.000	97.000	77.000	21.000	390.000	89.000	20.000	4.000	8.000	14.000	157.000	27.000	7.000	16.000	3.000	1.000	1.000	5.000	12	
13 -	0.000	2.000	4.000	3.000	10.000	5.000	19.000	7.000	9.000	15.000	4.000	2.000	83.000	708.000	53.000	48.000	3.000	3.000	5.000	7.000	13	
14 -	2.000	1.000	3.000	3.000	0.000	35.000	7.000	16.000	1.000	8.000	7.000	2.000	59.000	8.000	797.000	20.000	13.000	0.000	0.000	5.000	14	
15 -	3.000	0.000	3.000	1.000	1.000	2.000	6.000	1.000	1.000	16.000	3.000	1.000	34.000	24.000	11.000	773.000	5.000	16.000	91.000	5.000	15	- 20
16 -	1.000	0.000	1.000	0.000	0.000	2.000	7.000	8.000	0.000	24.000	5.000	8.000	17.000	5.000	6.000	14.000	426.000	12.000	4.000	370.000	16	20
17 -	209.000	0.000	0.000	0.000	3.000	2.000	6.000	3.000	8.000	13.000	0.000	2.000	21.000	3.000	0.000	18.000	33.000	614.000	3.000	2.000	17	
18 -	1.000	0.000	4.000	0.000	1.000	1.000	5.000	7.000	2.000	107.000	4.000	2.000	16.000	35.000	17.000	28.000	220.000	23.000	192.000	110.000	18	
19 -	3.000	2.000	2.000	0.000	1.000	4.000	1.000	1.000	0.000	47.000	2.000	0.000	14.000	6.000	8.000	421.000	13.000	17.000	13.000	73.000	19	
	i	9	18	15	3	2	6	11	4	7	ò	10	16	13	17	19	12	14	5	8	L	0

Homogeneity: 0.5583331319425638 Completeness: 0.5918004325527025 V-measure: 0.5745798561006596 Adjusted Rand Index: 0.42152880386103947 Adjusted mutual information score: 0.5731539609712347

All these 6 combinations of the contingency matrix and five clustering evaluations are shown as above.

Question 12:

As the comparison from the five clustering evaluations of each component respected to euclidean and cosine, cosine distance performs better. Since cosine distance is not affected by the dimensions. On the other hand, the euclidean distance will perform less well in high dimension features. For each metric choice, r = 20, cosine distance UMAP with the highest adjusted rand index (accuracy).

Question 13:

So far, K-means clustering with sparse TF-IDF has the worst performance, since it does not reduce the dimension. For SVD reduced and NMF reduced, we found out that components r = 100 and 20 respectively are the best choice for SVD reduced and NMF reduced.

Question 14:

Agglomerative, for "ward" linkage criteria, five clustering evaluations are shown below.

Homogeneity: 0.5664705707502917 Completeness: 0.5827393082216585 V-measure: 0.5744897856643124

Adjusted Rand Index: 0.425330910828589

Adjusted mutual information score: 0.5730958277607423

Agglomerative, for "single" linkage criteria, five clustering evaluations are shown below.

Homogeneity: 0.016304604013505815 Completeness: 0.36263221868186957 V-measure: 0.03120612394483502

Adjusted Rand Index: 0.0005136327378515986

Adjusted mutual information score: 0.026289810915222936

Question 15:

HDBSCAN five clustering evaluations with respect to min_clsuter_size of 20, 100, 200 are shown below:

20
Homogeneity: 0.4213436898781575
Completeness: 0.43394333368258675
V-measure: 0.42755070607911316
Adjusted Rand Index: 0.06949144572608194
Adjusted mutual information score: 0.4149583559942783
100

Homogeneity: 0.4336789202796464 Completeness: 0.6375099981242074 V-measure: 0.5162014709150505

Adjusted Rand Index: 0.21859244455476776

Adjusted mutual information score: 0.5151237897905137

200

Homogeneity: 0.41815386406777416 Completeness: 0.6154862906128278 V-measure: 0.49798369294200345

Adjusted Rand Index: 0.2115996277850882

Question 16:

The contingency matrix for HDBSCAN is shown down below:

