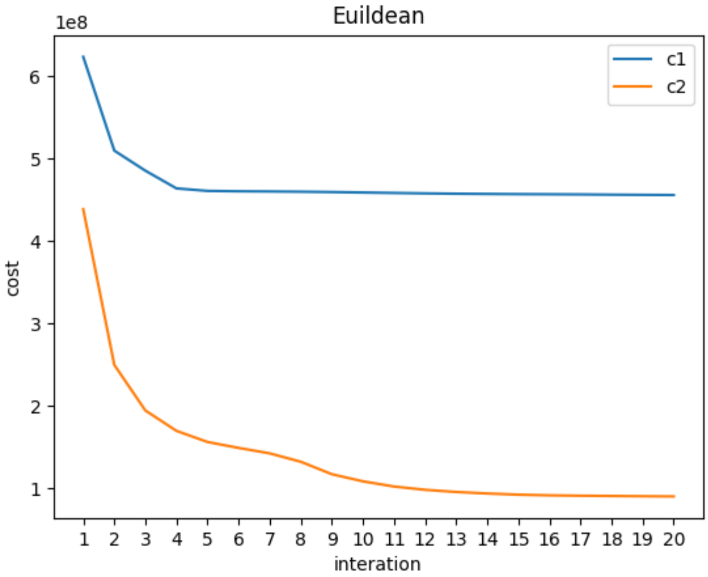
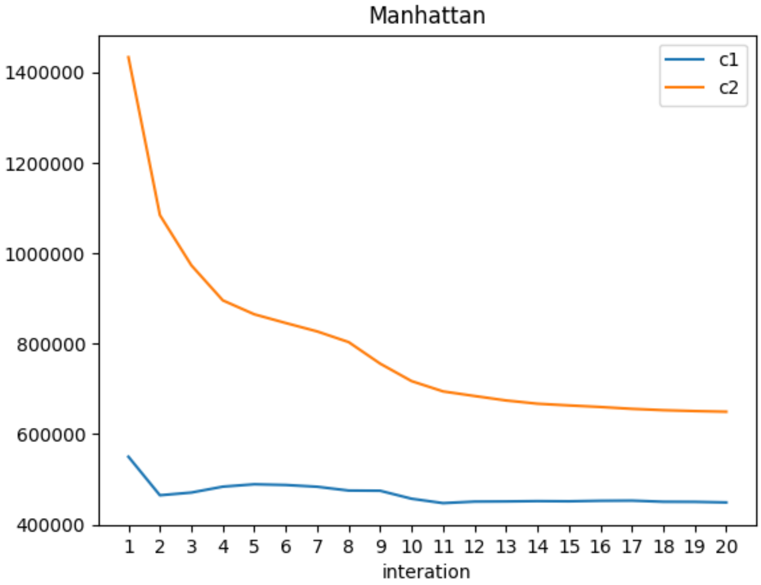
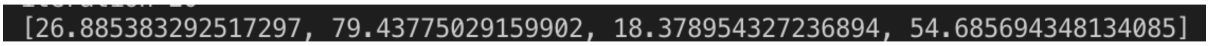
MDA_HW3　　108062633　　洪翊恩
Gitlab: https://gitlab.com/Apr5048/nthu_mda_hw3/



