EE 219 Project 4: Clustering

Winter 2017

March 8th, 2017

Arunav Singh (304760844)
John Moon (204774912)
Jiawen Yu (204753330)

1) Finding a good representation of the data is fundamental to the task of clustering. Following the steps in project 2, transform the documents into TF-IDF vectors.

See 'part_1_2.py' for the extraction of the TF-IDF vectors.

2) Apply K-means clustering with k=2. Compare the clustering results with the known class labels. Inspect the confusion matrix to evaluate how well your clusters match the ground truth labels. Is there a permutation of the rows that makes confusion matrix look almost diagonal?

Applying the K-means clustering algorithm with k = 2 would occasionally yield good results but at the same time would yield poor results as shown in *Figure 1* and *Figure 2* below. This is due to the algorithm starting to cluster at different points. With a set amount of iterations the algorithm will not converge to the same cluster location every time.

	Predicted: CT	Predicted: RA			
Actual: CT	3750	153		V-measure	0.811955
Actual: RA	77	3902		Adjusted Mutual In	0.81172
				Completeness	0.812173
				Adjust Rand-Index	0.88667
				Homogeneity	0.811737
	CT	RA	Average		
Precision	0.979879801	0.962268804	0.971074303		
Recall	0.960799385	0.980648404	0.970723895		
Accuracy	0.970819589		0.970872595		

Figure 1: K-Means Clustering Algorithm k = 2, First Run

	Predicted: CT	Predicted: RA			
Actual: CT	1640	2339		V-measure	0.275446
Actual: RA	5	3898		Adjusted M	0.239403
				Completen	0.324137
				Adjust Ran	0.164318
				Homogene	0.239473
	CT	RA	Average		
Precision	0.996960486	0.624979958	0.810970222		
Recall	0.41216386	0.998718934	0.705441397		
Accuracy	0.70261355		0.739675056		

Figure 2: K-Means Clustering Algorithm k = 2, Second Run Adjusted

We noticed that sometimes the resulting confusion matrix would appear like it does in *Figure 3.* This is because the K-Means Algorithm's actual label (0 or 1) would differ from our ground truth labeling, even though it may be clustering well.

If we were to use this labeling convention we would get the following accuracy result as shown in *Figure 3.* We realized however that the K-means clustering algorithm for k = 2 can only have an accuracy above 50% since if the confusion matrix yields something lower, it is because it has a different labeling convention.

	Predicted: CT	Predicted: RA	
Actual: CT	5	3898	
Actual: RA	1640	2339	
	CT	RA	Average
Precision	0.003039514	0.375020042	0.189029778
Recall	0.001281066	0.58783614	0.294558603
Accuracy	0.29738645		0.260324944

Figure 3: K-Means Clustering Algorithm k = 2, Second Run, Non Adjusted

In the case of *Figure 3*, we simply changed the order of the rows to obtain the true, adjusted confusion matrix and its accuracy, precision, and recall values. Fortunately, the purity metrics take this labeling convention variation into account thus making it a good metric to rely on going forward assuming the dimensionality reduction method we use is stable. The purity metrics in this case can be seen to vary heavily across successive runs thus requiring extra work to better the performance of our classifier.

3. Now, use the following two methods for reducing the dimension of the data by sweeping over the dimension parameter in each.

We attempted to apply many different dimensionality reduction techniques including: LSI, NMF, NMF with logarithmic non-linear transformation, and NMF with normalization, before passing through our k-means algorithm.

In order to see how each of these stacked up against each other we swept across the dimensionality parameter for each of the methods and observed their respective purity metrics and confusion matrices. We looked at the adjusted-rand index in particular, as it is the measure that gauges the accuracy of the clustering. We observed that for dimensions > 20, our performance measures decreased and hence swept the dimension parameter from one to twenty.

Figure 4 shows the performance of our k-means algorithm after performing LSI across various dimensions.

Dimension	Adjusted Rand	Homogeneity	Completeness	V-measure	Adjusted Mutual Information Score
1	0.001118751	0.00105031	0.001085047	0.001067	0.000958854
2	0.181409823	0.2510509	0.330339931	0.285289	0.25098233
3	0.140550724	0.21917551	0.309589464	0.256653	0.219104014
4	0.008246867	0.04270684	0.17346106	0.068539	0.042619075
5	0.136019758	0.2150244	0.306464098	0.252728	0.214952521
6	0.109877985	0.10059787	0.113696056	0.106747	0.100515527
7	0.136207067	0.21519684	0.306593947	0.252891	0.215124977
8	0.138087242	0.21692373	0.307894214	0.254525	0.216852031
9	0.850719324	0.76947078	0.77073929	0.770105	0.76944967
10	0.136582072	0.21554185	0.306853741	0.253217	0.215470024
11	0.137333608	0.2153844	0.305938486	0.252797	0.215312555
12	0.136582072	0.21554185	0.306853741	0.253217	0.215470024
13	0.871916055	0.79115481	0.791433519	0.791294	0.79113569
14	0.00687195	0.03679484	0.159711069	0.05981	0.036706521
15	0.136582072	0.21554185	0.306853741	0.253217	0.215470024
16	0.137333627	0.21623243	0.307373717	0.253871	0.216160664
17	0.863877751	0.78401885	0.784876521	0.784447	0.783999073
18	0.137898645	0.21675084	0.307764041	0.254361	0.216679121
19	0.137145526	0.21521179	0.30580802	0.252633	0.215139932
20	0.137145526	0.21521179	0.30580802	0.252633	0.215139932

Figure 4: Purity Metrics for K-Means algorithm k = 2 with LSI across various dimensions, first run

We see that at certain dimensions we achieve high performance; however this performance varied across successive runs, as seen by Figure 8. For some dimensions, the LSI would prove effective one run but poor the next.

