Derya Tınmaz 2380947

CENG 499 - Introduction to Machine Learning

Homework 2 Report

Part 1 - KNN

K	Distance Metric	Accuracy Score	Confidence Interval
11	cosine	0.947	(0.937, 0.956)
13	cosine	0.944	(0.941, 0.947)
11	minkowski (p=2)	0.959	(0.951, 0.966)
13	minkowski (p=2)	0.964	(0.961, 0.967)
11	mahalanobis	0.881	(0.876, 0.886)
13	mahalanobis	0.865	(0.860, 0.870)

In this part we were expected to perform grid search on KNN algorithm. We had two hyper parameters: K value (number of neighbors to consider) and distance metric. Our dataset consist of 150 examples, since the optimal k value usually found is the square root of number of examples in the dataset and to become odd number I choose 11 and 13 as k values for my grid search. I tried the combination of these k values with 3 different distance metrics. And also we were expected to perform grid search with cross validation technique. I performed 10-fold stratified cross validation 5 times for each hyper parameter combination. Then, I calculated the confidence intervals of accuracy scores for each. As you can see from table when k is 13 and distance metric is Minkowski (p=2) distance, accuracy score is the highest, and it has a narrow confidence interval. That is why I chose these hyper parameters for my final KNN model.

Derya Tınmaz 2380947

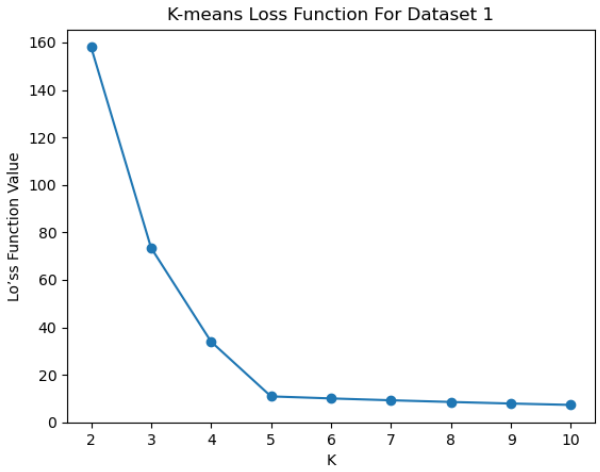
Part 2 - K Means

In this part we were expected to implement K Means and K Means++ clustering algorithms, and select the most suitable number of clusters for given two datasets using elbow method. For each dataset, after implementing the algorithms I calculated the loss values of final clusters for K values up to 10. Then I plotted the loss value versus K graphs for each to detect the elbow point for optimum K (number of clusters) value.

The suitable K values for individual datasets are same for K means and K means++ for my implementation, because their loss values for different k's are too close to each other, but in K Means++ the confidence interval is narrower than K Means due to better initialization of means.

Part 2.1 - K Means & Dataset 1:

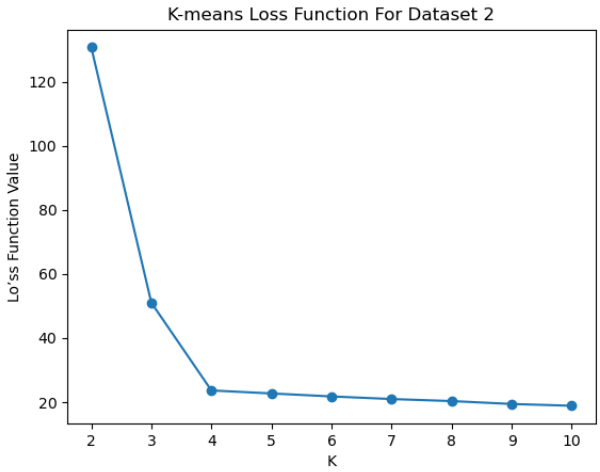
K	Loss value	Confidence interval
1	331.058	(331.058, 331.058)
2	157.994	(157.994, 157.994)
3	73.475	(73.475, 73.475)
4	33.956	(33.956, 33.956)
5	10.893	(10.893, 10.893)
6	10.040	(9.991, 10.090)
7	9.258	(9.254, 9.262)
8	8.547	(8.517, 8.577)
9	7.889	(7.824, 7.953)
10	7.310	(7.286, 7.334)



For dataset 1 the most suitable K value (number of clusters) is 5 with using K means algorithm. The loss function value does not decrease drastically after 5 (elbow point), but k=4 to k=5 there is a drastic decrease.