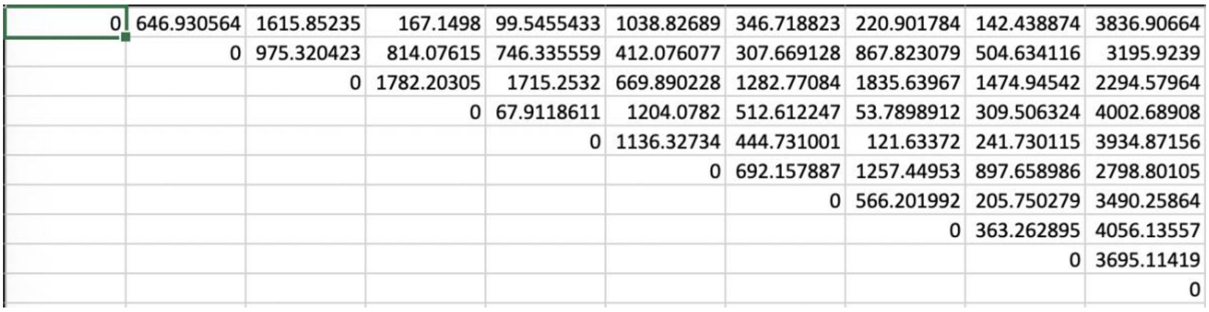


Percentage improvement values
（依序為 c1+Euclidean、c2+Euclidean、c1+Manhattan、c2+Manhattan）



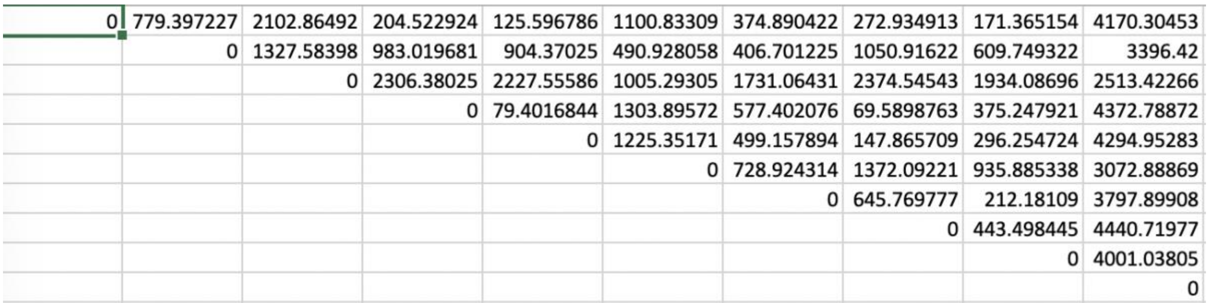
```
[26.885383292517297, 79.43775029159902, 18.378954327236894, 54.685694348134085]
```

c1 20 次 Eulidean 法後的 Eulidean Distance



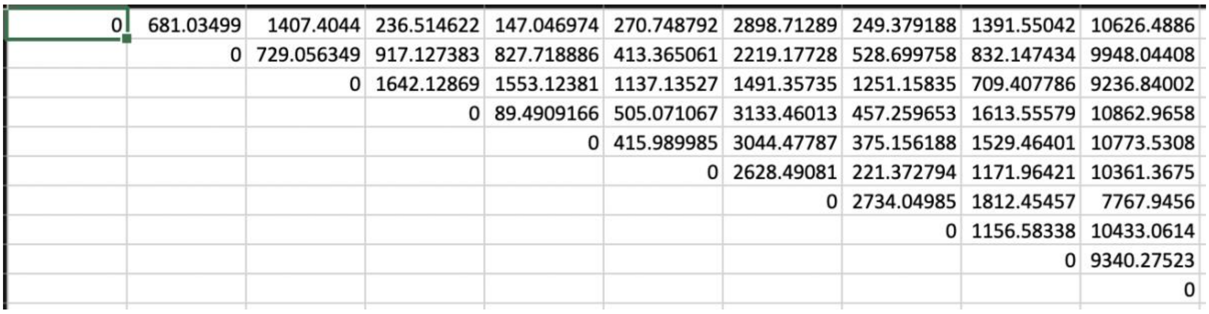
0	646.930564	1615.85235	167.1498	99.5455433	1038.82689	346.718823	220.901784	142.438874	3836.90664
	0	975.320423	814.07615	746.335559	412.076077	307.669128	867.823079	504.634116	3195.9239
		0	1782.20305	1715.2532	669.890228	1282.77084	1835.63967	1474.94542	2294.57964
			0	67.9118611	1204.0782	512.612247	53.7898912	309.506324	4002.68908
				0	1136.32734	444.731001	121.63372	241.730115	3934.87156
					0	692.157887	1257.44953	897.658986	2798.80105
						0	566.201992	205.750279	3490.25864
							0	363.262895	4056.13557
								0	3695.11419
									0