We then repeated the process for NMF. *Figure 9* and *10* show the results across two runs of NMF. Here we see that across two runs, the performance measures varied heavily across all dimension parameters. In attempt to stabilize and improve the result, we attempted to apply a non-linear transformation (logarithmic) after reducing the dimension, before feeding it into the K-means algorithm.

We noticed a substantial improvement in relative performance across most of the dimensions with the average performance across all dimensions higher. *Figure 11* and *12* shows the results of two runs of NMF with a logarithmic non-linear transformation.

In order to get more stability, we swept across the dimension value again but with various values for the alpha regularization parameter. With alpha set to 0.0018 we found the best results. We obtained the following results across two runs are showed in *Figure 5* and *6*.

We can see that at a dimension of 18, we achieve the best performance across both runs. We then ran the algorithm several times at this dimension to see how consistent the results were. The result is in *Figure 7*. We can say with high

confidence that these settings will yield a relatively high performance across each run.

Dimension	Adjusted Rand	Homogeneity	Completeness	V-measure	Adjusted Mutual Information Score
1	0.004095278	0.00301128	0.003152291	0.00308	0.00292
2	0.498966258	0.44845599	0.46954785	0.45876	0.448405498
3	0.325000072	0.33643868	0.378689906	0.356316	0.336377926
4	0.313527276	0.30401154	0.335990243	0.319202	0.30394782
5	0.717791492	0.61071976	0.610780928	0.61075	0.610684122
6	0.307305823	0.31143066	0.350187326	0.329674	0.311367616
7	0.746891895	0.64216068	0.642285419	0.642223	0.64212792
8	0.701968935	0.59723917	0.598484627	0.597861	0.597202294
9	0.010699368	0.00859838	0.00892836	0.00876	0.008507611
10	0.710072028	0.60660037	0.608080652	0.60734	0.606564351
11	0.696027981	0.5899345	0.590834442	0.590384	0.589896957
12	0.305619872	0.31337028	0.353928306	0.332417	0.313307421
13	0.735965929	0.6310225	0.631445857	0.631234	0.630988723
14	0.73945349	0.63604342	0.636855839	0.636449	0.636010098
15	0.236024356	0.26572499	0.318072571	0.289552	0.265657766
16	0.734660248	0.63419704	0.63595768	0.635076	0.634163546
17	0.224335315	0.26571585	0.325228128	0.292475	0.26564862
18	0.750844957	0.64761074	0.648140697	0.647876	0.647578482
19	0.208515429	0.2632603	0.33143007	0.293438	0.263192843
20	0.210839901	0.26360196	0.330459361	0.293268	0.263534534

Figure 5: Purity Metrics for K-Means algorithm k = 2 with NMF with Log NLT and alpha regularization, first run

Dimension	Adjusted Rand	Homogeneity	Completeness	V-measure	Adjusted Mutual Information Score
1	0.00422786	0.00309712	0.003218236	0.003157	0.003005851
2	0.505801738	0.45149385	0.471154664	0.461115	0.451443633
3	0.760994594	0.65973682	0.659874292	0.659806	0.659705667
4	0.310974452	0.3023657	0.33473395	0.317728	0.302301828
5	0.291201531	0.29661877	0.335820921	0.315005	0.296554373
6	0.755249465	0.65121161	0.651211608	0.651212	0.651179676
7	0.745576529	0.64078836	0.640945559	0.640867	0.640755473
8	0.703670984	0.59897119	0.600196328	0.599583	0.598934478
9	0.708789493	0.60397476	0.605070096	0.604522	0.603938504
10	0.241481183	0.2710417	0.323290973	0.29487	0.270974963
11	0.233074248	0.26484508	0.318446312	0.289183	0.264777773
12	0.722098145	0.61694971	0.617661404	0.617305	0.616914644
13	0.708362236	0.60268825	0.603516371	0.603102	0.602651875
14	0.715643028	0.61170258	0.612929545	0.612315	0.611667034
15	0.727283129	0.62537706	0.626930406	0.626153	0.625342766
16	0.021759836	0.0157829	0.015807798	0.015795	0.015692793
17	0.738580833	0.63582092	0.636846454	0.636333	0.635787582
18	0.750844959	0.64795333	0.648587808	0.64827	0.6479211
19	0.21718041	0.26821006	0.333717531	0.297399	0.268143057
20	0.734225254	0.63081564	0.631767091	0.631291	0.630781837

Figure 6: Purity Metrics for K-Means algorithm k = 2 with NMF with Log NLT and alpha regularization, second run

Dimension = 18					
Iteration	Adjusted Rand	Homogeneity	Completeness	V-measure	Adjusted Mutual Information Score
1	0.746891905	0.64367565	0.644333189	0.644004	0.643643033
2	0.746891895	0.64216068	0.642285419	0.642223	0.64212792
3	0.759224558	0.65825315	0.659168447	0.65871	0.65822186
4	0.752164961	0.64932596	0.649934851	0.64963	0.64929386
5	0.214115763	0.26285042	0.327132644	0.29149	0.262782931
6	0.236271504	0.28095491	0.34199156	0.308483	0.280889078
7	0.739889999	0.63440334	0.634425884	0.634415	0.63436987
8	0.74557655	0.64443111	0.645711061	0.64507	0.644398562
9	0.223854725	0.26617868	0.326268925	0.293176	0.266111488
10	0.725120441	0.61851046	0.618555885	0.618533	0.618475534
11	0.744700275	0.64181531	0.642635105	0.642225	0.641782521
12	0.745138333	0.64055571	0.640816385	0.640686	0.6405228
13	0.232584753	0.27710155	0.338186088	0.304612	0.277035363
14	0.752164964	0.6497705	0.650510275	0.65014	0.649738434
15	0.730750064	0.62709712	0.628064606	0.62758	0.627062984
16	0.752605217	0.64930278	0.649753744	0.649528	0.649270672