13	Ó	3	15	12	18	9	19	17	8	14	i	4	16	2	7	11	5	6	10	20	30)	95
0.000	263.000	1000	0.000	0.000	0.000	17.000	0.000	0.000	2.000	0.000	3.000	0.000	0.000	11.000	2.000	241.000	79.000	4.000	5.000	0.000	20	
0.000	205.000	3.000	0.000	0.000	0.000	19.000	0.000	0.000	7.000	0.000	3.000	2.000	0.000	37.000	7.000	11.000	312.000	1.000	168.000	0.000	19	
0.000	222.000	207.000	0.000	0.000	0.000	19.000	0.000	0.000	7.000	0.000	3.000	2.000	0.000	3.000	0.000	4.000	18.000	454.000	1.000	0.000	18	
0.000	105.000	1.000	0.000	0.000	0.000	30.000	0.000	0.000	2.000	0.000	6.000	4.000	0.000	5.000	3.000	6.000	747.000	1.000	0.000	0.000	17	
0.000	338.000	3.000	0.000	0.000	0.000	33.000	0.000	0.000	2.000	0.000	1.000	1.000	0.000	11.000	3.000	583.000	6.000	6.000	10.000	0.000	16	
0.000	161.000	1.000	0.000	0.000	0.000	51.000	0.000	0.000	8.000	0.000	4.000	1.000	0.000	9.000	725.000	6.000	20.000	1.000	0.000	0.000	15	
0.000	199.000	1.000	0.000	0.000	0.000	88.000	0.000	0.000	8.000	0.000	4.000	2.000	0.000	679.000	0.000	7.000	1.000	1.000	0.000	0.000	14	
0.000	395.000	0.000	0.000	0.000	0.000	539.000	0.000	0.000	24.000	0.000	5.000	8.000	0.000	2.000	7.000	1.000	2.000	1.000	0.000	0.000	13	
0.000	79.000	1.000	0.000	0.000	0.000	63.000	0.000	0.000	0.000	0.000	4.000	781.000	0.000	2.000	3.000	2.000	53.000	3.000	0.000	0.000	12	
0.000	27.000	0.000	0.000	0.000	0.000	21.000	0.000	0.000	3.000	0.000	940.000	0.000	0.000	1.000	0.000	2.000	5.000	0.000	0.000	0.000	11	
0.000	47.000	0.000	0.000	0.000	0.000	48.000	0.000	0.000	6.000	0.000	886.000	0.000	0.000	2.000	0.000	2.000	1.000	1.000	1.000	0.000	10	
0.000	101.000	1.000	0.000	0.000	0.000	54.000	0.000	0.000	810.000	0.000	14.000	1.000	0.000	3.000	2.000	4.000	5.000	0.000	1.000	0.000	9	
0.000	194.000	0.000	0.000	0.000	0.000	70.000	0.000	0.000	706.000	0.000	8.000	1.000	0.000	3.000	1.000	1.000	5.000	1.000	0.000	0.000	8	
0.000	152.000	0.000	0.000	0.000	0.000	731.000	0.000	0.000	48.000	0.000	31.000	1.000	0.000	3.000	3.000	2.000	2.000	0.000	2.000	0.000	7	
0.000	84.000	0.000	0.000	0.000	0.000	898.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000	0.000	3.000	0.000	1.000	0.000	0.000	0.000	6	
0.000	167.000	0.000	0.000	0.000	0.000	785.000	0.000	0.000	4.000	0.000	0.000	1.000	0.000	0.000	4.000	1.000	0.000	1.000	0.000	0.000	5	
0.000	98.000	0.000	0.000	0.000	0.000	875.000	0.000	0.000	3.000	0.000	2.000	2.000	0.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	4	
0.000	96.000	0.000	0.000	0.000	0.000	871.000	0.000	0.000	0.000	0.000	10.000	1.000	0.000	1.000	5.000	0.000	0.000	1.000	0.000	0.000	3	
0.000	156.000	1.000	0.000	0.000	0.000	803.000	0.000	0.000	3.000	0.000	3.000	1.000	0.000	0.000	3.000	0.000	2.000	0.000	1.000	0.000	2	
0.000	407.000	1000	0.000	0.000	0.000	12.000	0.000	0.000	0.000	0.000	1.000	1.000	0.000	3.000	1.000	354.000	6.000	8.000	5.000	0.000	1	
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" means the noise that has not been classified into any class labels. By observing the HBDSCAN contingency matrix, we have 11 clusters. For example, cluster 9 has 2-7 classes combined.