c1 20 次 Eulidean 法後的 Manhattan Distance



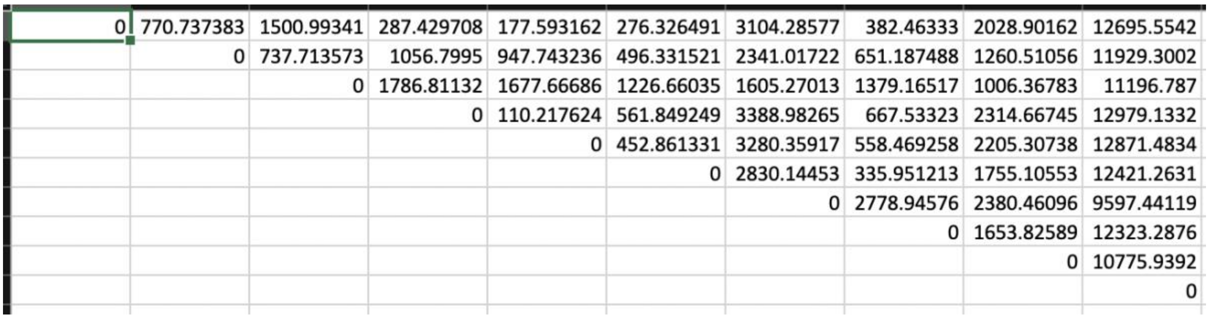
0	779.397227	2102.86492	204.522924	125.596786	1100.83309	374.890422	272.934913	171.365154	4170.30453
	0	1327.58398	983.019681	904.37025	490.928058	406.701225	1050.91622	609.749322	3396.42
		0	2306.38025	2227.55586	1005.29305	1731.06431	2374.54543	1934.08696	2513.42266
			0	79.4016844	1303.89572	577.402076	69.5898763	375.247921	4372.78872
				0	1225.35171	499.157894	147.865709	296.254724	4294.95283
					0	728.924314	1372.09221	935.885338	3072.88869
						0	645.769777	212.18109	3797.89908
							0	443.498445	4440.71977
								0	4001.03805
									0

c1 20 次 Manhattan 法後的 Eulidean Distance



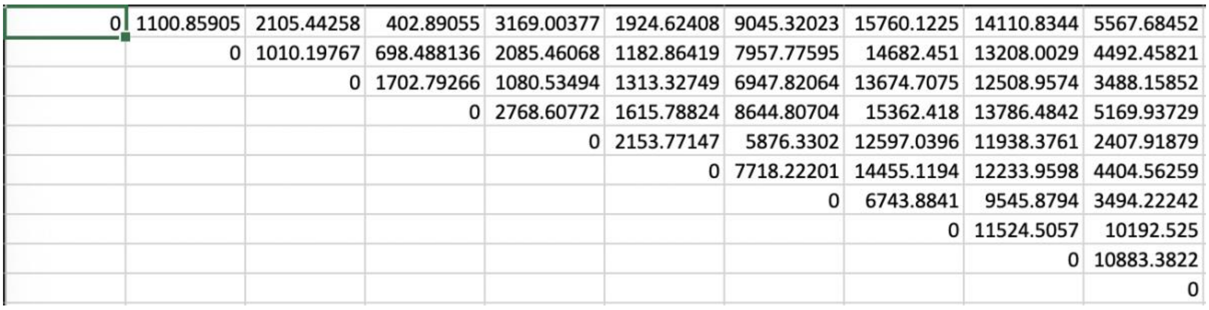
0	681.03499	1407.4044	236.514622	147.046974	270.748792	2898.71289	249.379188	1391.55042	10626.4886
	0	729.056349	917.127383	827.718886	413.365061	2219.17728	528.699758	832.147434	9948.04408
		0	1642.12869	1553.12381	1137.13527	1491.35735	1251.15835	709.407786	9236.84002
			0	89.4909166	505.071067	3133.46013	457.259653	1613.55579	10862.9658
				0	415.989985	3044.47787	375.156188	1529.46401	10773.5308
					0	2628.49081	221.372794	1171.96421	10361.3675
						0	2734.04985	1812.45457	7767.9456
							0	1156.58338	10433.0614
								0	9340.27523
									0