Figure 7: Purity Metrics of Dimension = 18 for K-Means algorithm k = 2 with NMF with Log NLT and alpha regularization

Dimension	Adjusted Rand	Homogeneity	Completeness	V-measure	Adjusted Mutual Information Score
1	0.00122835	0.00112824	0.001162252	0.001145	0.001036789
2	0.18314373	0.25251543	0.33146003	0.286652	0.252446991
3	0.140360453	0.21900202	0.309458861	0.256489	0.21893051
4	0.790046763	0.71333921	0.718063867	0.715694	0.713312962
5	0.136019758	0.2150244	0.306464098	0.252728	0.214952521
6	0.100331429	0.08986992	0.100454677	0.094868	0.08978659
7	0.861520627	0.78128023	0.782193313	0.781737	0.781260203
8	0.137145545	0.21605972	0.307243674	0.253707	0.215987936
9	0.850719324	0.76912036	0.770333782	0.769727	0.769099225
10	0.135832578	0.214852	0.306334282	0.252564	0.21478011
11	0.850251249	0.76968494	0.771080798	0.770382	0.769663854
12	0.006956947	0.03778196	0.163995755	0.061415	0.037693733
13	0.007042396	0.03733392	0.160399255	0.06057	0.03724566
14	0.007433371	0.03734859	0.155362087	0.06022	0.037260334
15	0.1388429	0.2167669	0.306983374	0.254105	0.216695183
16	0.007566035	0.03775058	0.155900383	0.060783	0.037662363
17	1.17131E-05	1.552E-05	4.2179E-05	2.27E-05	-7.60941E-05
18	0.872864188	0.7947655	0.795446834	0.795106	0.794746708
19	0.877612566	0.80011208	0.800639536	0.800376	0.800093783
20	0.139032159	0.21778889	0.308545572	0.255343	0.217717264

Figure 8: Purity Metrics for K-Means algorithm k = 2 with LSI across various dimensions, second run

Dimension	Adjusted Rand	Homogeneity	Completeness	V-measure	Adjusted Mutual Information Score
1	0.001154797	0.00107806	0.001113354	0.001095	0.000986604
2	0.165555375	0.2405487	0.324949624	0.276451	0.240479159
3	0.185757182	0.1921381	0.226823057	0.208045	0.192064139
4	0.137333627	0.21623243	0.307373717	0.253871	0.216160664
5	0.103583454	0.1838766	0.282943846	0.222898	0.183801866
6	0.653916363	0.57080288	0.576593543	0.573684	0.570763588
7	0.793207842	0.6958049	0.695863242	0.695834	0.695777051
8	0.137710884	0.20982219	0.306978264	0.249268	0.20974984
9	0.860578676	0.77878791	0.779497022	0.779142	0.778767657
10	0.071019417	0.13660422	0.243774829	0.175092	0.136525156
11	0.183354122	0.14959871	0.156573877	0.153007	0.149520851
12	0.120921996	0.19775544	0.290262707	0.235241	0.19768198
13	0.134712127	0.21067332	0.300149225	0.247575	0.210601043
14	0.039672866	0.06402939	0.114051366	0.082015	0.063943678
15	0.116201957	0.1932275	0.286787965	0.23089	0.193153629
16	0.050555274	0.10689595	0.210612086	0.141814	0.10681415
17	0.135645545	0.21558858	0.307734271	0.253549	0.215516751
18	0.033369878	0.0928072	0.217811867	0.130156	0.092724097
19	0.036600729	0.09412572	0.211635371	0.1303	0.094042745
20	0.073192128	0.07590694	0.093088334	0.083624	0.07582233

Figure 9: Purity Metrics for K-Means algorithm k = 2 with NMF across various dimensions, first run

Dimension	Adjusted Rand	Homogeneity	Completeness	V-measure	Adjusted Mutual Information Score
1	0.001173024	0.00109285	0.001128721	0.00111	0.001001397
2	0.16700446	0.24098905	0.324587008	0.27661	0.24091955
3	0.18553847	0.1923178	0.227274966	0.20834	0.192243848
4	0.82516636	0.73534546	0.736273104	0.735809	0.735321231
5	0.103746936	0.18404106	0.283068586	0.223058	0.183966348
6	0.005393076	0.03346746	0.162901792	0.055527	0.033378812
7	0.111745553	0.19198412	0.289084004	0.230735	0.191910131
8	0.111915346	0.19215062	0.289209927	0.230895	0.19207665
9	0.111236946	0.19148486	0.288706384	0.230254	0.191410829
10	0.129728606	0.2023403	0.301321401	0.242105	0.202267257
11	0.006454787	0.03427702	0.15112946	0.05588	0.034188471
12	0.046856557	0.11019587	0.228024821	0.148586	0.110114375
13	0.121275383	0.1988179	0.291822167	0.236505	0.198744534
14	0.1620621	0.2259835	0.303390079	0.259027	0.225912634
15	0.105554342	0.17250182	0.271172176	0.210865	0.172426043
16	0.12863289	0.2036462	0.293378944	0.240412	0.203573283
17	0.134712238	0.21573078	0.308759417	0.253995	0.215658972
18	0.142652264	0.22426949	0.316201079	0.262417	0.224198465
19	6.76133E-05	6.6697E-05	0.000221002	0.000102	-2.49387E-05
20	0.143613165	0.21076101	0.303158278	0.248654	0.210688739

Figure 10: Purity Metrics for K-Means algorithm k = 2 with NMF across various dimensions, second run

