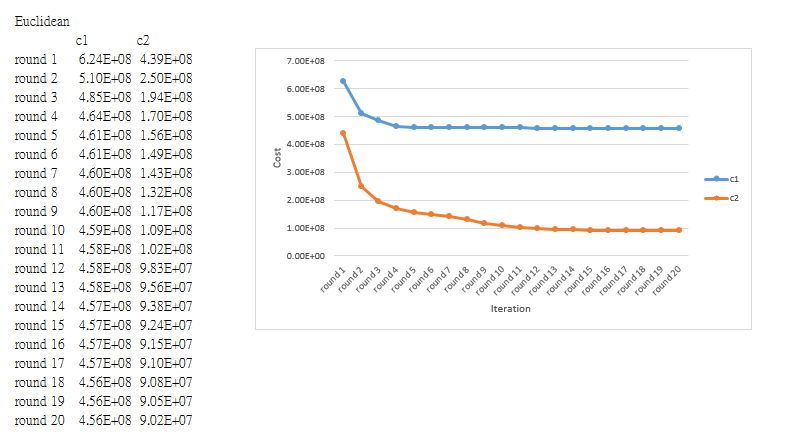
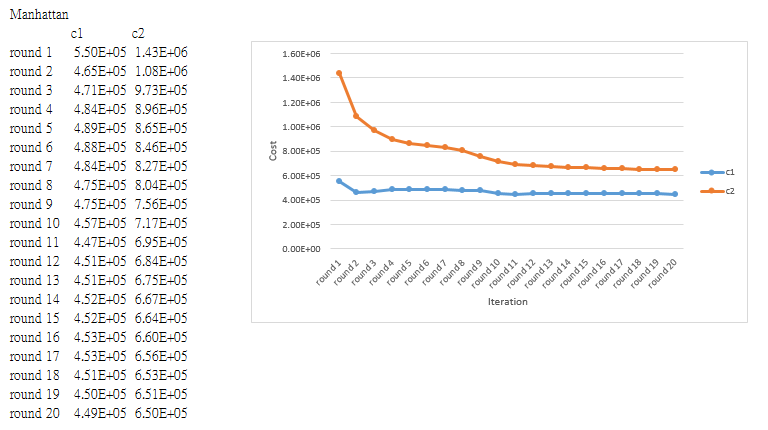
Report

指令：yarn jar kmeansjava-1.0-SNAPSHOT.jar org.apache.hadoop.examples.Kmeans c1.txt c2.txt

1. (a) 和 (b) 的 cost vs. iteration 圖





2. 做完 10 次 iteration 之後 (a) 和 (b) 的 Percentage improvement values

**Euclidean：**

C1：

(623660345.306413 - 459021103.342289) / 623660345.306413 = 0.263988632920431

C2：

(438747790.027914 - 108547377.178569) / 438747790.027914 = 0.752597324372476

從cost改善的情況可以發現，C2比C1有更明顯的改善；而從cost來看，C2一直都比C1好。所以，C2是比較好的initialization。

**Manhattan：**

C1：

(550117.142000001 - 457232.920115077) / 550117.142000001 = 0.168844442016904

C2：

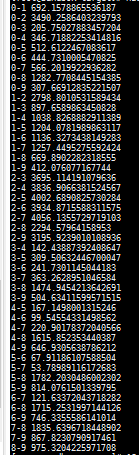
(1433739.31 - 717332.902543228) / 1433739.31 = 0.499676895555561

從cost改善的情況可以發現，C2比C1有更明顯的改善；但若從cost來看，C1一直都比C2好。所以，若以cost改善的觀點來看，C2是比較好的initialization；而從cost來看，C1比較好。

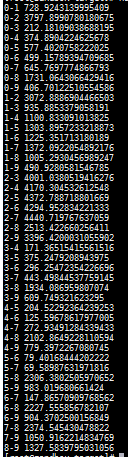
3. 做完 20 次 iterations 之後，所有 Cluster Centroids 兩兩之間的 Euclidean 和 Manhattan Distance

C1

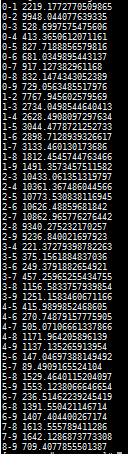
Euclidean-Euclidean distance



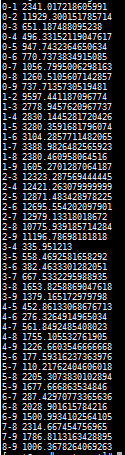
Euclidean-Manhattan distance



Manhattan-Euclidean distance

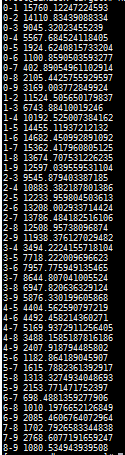


Manhattan-Manhattan distance

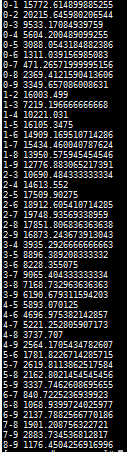


C2

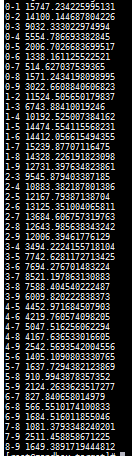
Euclidean-Euclidean distance



Euclidean-Manhattan distance



Manhattan-Euclidean distance



Manhattan-Manhattan distance