c1 20 次 Manhattan 法後的 Manhattan Distance



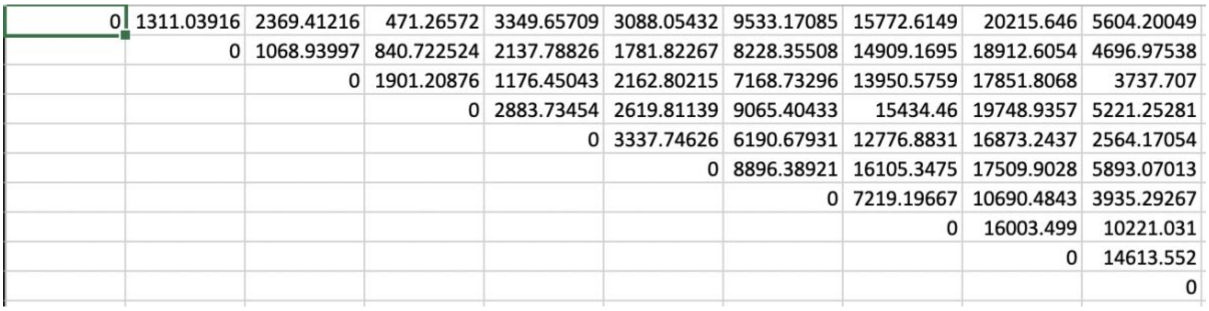
0	770.737383	1500.99341	287.429708	177.593162	276.326491	3104.28577	382.46333	2028.90162	12695.5542
	0	737.713573	1056.7995	947.743236	496.331521	2341.01722	651.187488	1260.51056	11929.3002
		0	1786.81132	1677.66686	1226.66035	1605.27013	1379.16517	1006.36783	11196.787
			0	110.217624	561.849249	3388.98265	667.53323	2314.66745	12979.1332
				0	452.861331	3280.35917	558.469258	2205.30738	12871.4834
					0	2830.14453	335.951213	1755.10553	12421.2631
						0	2778.94576	2380.46096	9597.44119
							0	1653.82589	12323.2876
								0	10775.9392
									0

c2 20 次 Eulidean 法後的 Eulidean Distance



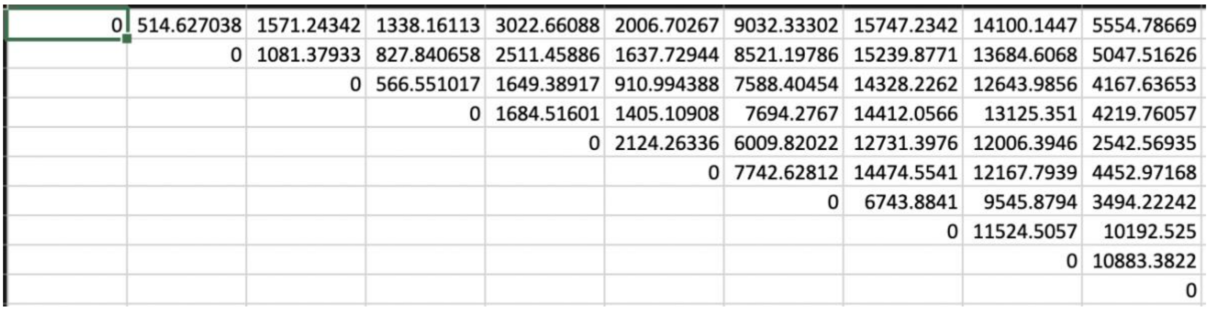
0	1100.85905	2105.44258	402.89055	3169.00377	1924.62408	9045.32023	15760.1225	14110.8344	5567.68452
	0	1010.19767	698.488136	2085.46068	1182.86419	7957.77595	14682.451	13208.0029	4492.45821
		0	1702.79266	1080.53494	1313.32749	6947.82064	13674.7075	12508.9574	3488.15852
			0	2768.60772	1615.78824	8644.80704	15362.418	13786.4842	5169.93729
				0	2153.77147	5876.3302	12597.0396	11938.3761	2407.91879
					0	7718.22201	14455.1194	12233.9598	4404.56259
						0	6743.8841	9545.8794	3494.22242
							0	11524.5057	10192.525
								0	10883.3822
									0