Dimension	Adjusted Rand	Homogeneity	Completeness	V-measure	Adjusted Mutual Information Score
1	0.00419483	0.00308775	0.00323709	0.003161	0.002996477
2	0.498607782	0.44817997	0.469304506	0.458499	0.448129445
3	0.761880391	0.66065817	0.660786507	0.660722	0.660627108
4	0.720804823	0.61904738	0.620794687	0.61992	0.619012504
5	0.290653951	0.29650331	0.335924398	0.314985	0.2964389
6	0.755690625	0.65172484	0.651706589	0.651716	0.651674704
7	0.01963652	0.01433603	0.014337232	0.014337	0.014245793
8	0.312391534	0.30736867	0.341597922	0.323581	0.307305255
9	0.704096847	0.60360779	0.606069392	0.604836	0.603571504
10	0.71564301	0.60920729	0.609623178	0.609415	0.609171512
11	0.717791501	0.61163757	0.61210712	0.611872	0.611602019
12	0.274200324	0.29621606	0.344109519	0.318372	0.296151621
13	0.717791519	0.61431029	0.61562861	0.614969	0.614274981
14	0.235039034	0.2685947	0.323248047	0.293398	0.268527739
15	0.741637397	0.63854636	0.639392594	0.638969	0.638513268
16	0.000667516	0.00066214	0.000676786	0.000669	0.000570646
17	0.234055966	0.27826336	0.33911791	0.305691	0.278197279
18	0.735530588	0.63472725	0.636373981	0.63555	0.634693811
19	0.729015548	0.62327471	0.623609369	0.623442	0.623240224
20	0.000442176	0.00042852	0.000429278	0.000429	0.000337006

Figure 11: Purity Metrics for K-Means algorithm k = 2 with NMF with Log Non-linear Trans, first run

Dimension	Adjusted Rand	Homogeneity	Completeness	V-measure	Adjusted Mutual Information Score
1	0.004161483	0.00306071	0.003204662	0.003131	0.002969438
2	0.497533132	0.44763985	0.468967851	0.458056	0.447589276
3	0.314096752	0.33142892	0.376551852	0.352552	0.331367708
4	0.014648607	0.01131835	0.011564252	0.01144	0.01122784
5	0.290927662	0.29611828	0.335195969	0.314448	0.296053839
6	0.755249465	0.6512089	0.651213617	0.651211	0.651176966
7	0.027079395	0.0197862	0.019793525	0.01979	0.019696463
8	0.301982404	0.30153328	0.337875904	0.318672	0.301469332
9	0.710927679	0.60607752	0.607120544	0.606599	0.606041456
10	0.716072458	0.61066507	0.611425844	0.611045	0.61062943
11	0.697298909	0.59052759	0.591173598	0.59085	0.590490105
12	0.707508128	0.60286766	0.604041531	0.603454	0.602831303
13	0.244484021	0.27749956	0.331724462	0.302199	0.277433414
14	0.735965934	0.63184516	0.632536361	0.632191	0.631811451
15	0.747769459	0.64440807	0.644994822	0.644701	0.644375512
16	0.020211541	0.0146539	0.01469192	0.014673	0.014563693
17	0.735095338	0.63162101	0.63253081	0.632076	0.631587287
18	0.229656263	0.27572599	0.337807421	0.303626	0.275659675
19	0.750844951	0.64671244	0.646934578	0.646823	0.646680092
20	0.735530584	0.63391281	0.635338476	0.634625	0.633879297

Figure 12: Purity Metrics for K-Means algorithm k = 2 with NMF with Log Non-linear Trans, second run

To get a visual sense on the NMF embedding of the data, try applying NMF to the data matrix with ambient parameter 2 and plot the resulting points to choose the appropriate non-linear transformation. Can you justify why logarithm is a good candidate for your TFxIDF data?

Quantitatively, we saw before that performance measures improve whilst adding a non-linear transform such as the logarithmic non-linear transform. We then plotted the clustering algorithm in the two-dimensional space to visualize its performance. Figure 13 shows the clustering algorithm using NMF at dimension = 2 without a logarithmic transformation of the features and *Figure 14* show the same but with a logarithmic transformation of the features. We see here that the points are more spread out than being stacked on top of each other, essentially achieving better separation to begin with. This is because of the nature of the logarithm function. It took the large concentration of points around the origin and mapped it to a larger space, as shown below. Figure 15 shows the ground truth of our data in two dimensions. Normalization is not a good candidate as it does not scatter the points as much as log.

