Data Mining Assignment - Clustering

Due : 11/11/13

Gaurav Ramesh(ramesh.48@osu.edu), Adithya Bhat(bhat.51@osu.edu)

K-Means Clustering

Number of Clusters considered : (8, 16, 32, 64)
Number of Documents used for Clustering : 21000 (whole dataset)
Dimension of Feature Vector : 938
Distance Metric used : Euclidean Distance

1. means/centroid calculation stops, when the number of cluster points changing between subsequent iterations reduces to 25% of the total number of clusters
e.g. for 16 clusters, convergence point is when the change in cluster centroids in successive iterations drops below 4 clusters

2. Cluster Entropy calculation :
    e = sigma [weight * clusterEntropy]

where, weight = number of documents in cluster / total number of documents *
    clusterEntropy = - sigma [ratio * log (ratio)],
    ratio = documents of class i / documents in the cluster

* document count here only include the documents with classes

Analyzing the output :

1. at the terminal/console, gives the **number of clusters, time taken for clustering, entropy**
2. creates two graphs, one showing **clusters vs running time**, and the other showing **clusters vs entropy**

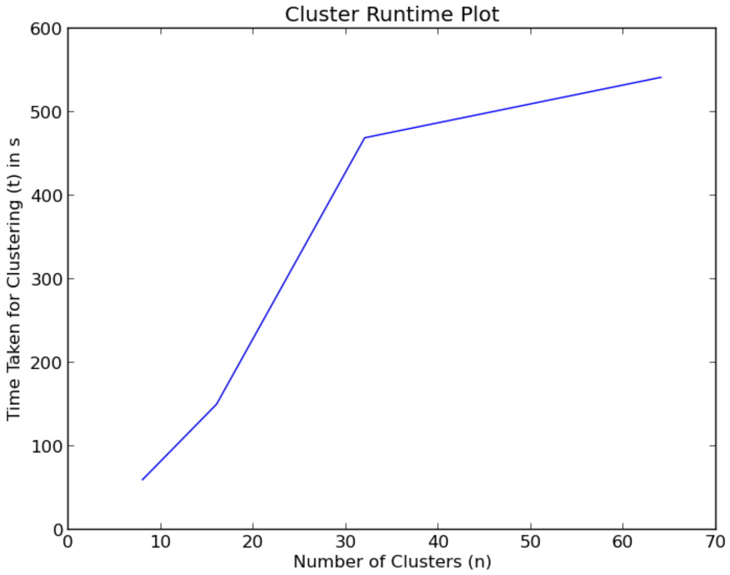


Fig 1.1 Graph showing number of clusters vs time taken for clustering

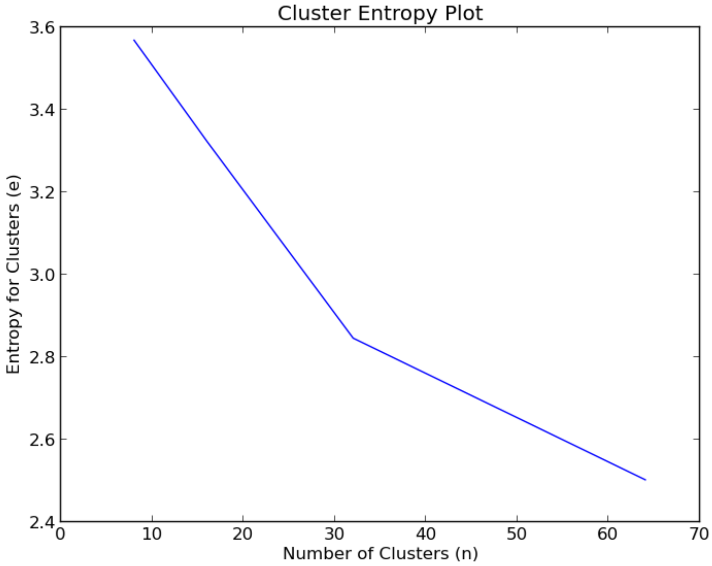
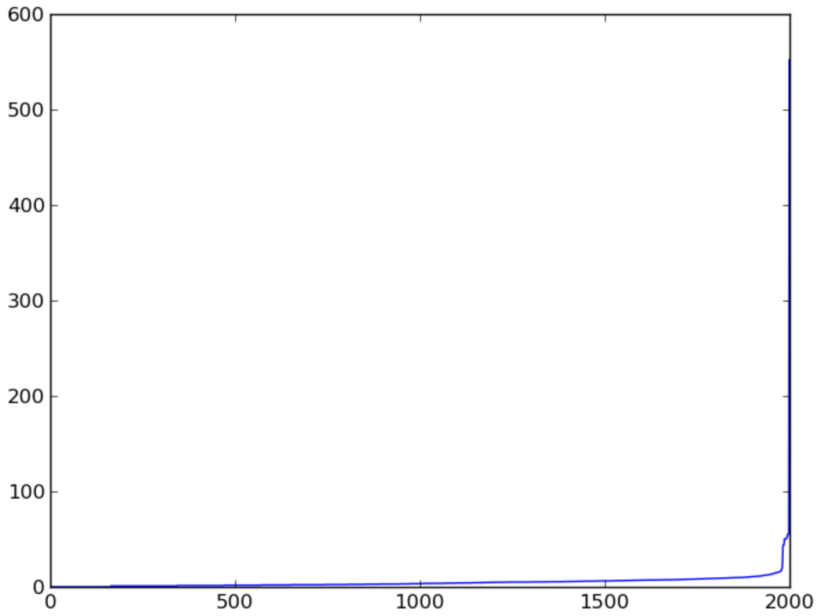