c2 20 次 Eulidean 法後的 Manhattan Distance



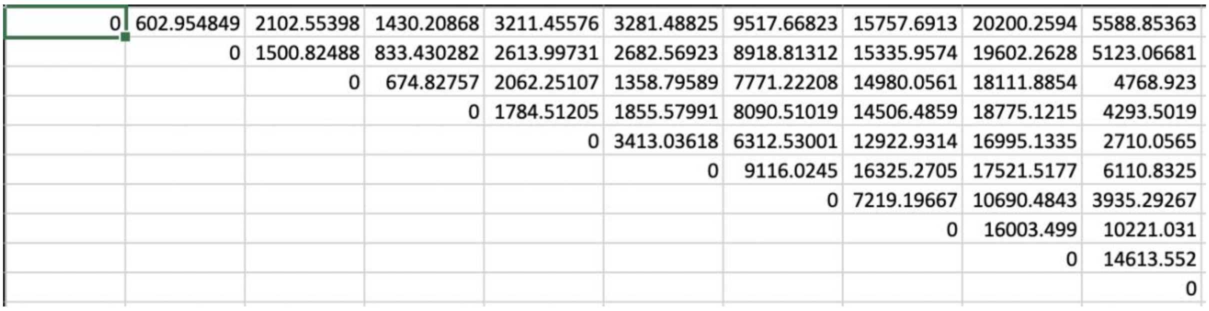
0	1311.03916	2369.41216	471.26572	3349.65709	3088.05432	9533.17085	15772.6149	20215.646	5604.20049
	0	1068.93997	840.722524	2137.78826	1781.82267	8228.35508	14909.1695	18912.6054	4696.97538
		0	1901.20876	1176.45043	2162.80215	7168.73296	13950.5759	17851.8068	3737.707
			0	2883.73454	2619.81139	9065.40433	15434.46	19748.9357	5221.25281
				0	3337.74626	6190.67931	12776.8831	16873.2437	2564.17054
					0	8896.38921	16105.3475	17509.9028	5893.07013
						0	7219.19667	10690.4843	3935.29267
							0	16003.499	10221.031
								0	14613.552
									0

c2 20 次 Manhattan 法後的 Eulidean Distance



0	514.627038	1571.24342	1338.16113	3022.66088	2006.70267	9032.33302	15747.2342	14100.1447	5554.78669
	0	1081.37933	827.840658	2511.45886	1637.72944	8521.19786	15239.8771	13684.6068	5047.51626
		0	566.551017	1649.38917	910.994388	7588.40454	14328.2262	12643.9856	4167.63653
			0	1684.51601	1405.10908	7694.2767	14412.0566	13125.351	4219.76057
				0	2124.26336	6009.82022	12731.3976	12006.3946	2542.56935
					0	7742.62812	14474.5541	12167.7939	4452.97168
						0	6743.8841	9545.8794	3494.22242
							0	11524.5057	10192.525
								0	10883.3822
									0