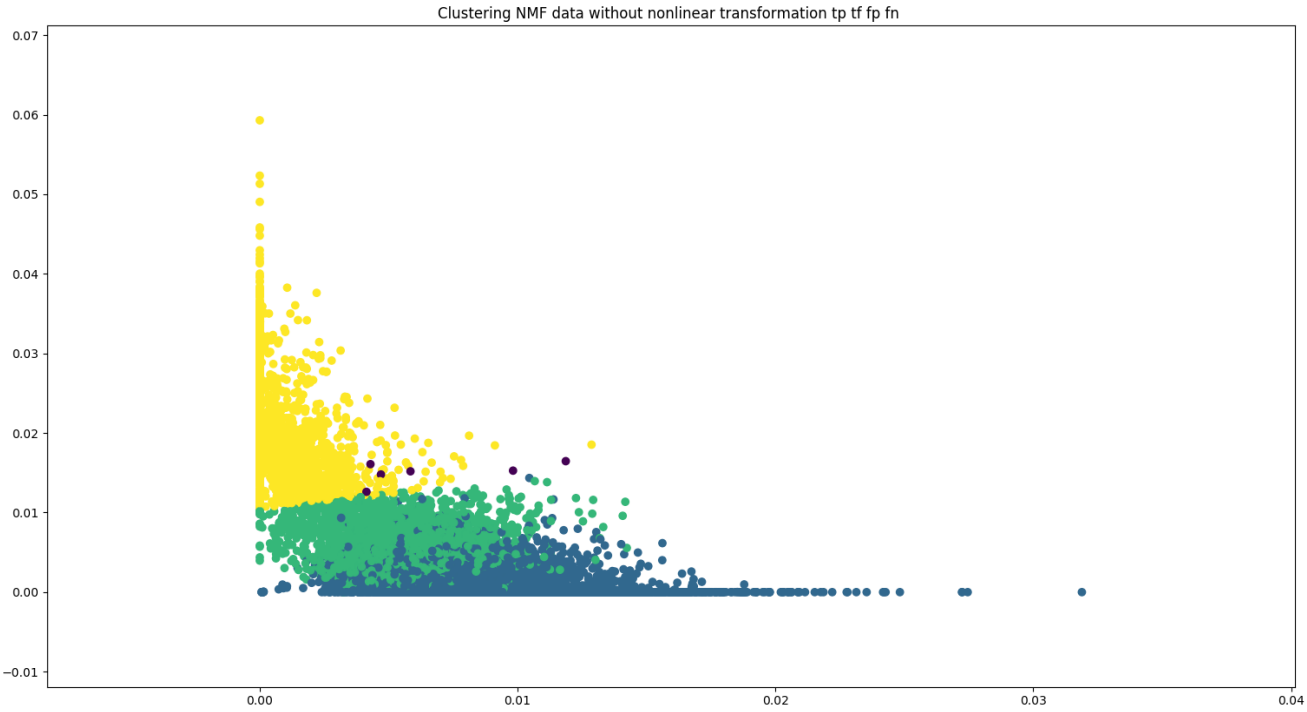

Figure 13: Plot of True Positives (yellow), True Negatives (dark blue), False Positives (light blue) and False Negatives (purple) for K-Means algorithm k = 2 with NMF without Log Non-linear Trans

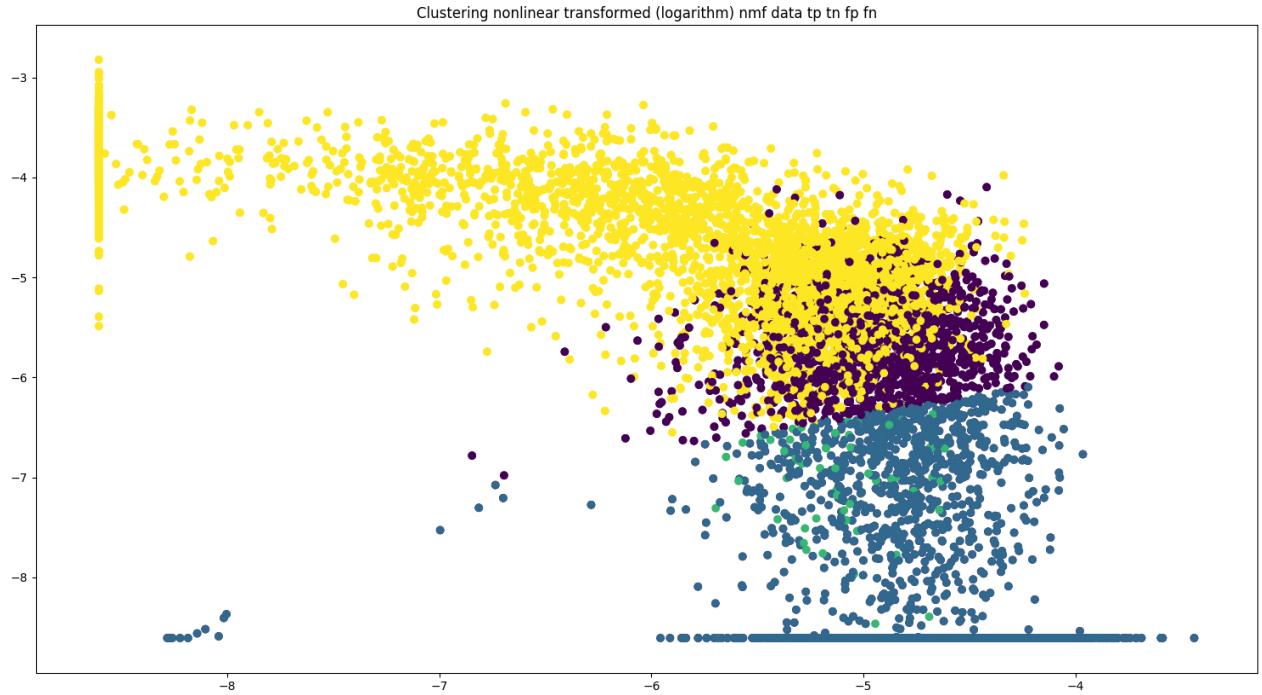

Figure 14: Plot of True Positives (yellow), True Negatives (dark blue), False Positives (light blue) and False Negatives (purple) for K-Means algorithm k = 2 with NMF with Log Non-linear Trans

Report the measures of purity introduced in part 2 for the best final data representation you use.

The best final data representation as described above was achieved by reducing the dimensionality down to 18, with NMF with a regularization parameter alpha, and then applying a logarithmic non-linear transform. The achieved purity metrics were as follows:

Homogeneity = 0.6479

Completeness = 0.6485

V-Measure = 0.6482

Adjusted Rand = 0.7508

Adjusted Mutual Information = 0.6479

4) Visualize the performance of your clustering by projecting final data vectors onto 2 dimensions and color-coding the classes. Can you justify why a non-linear transform is useful?