Question 17:

```
{'SVD5 + K_MEAN + 10': 0.09679702072269666, 
'SVD5 + K_MEAN + 20': 0.13344166546423533, 
'SVD5 + K_MEAN + 50': 0.1089927051735795, 
'SVD20 + K_MEAN + 10': 0.08588229350837168, 
'SVD20 + K_MEAN + 20': 0.12297443619840195, 
'SVD20 + K_MEAN + 50': 0.15404690479353797,
```

```
'SVD200 + K MEAN + 10': 0.09918873421997113,
'SVD200 + K MEAN + 20': 0.0985557850547592,
'SVD200 + K MEAN + 50': 0.13006995979150218,
'SVD5 + AGG + 20': 0.11567593953990528,
'SVD20 + AGG + 20': 0.16830745559066143,
'SVD200 + AGG + 20': 0.12443854396797484,
'SVD5 + HDB + 100': -9.419666101039764e-06,
'SVD20 + HDB + 100': 0.0,
'SVD200 + HDB + 100': 0.0,
'SVD5 + HDB + 200': 0.0,
'SVD20 + HDB + 200': 0.0,
'SVD200 + HDB + 200': 0.0,
'NMF200 + K MEAN + 10': 0.004249134517982495,
'NMF200 + K MEAN + 20': 0.035679745688306376,
'NMF200 + K MEAN + 50': 0.04527096828501377,
'NMF5 + K MEAN + 10': 0.08986787288671771,
'NMF5 + K MEAN + 20': 0.1035619753598765,
'NMF5 + K MEAN + 50': 0.09109340339245127,
'NMF20 + K MEAN + 10': 0.03177639947680652,
'NMF20 + K MEAN + 20': 0.10406823198250942,
'NMF20 + K MEAN + 50': 0.13559778428965924,
'NMF5 + AGG + 50': 0.10533647441851773,
'NMF20 + AGG + 50': 0.10990489971472493,
'NMF200 + AGG + 50': 0.04625650654815621,
'NMF5 + HDB + 100': 0.034730587914443604,
'NMF20 + HDB + 100': 0.0,
'NMF200 + HDB + 100': 0.0,
'NMF5 + HDB + 200': 0.0,
'NMF20 + HDB + 200': 0.0,
'NMF200 + HDB + 200': 0.0,
'UMAP5 + K MEAN + 10': 0.36760645530133773,
'UMAP5 + K MEAN + 20': 0.4690091818375202,
'UMAP5 + K MEAN + 50': 0.402091823310506,
'UMAP20 + K MEAN + 10': 0.3307463748469388,
'UMAP20 + K MEAN + 20': 0.46765868857262133,
'UMAP20 + K MEAN + 50': 0.3696683148548335,
'UMAP200 + K MEAN + 10': 0.32794978430798466,
'UMAP200 + K MEAN + 20': 0.4199456856826519,
'UMAP200 + K MEAN + 50': 0.3986889321980805,
'UMAP200 + AGG + 20': 0.4440650914683766,
'UMAP5 + AGG + 20': 0.4207952176167666,
'UMAP20 + AGG + 20': 0.4366819550548756,
'UMAP5 + HDB + 100': 0.20593084183962104,
'UMAP20 + HDB + 100': 0.19523198997207222,
```

```
'UMAP200 + HDB + 100': 0.20379728827942695,

'UMAP5 + HDB + 200': 0.22302916559650812,

'UMAP20 + HDB + 200': 0.2220502699941576,

'UMAP200 + HDB + 200': 0.22108064519743173}

'None + K_MEAN + 10': 0.10635425868004754,

'None + K_MEAN + 20': 0.11499566821171768,

'None + K_MEAN + 50': 0.12343873501123943,

'None + HDB + 100': 0.0,

'None + HDB + 200': 0.0
```

Thus, from the result we get from above, the best adjusted rand index we get from the combination of UMAP20 + K_MEAN + 20': 0.46765868857262133, which is UMAP with 20 components with k mean and parameter with 20.

Question 19:

We know that VGG is a pre-trained model for image detection on different datasets. We also assume that it can extract its features from the other datasets. Since the basic structure of the VGG is a neural network, the gradient in the hidden layers can make sure we are able to capture some features of the image in the dataset. Furthermore, the image dataset is large, and we can make sure VGG has the ability to learn and have discriminative power to the custom dataset.

Question 20:

It first downloads the flower dataset and extracts its images. Then, it defines a class of FeatureExtractor, with _init_, and forward function in it. It extracts the feature layers and average pooling layer from the VGG, reduces the obtained feature map size into a one-dimensional vector, then passes it through the final fully-connected layer to get the final features.

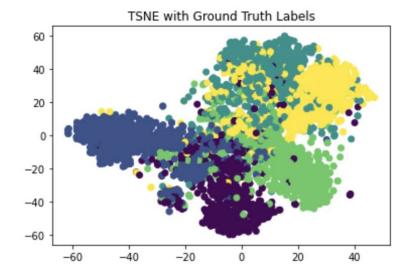
Question 21:

There are many different pixels in the original images. It has (250,320), (320,263). After VGG transformation, each picture has (224,224). In the end, it extracts 4096 features for each picitures.