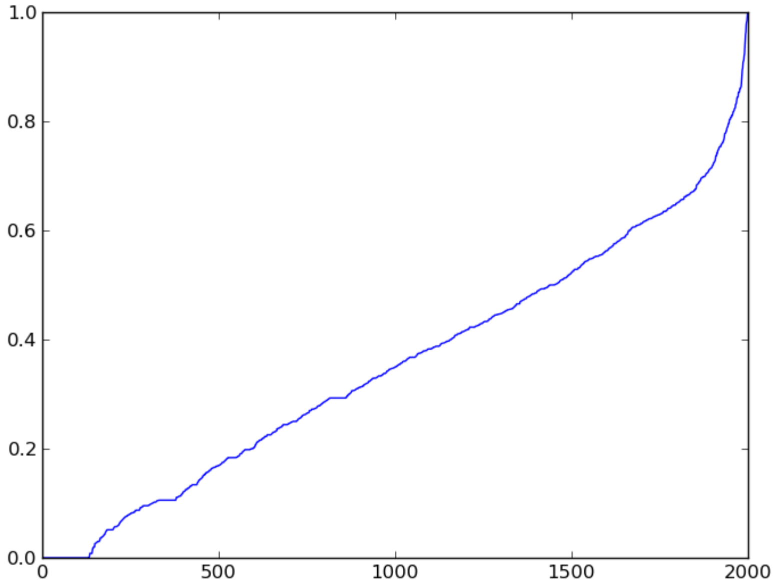
Fig 1.2 Graph showing number of clusters vs Entropy of clustering

DBSCAN

In order to choose epsilon and minpoints, we selected some 2000 documents and found each points 4th nearest neighbour and plotted the graph of distance against the point.
The graph spikes at some point and we considered that as the epislon.
We did this for Euclidean and Cosine distance measure and found 0.4 and 10 as respective epsilon distance.