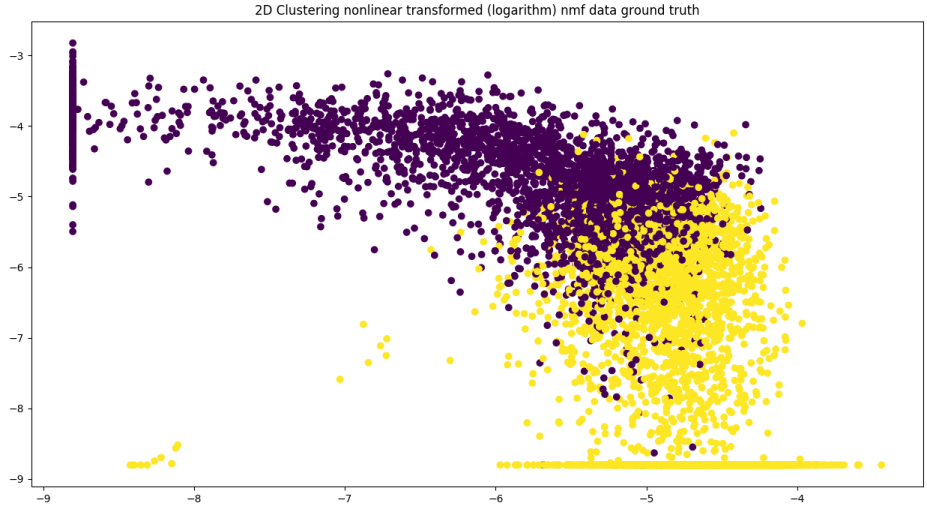

Figure 15: Ground truth of data in 2D

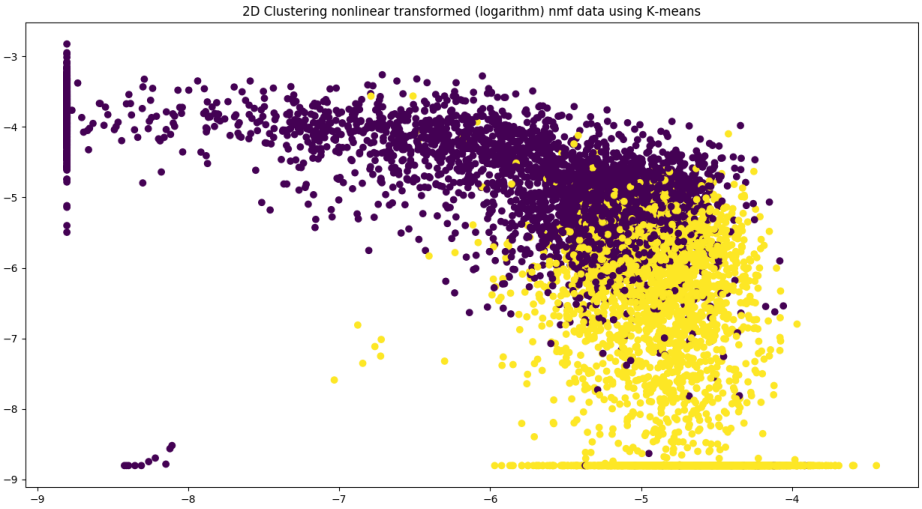

Figure 16: Projection of K-means clustering algorithm of final data representation onto 2D

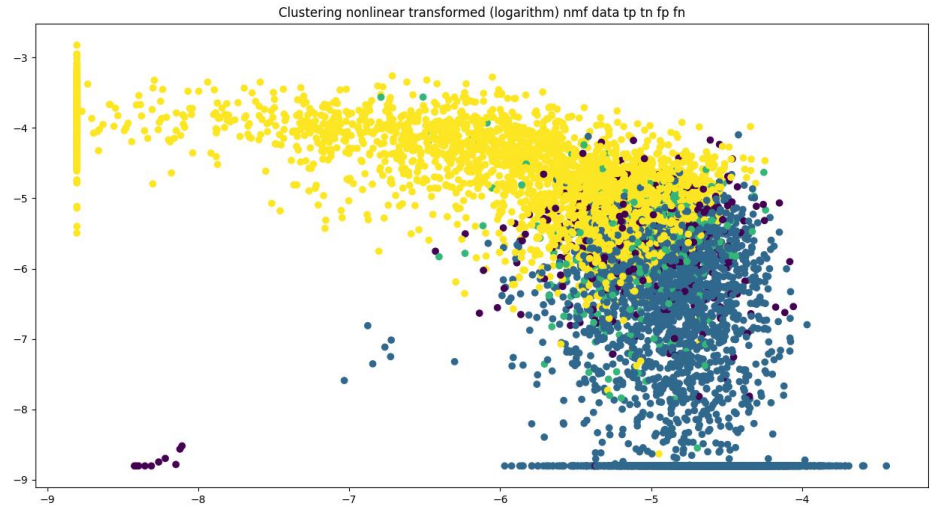

Figure 17: Projection of K-means cluster algorithm of final data representation on 2D with Tp, Fp, Fn, Tn

Non-linear transforms are useful because as seen in the previous sections, they improve the stability of the results. In the case of the logarithm, it was able to map a densely populated region into a bigger space thus aiding the K-means algorithm when it came time to cluster. Figure 15 shows the ground truth of the data in 2D and *Figure 16* and *17* shows the result of our K-means algorithm labeling in a higher dimension and then projecting in a 2D space.

5) In this part we want to examine how purely we can retrieve all the 20 original sub-class labels with clustering.

For this part we ran our K-means algorithm with k = 20 as our goal was not retrieve all 20 original sub-class labels. We applied the same methods as before (LSI, NMF, NMF with Log NLT) and found that NMF with a logarithmic nonlinear transform was yielding relatively strong and objectively stable results. The results from successive runs can be seen in Figures 18 and 19. When we tried to project our results onto two dimensions, it was too difficult to visually see the performance on k = 20 clusters. The same held for part 6) with k = 6 clusters.