Question 22:

Compared with the sparse TF-IDF, extracted features are dense.

Question 23:



The above plot is the TSNE visualization in 2-D. We can see that green, blue, and dark red has clear boundary line between each other, even though there are some spots landing into the other color regions. The yellow and dark green overlap in the region ([0-20], [20,40]). So they might have similarities in that clusters.

Question 24:

```
{'SVD50 + K MEAN + 5': 0.19146937613851173,
'SVD50 + AGG + 5': 0.2334745470751877,
'SVD50 + HDB + 5sample 5': 0.005498837928082272,
'SVD50 + HDB + 5sample 20': 0.0,
'SVD50 + HDB + 50sample 5': 0.0,
'SVD50 + HDB + 50sample 20': 0.0,
'SVD50 + HDB + 100sample 5': 0.0,
'SVD50 + HDB + 100sample 20': 0.0,
'UMAP50 + K MEAN + 5': 0.4676822988641824,
'UMAP50 + AGG + 5': 0.48631519971009973,
'UMAP50 + HDB + 5sample 5': 0.09411014415380671,
'UMAP50 + HDB + 5sample 10': 0.09494009559863244,
'UMAP50 + HDB + 10sample 5': 0.09494009559863244,
'UMAP50 + HDB + 10sample 10': 0.09483164455701254,
'AUTO50 + K MEAN + 5': 0.1988387986337919,
'AUTO50 + AGG + 5': 0.14208845180661883,
'AUTO50 + HDB + 5sample 5': 0.020317450403088272,
'AUTO50 + HDB + 5sample 10': 0.023403788945569347,
'AUTO50 + HDB + 10sample 5': 0.001313600583448493,
'AUTO50 + HDB + 10sample 10': 0.0.
'NONE50 + K MEAN + 5': 0.19447072328729068,
'NONE + K MEAN + 5': 0.19447072328729068,
'NONE + AGG + 5': 0.18855278251971858,
```

```
'NONE + HDB + 5sample 5': 0.006705947729476718,
'NONE + HDB + 5sample 10': 0.0,
'NONE + HDB + 10sample 5': 0.006705947729476718,
'NONE + HDB + 10sample 10': 0.0}
```

Thus, from the result we get above, the best-adjusted rand index we get from the combination of 'UMAP50 + AGG + 5': 0.48631519971009973, which is UMAP with 50 components and AGG with parameter 5.

Question 25:

```
class MLP(torch.nn.Module):
   def __init__(self, num_features):
       super().__init__()
       self.model = nn.Sequential(
           nn.Linear(num_features, 1280),
           nn.ReLU(True),
           nn.Linear(1280, 640).
           nn.ReLU(True),
           nn.Linear(640, 5),
           nn.LogSoftmax(dim=1)
        self.cuda()
    def forward(self, X):
       return self.model(X)
    def train(self, X, y):
       X = torch.tensor(X, dtype=torch.float32, device='cuda')
       y = torch.tensor(y, dtype=torch.int64, device='cuda')
        self.model.train()
       criterion = nn.NLLLoss()
       optimizer = torch.optim.Adam(self.parameters(), lr=1e-3, weight_decay=1e-5)
        dataset = TensorDataset(X, y)
       dataloader = DataLoader(dataset, batch_size=128, shuffle=True)
        for epoch in tqdm(range(100)):
           for (X_, y_) in dataloader:
               *******************************
               optimizer.zero_grad()
               criterion(self(X_), y_).backward()
               optimizer.step()
               ******************************
    def eval(self, X_test, y_test):
        dataset = TensorDataset(torch.tensor(X_test, dtype=torch.float32, device='cuda'), torch.tensor(y_test, dtype=torch.int64, device='cuda'))
       dataloader = DataLoader(dataset, batch_size=64, shuffle=True)
        count = 0
        for (X,y) in dataloader :
         output = self(X)
         nums += v.shape[0]
         count += (torch.max(output,1)[1]==y).sum().item()
       return count/nums
```

As the above result showed, the test accuracy of the MLP classifier on the original VGG features was 0.8923; the test accuracy of the MLP classifier on the reduced VGG features was 0.8896. Since the dimension is reduced, the performance does lower a little, but it is not significant. We can see, the accuracy of MLP is way better than what we got from Q24. This means that MLP improves the performance of classification tasks.