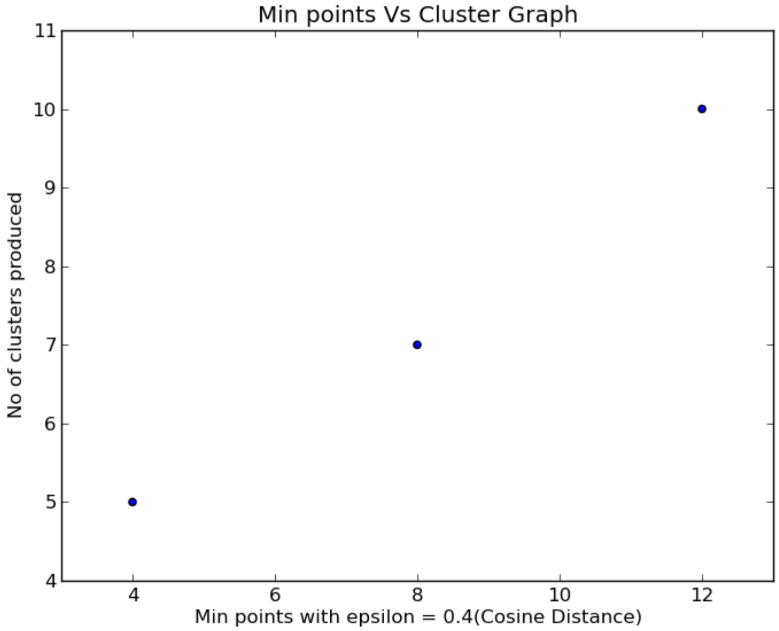


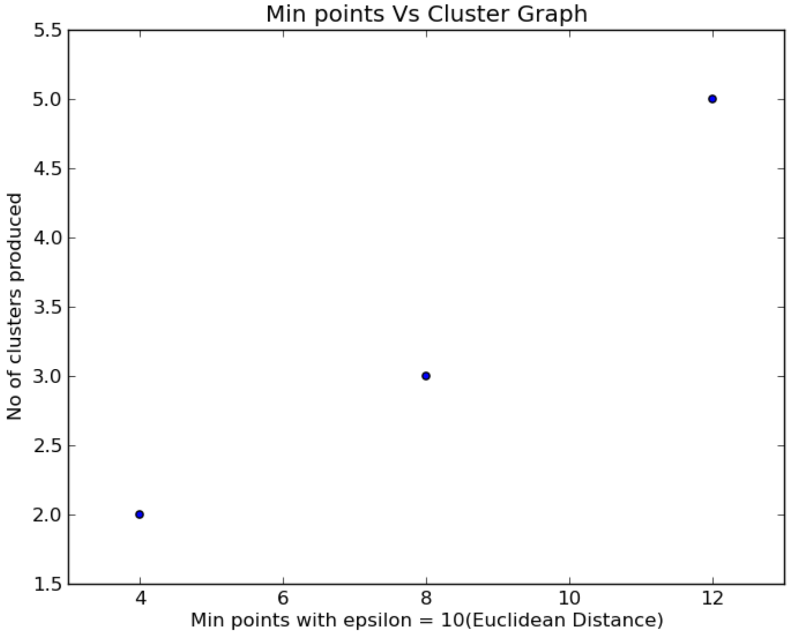
In this graph the y-axis is Euclidean distance and x-axis are points