Dimension	Adjusted Rand	Homogeneity	Completeness	V-measure	Adjusted Mutual Information Score
1	0.009129226	0.03947832	0.043791116	0.041523	0.036328462
2	0.066098238	0.20204808	0.213458782	0.207597	0.199470451
3	0.089193946	0.24527004	0.258453595	0.251689	0.24282984
4	0.134297883	0.30590426	0.313048809	0.309435	0.30366336
5	0.127725414	0.30641753	0.314176417	0.310248	0.304177971
6	0.140574818	0.33310167	0.341894365	0.337441	0.330948085
7	0.144167639	0.328254	0.333958688	0.331082	0.326085614
8	0.181331494	0.36564006	0.368015222	0.366824	0.363593004
9	0.215292792	0.39244948	0.393927279	0.393187	0.390489092
10	0.200604486	0.40096131	0.410038595	0.405449	0.399026516
11	0.223309885	0.3962344	0.405146638	0.400641	0.394284809
12	0.211931071	0.37889516	0.382732847	0.380804	0.376890689
13	0.241001029	0.40202417	0.411178551	0.40655	0.400093128
14	0.219503509	0.39883905	0.408275011	0.403502	0.396897353
15	0.255422076	0.4212489	0.429463267	0.425316	0.419380203
16	0.249534389	0.4009347	0.410501379	0.405662	0.399000329
17	0.236873809	0.40707798	0.414450792	0.410731	0.405163452
18	0.240784016	0.41809168	0.429360454	0.423651	0.416212256
19	0.269503284	0.43763498	0.442220149	0.439916	0.435820037
20	0.265474993	0.43848319	0.440816694	0.439647	0.436671262

Figure 18: K-means algorithm k = 20 with NMF with Log NLT, run one

Dimension	Adjusted Rand	Homogeneity	Completeness	V-measure	Adjusted Mutual Information Score
1	0.009029833	0.03988047	0.043577976	0.041647	0.036743732
2	0.066246143	0.20187574	0.211653064	0.206649	0.199276591
3	0.087493417	0.24418999	0.265758835	0.254518	0.24174324
4	0.129668637	0.30700525	0.317192324	0.312016	0.30476733
5	0.127672474	0.30805605	0.31545507	0.311712	0.305821785
6	0.142414246	0.33118706	0.337063927	0.3341	0.329028001
7	0.152221397	0.34341593	0.348576958	0.345977	0.34129674
8	0.183974805	0.37025432	0.373154258	0.371699	0.368222092
9	0.205777636	0.38563516	0.388473194	0.387049	0.383652454
10	0.215232289	0.4130517	0.420944121	0.416961	0.411156151
11	0.216175649	0.40305414	0.409026097	0.406018	0.401126971
12	0.231174519	0.40749132	0.410171582	0.408827	0.405579314
13	0.230994477	0.41833518	0.42803241	0.423128	0.41645677
14	0.224579851	0.40634486	0.414807509	0.410533	0.404427931
15	0.223840569	0.40176554	0.404524698	0.40314	0.399835007
16	0.245772404	0.41692207	0.427319257	0.422057	0.41503889
17	0.240916712	0.39481595	0.403096876	0.398913	0.39286197
18	0.256258181	0.43508908	0.442155805	0.438594	0.433265274
19	0.255345868	0.43390527	0.43787116	0.435879	0.432078346
20	0.253677031	0.41415436	0.419967783	0.417041	0.412263209

Figure 19: K-means algorithm k = 20 with NMF with Log NLT, run two

We can see from the above figures that our performance tends to peak around a dimension values of 19. As a result our final representation of these 20 subgroups would involve reducing the dimensionality down to 19 and performing a

logarithmic nonlinear transformation to the data before applying our K-means clustering algorithm with k = 20.

From above we can see we obtained the following purity metrics:

Homogeneity = 0.4376

Completeness = 0.4422

V-Measure = 0.4399

Adjusted Rand = 0.2695

Adjusted Mutual Information = 0.4358

6) Evaluate the performance of your clustering in retrieving the topic-wise classes. Note that again, you need to find a proper representation of your data through dimensionality reduction and feature transformation.

For this part, our target was classifying all the newsgroups into six clusters and therefore set k = 6 in our K-means algorithm. Right away we tried the same approach by using NMF with a logarithmic NLT. *Figure 20* and *21* show successive runs of this approach.

Dimension	Adjusted Rand	Homogeneity	Completeness	V-measure	Adjusted Mutual Information Score
1	0.01163845	0.03428638	0.062807044	0.044358	0.033469852
2	0.075316742	0.17251618	0.298832883	0.218749	0.171816975
3	0.088570844	0.18904371	0.340384057	0.243083	0.188358096
4	0.125475515	0.23808243	0.405720887	0.300076	0.237438752
5	0.166069244	0.27577671	0.476982751	0.349489	0.275164853
6	0.161889746	0.2816389	0.483115124	0.355837	0.281032011
7	0.161622123	0.27090622	0.460956166	0.341255	0.270290276
8	0.180459084	0.27836138	0.469532867	0.349514	0.277751763
9	0.148835575	0.26061291	0.455370823	0.331503	0.259988185
10	0.17937774	0.29090025	0.494195035	0.366227	0.29030121
11	0.176006034	0.27955288	0.474228541	0.351752	0.278944261
12	0.174575736	0.28398789	0.48341702	0.357789	0.283383004
13	0.192312862	0.29485289	0.506311486	0.372676	0.294257144
14	0.140883061	0.2418352	0.420469083	0.307062	0.241194627
15	0.164456051	0.27229562	0.461627743	0.34254	0.271680864
16	0.192016327	0.29576815	0.509705523	0.374325	0.295173175
17	0.149069668	0.25105435	0.424958111	0.315638	0.250421656
18	0.186597028	0.29431566	0.49637031	0.369526	0.293719526
19	0.177548009	0.28641282	0.49259025	0.362217	0.285809969
20	0.164444724	0.27320074	0.463855425	0.34387	0.272586746