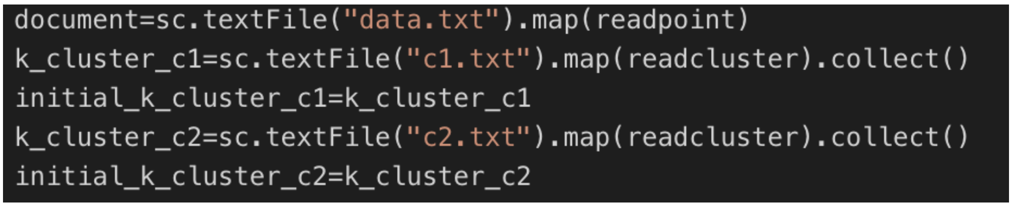
c2 20 次 Manhattan 法後的 Manhattan Distance



0	602.954849	2102.55398	1430.20868	3211.45576	3281.48825	9517.66823	15757.6913	20200.2594	5588.85363
	0	1500.82488	833.430282	2613.99731	2682.56923	8918.81312	15335.9574	19602.2628	5123.06681
		0	674.82757	2062.25107	1358.79589	7771.22208	14980.0561	18111.8854	4768.923
			0	1784.51205	1855.57991	8090.51019	14506.4859	18775.1215	4293.5019
				0	3413.03618	6312.53001	12922.9314	16995.1335	2710.0565
					0	9116.0245	16325.2705	17521.5177	6110.8325
						0	7219.19667	10690.4843	3935.29267
							0	16003.499	10221.031
								0	14613.552
									0

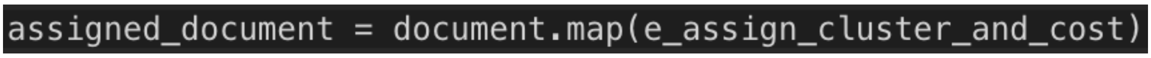
步驟

1. 用 Mapper 讀檔案產生兩組 clusters 的初始值 c1,c2 和 documents



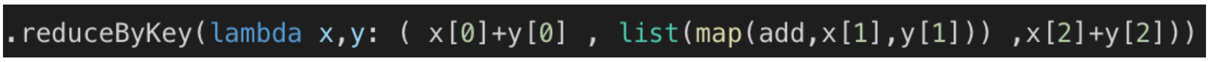
```
document=sc.textFile("data.txt").map(readpoint)
k_cluster_c1=sc.textFile("c1.txt").map(readcluster).collect()
initial_k_cluster_c1=k_cluster_c1
k_cluster_c2=sc.textFile("c2.txt").map(readcluster).collect()
initial_k_cluster_c2=k_cluster_c2
```

2. 用 Mapper 把 document 對 2 組 cluster 計算距離用 Eulidean 或 Manhattan 然後 assign 每個 document 到最近的 cluster ，並回傳 assign 的 cluster、跟 cost



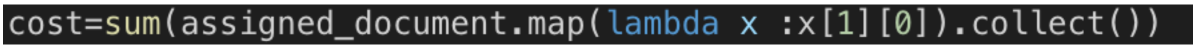
```
assigned_document = document.map(e_assign_cluster_and_cost)
```

3. 用 Reducer 把 assign 後的 document 針對每個 cluster 計算其中所有 document(list)的加總跟數量和這些點對該 cluster 的 cost 和



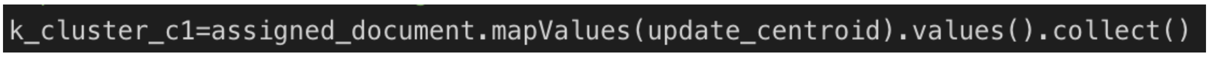
```
.reduceByKey(lambda x,y: ( x[0]+y[0] , list(map(add,x[1],y[1])) ,x[2]+y[2]))
```

4. 計算該次迭代所有的 cost



```
cost=sum(assigned_document.map(lambda x :x[1][0]).collect())
```

5. 用 Mapper 更新 每個 cluster 的 centroid



```
k_cluster_c1=assigned_document.mapValues(update_centroid).values().collect()
```

6. 重複 2~5 直到做了 20 次迭代