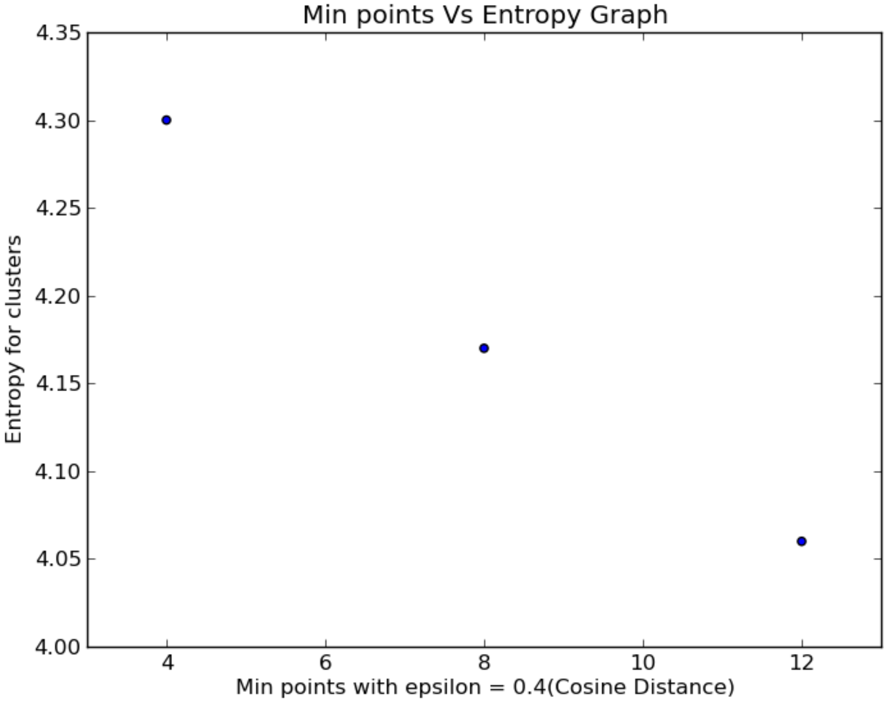


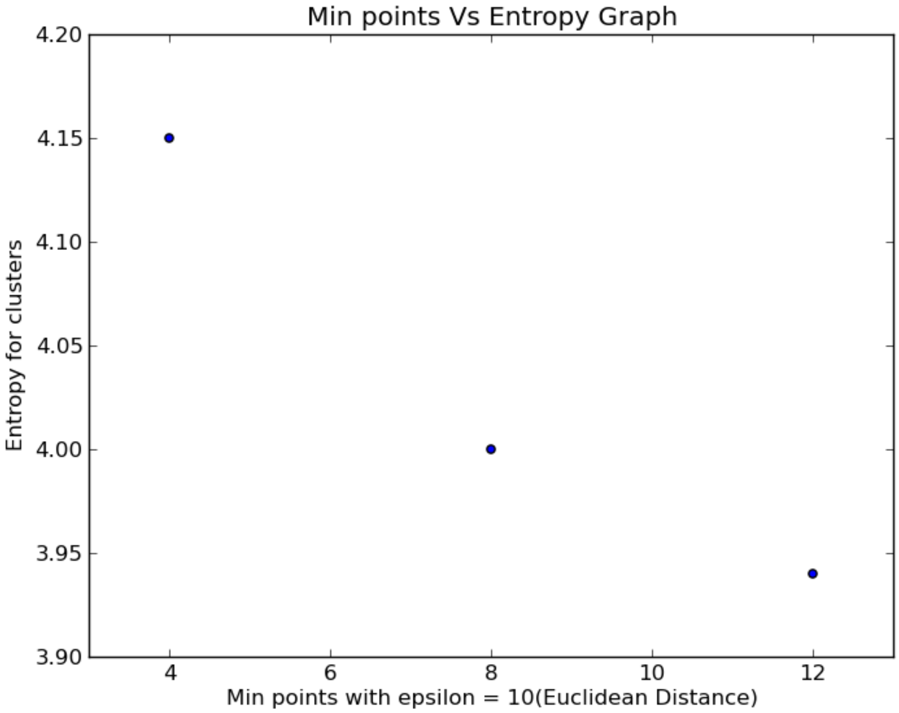
In this graph the y-axis is cosine distance and x-axis are points

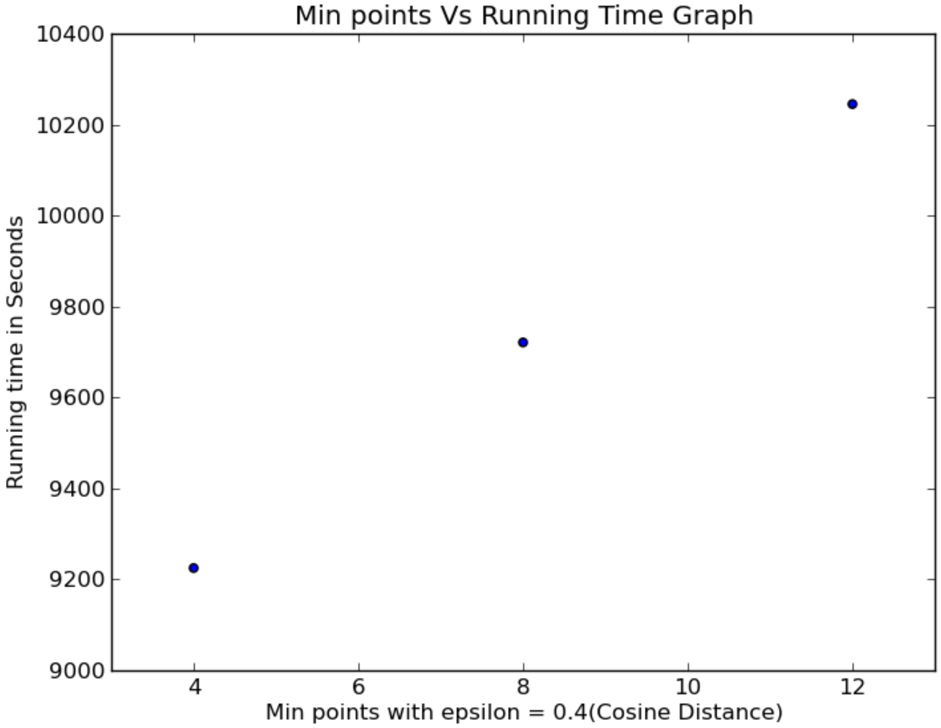
The following are the graphs for various run of DB scan. We ran DB scan for 10000 random sample for cosine distance and it took around 150 min so we ran just for 2000 samples for euclidean distance.






Min points Vs Entropy Graph
Entropy for clusters
4.35
4.30
4.25
4.20
4.15
4.10
4.05
4.00
4
6
8
10
12
Min points with epsilon = 0.4(Cosine Distance)


Min points Vs Entropy Graph
Entropy for clusters
4.20
4.15
4.10
4.05
4.00
3.95
3.90
4
6
8
10
12
Min points with epsilon = 10(Euclidean Distance)


Min points Vs Running Time Graph
Running time in Seconds
10400
10200
10000
9800
9600
9400
9200
9000
4
6
8
10
12
Min points with epsilon = 0.4(Cosine Distance)