Figure 20: K-means algorithm k = 6 with NMF with Log NLT, run one

Dimension	Adjusted Rand	Homogeneity	Completeness	V-measure	Adjusted Mutual Information Score
1	0.011613015	0.0342626	0.062765198	0.044327	0.033446057
2	0.074088904	0.17307635	0.315397507	0.223504	0.17237745
3	0.081191482	0.17836965	0.30935264	0.226273	0.177675439
4	0.135712291	0.23100749	0.392142898	0.290742	0.230357862
5	0.145239687	0.25515732	0.438377405	0.322566	0.254528066
6	0.161892934	0.28215038	0.484247026	0.356553	0.281543912
7	0.135876922	0.24484148	0.419043788	0.309087	0.244203501
8	0.166743317	0.27807143	0.476419428	0.351174	0.277461526
9	0.153457606	0.26338428	0.451455996	0.33268	0.26276197
10	0.203341196	0.30849646	0.522571253	0.387962	0.307912294
11	0.141939504	0.24699763	0.418431412	0.310631	0.246361507
12	0.158009026	0.26399072	0.448117352	0.33225	0.263368951
13	0.158730913	0.26255467	0.44704115	0.330816	0.261931678
14	0.174206187	0.28323297	0.481347079	0.356623	0.282627449
15	0.193954874	0.30731963	0.524088562	0.387446	0.306734446
16	0.195645197	0.29993668	0.51775764	0.379835	0.29934523
17	0.15889612	0.24685983	0.422776813	0.311711	0.246223543
18	0.168250923	0.2776934	0.470876765	0.349358	0.277083212
19	0.123674326	0.22334684	0.385303325	0.282777	0.222690666
20	0.161857495	0.26643851	0.463977059	0.338496	0.2658187

Figure 21: K-means algorithm k = 6 with NMF with Log NLT, run two

In an attempt to achieve better performance, we normalized the data after reducing the dimensionality of it using NMF and then passed it through our K-means algorithm. We found a stable result that achieved better performance than the previous approach. The results of two successive runs can be seen in Figures 22 and 23.

Dimension	Adjusted Rand	Homogeneity	Completeness	V-measure	Adjusted Mutual Information Score
1	0	4.8877E-16	1	9.78E-16	1.00908E-15
2	0.1496931	0.17521793	0.204298483	0.188644	0.174896381
3	0.127230754	0.17882779	0.174209231	0.176488	0.173895709
4	0.128478507	0.20772963	0.210681086	0.209195	0.207420808
5	0.244324342	0.29933122	0.309042558	0.304109	0.299058107
6	0.154114214	0.25000914	0.264421352	0.257013	0.249716783
7	0.183063401	0.27440172	0.281540992	0.277926	0.274118875
8	0.123022377	0.1870672	0.192972193	0.189974	0.186750292
9	0.127767539	0.20472769	0.203808944	0.204267	0.20350001
10	0.126998852	0.2118441	0.217431148	0.214601	0.211536886
11	0.15254467	0.2321986	0.230302673	0.231247	0.230005126
12	0.186004735	0.23925	0.237676765	0.238461	0.237381593
13	0.12485501	0.2136929	0.226411944	0.219869	0.213386411
14	0.177444992	0.2479539	0.240683155	0.244264	0.240395907
15	0.107349763	0.20088611	0.209314057	0.205014	0.200574628
16	0.164407995	0.27336426	0.280807846	0.277036	0.273081036
17	0.120563819	0.22904822	0.241276991	0.235004	0.228747706
18	0.157039099	0.19874514	0.202426854	0.200569	0.19843282
19	0.121111276	0.18271556	0.175743973	0.179162	0.175435003
20	0.170237202	0.26641499	0.264241006	0.265324	0.263956587

Figure 22: K-means algorithm k = 6 with NMF with Log NLT & normalized data, run one

Dimension	Adjusted Rand	Homogeneity	Completeness	V-measure	Adjusted Mutual Information Score
1	0	4.8877E-16	1	9.78E-16	1.00908E-15
2	0.150326812	0.17590553	0.203992779	0.188911	0.175584249
3	0.126901058	0.1787264	0.17402891	0.176346	0.173715466
4	0.128818808	0.20762235	0.210280303	0.208943	0.207313492
5	0.244554115	0.29941137	0.309183106	0.304219	0.299138283
6	0.146212699	0.24560105	0.249061363	0.247319	0.245306999
7	0.150281385	0.2475078	0.258834484	0.253044	0.247214463
8	0.122531079	0.2002461	0.200221619	0.200234	0.199909929
9	0.127907138	0.20506274	0.204121904	0.204591	0.203813123
10	0.133689667	0.17772857	0.172975961	0.17532	0.172662263
11	0.180803638	0.232147	0.224495134	0.228257	0.224202865
12	0.15367032	0.24431067	0.238665881	0.241455	0.238376013
13	0.153240474	0.23494912	0.233363756	0.234154	0.233066975
14	0.179045841	0.2652462	0.270304143	0.267751	0.264959804
15	0.117741606	0.1608514	0.165377086	0.163083	0.160524318
16	0.129252122	0.25657512	0.272419357	0.26426	0.256285348
17	0.159909228	0.24030619	0.23789561	0.239095	0.237601564
18	0.068132425	0.16275202	0.169943128	0.16627	0.162425685
19	0.173900223	0.24226012	0.248277923	0.245232	0.241964761
20	0.169565392	0.28294203	0.286770885	0.284844	0.282662549

Figure 23: K-means algorithm k = 6 with NMF with Log NLT & normalized data, run two

We see here that at dimension = 5, we get relatively good performance. We than ran this algorithm several more times with dimension = 5 and saw that we were

getting consistently good performance with values very similar to those obtained in the two runs listed above.

The final data representation for this part with k = 6 was as follows: NMF with Log NLT reduced to dimension = 5 with normalized values. The performance was as such:

Homogeneity = 0.299

Completeness = 0.309

V-Measure = 0.304

Adjusted Rand = 0.244

Adjusted Mutual Information = 0.299