Part 2.2 - K Means & Dataset 2:

K	Loss value	Confidence interval
1	555.807	(555.807, 555.807)
2	130.692	(130.692, 130.692)
3	51.026	(51.026, 51.026)
4	23.652	(23.652, 23.652)
5	22.708	(22.696, 22.719)
6	21.778	(21.760, 21.796)
7	20.964	(20.937, 20.991
8	20.331	(20.246, 20.416)
9	19.425	(19.395, 19.454)
10	18.869	(18.793, 18.946)

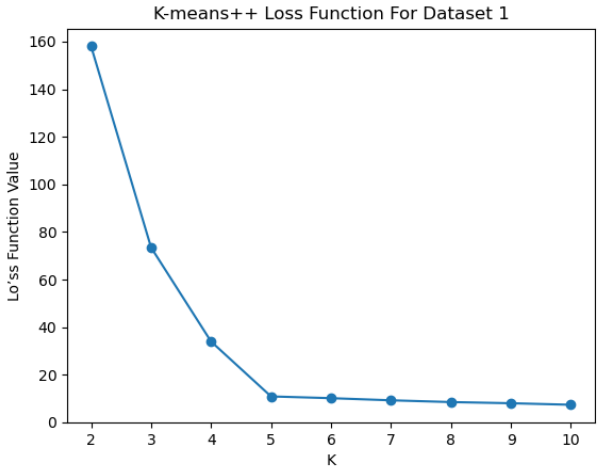


For dataset 2 the most suitable K value (number of clusters) is 4 with using K means algorithm. The loss function value does not decrease drastically after 4 (elbow point), but k=3 to k=4 there is a drastic decrease.

Derya Tınmaz 2380947

Part 2.3 - K Means++ & Dataset 1:

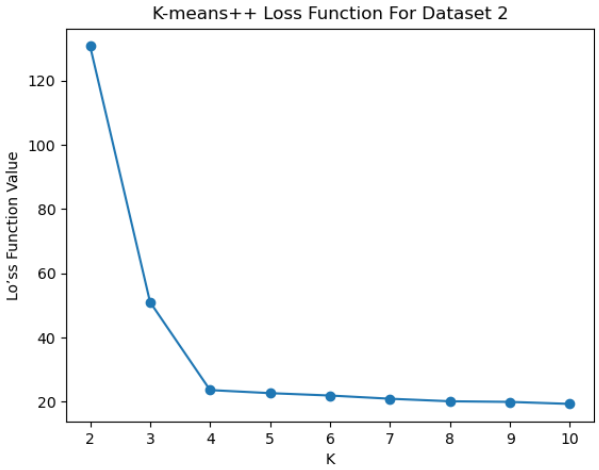
K	Loss value	Confidence interval
1	331.058	(331.058, 331.058)
2	157.994	(157.994, 157.994)
3	73.475	(73.475, 73.475)
4	33.956	(33.956, 33.956)
5	10.893	(10.893, 10.893)
6	10.158	(10.158, 10.158)
7	9.259	(9.259, 9.259)
8	8.523	(8.523, 8.523)
9	8.031	(8.031, 8.031)
10	7.390	(7.390, 7.390)



For dataset 1 the most suitable K value (number of clusters) is 5 with using K means++ algorithm. The loss function value does not decrease drastically after 5 (elbow point), but k=4 to k=5 there is a drastic decrease.

Part 2.4 - K Means++ & Dataset 2:

K	Loss value	Confidence interval
1	555.807	(555.807, 555.807)
2	130.692	(130.692, 130.692)
3	51.026	(51.026, 51.026)
4	23.652	(23.652, 23.652)
5	22.695	(22.695, 22.695)
6	21.952	(21.952, 21.952)
7	20.950	(20.950, 20.950)
8	20.163	(20.163, 20.163)
9	19.968	(19.968, 19.968)
10	19.882	(19.344, 19.344)



For dataset 2 the most suitable K value (number of clusters) is 4 with using K means++ algorithm. The loss function value does not decrease drastically after 4 (elbow point), but k=3 to k=4 there is a drastic decrease.

Worst-case running time analysis for K means:

- For each iteration we calculate cluster centers -> **K*d**
- **then** we assign clusters according to their distance to K cluster means -> **N*K*d**

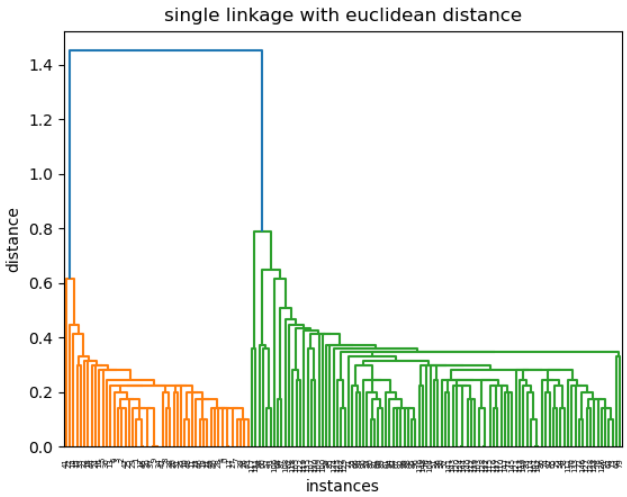
Overall for one iteration: **K*d + N*K*d**
Total= **I*(K*d + N*K*d)**
If we write it in big O notation-> **O(I*K*N*d)**
N: the number of data points
d: data sample vector dimension
K: cluster number
I: the number of iterations

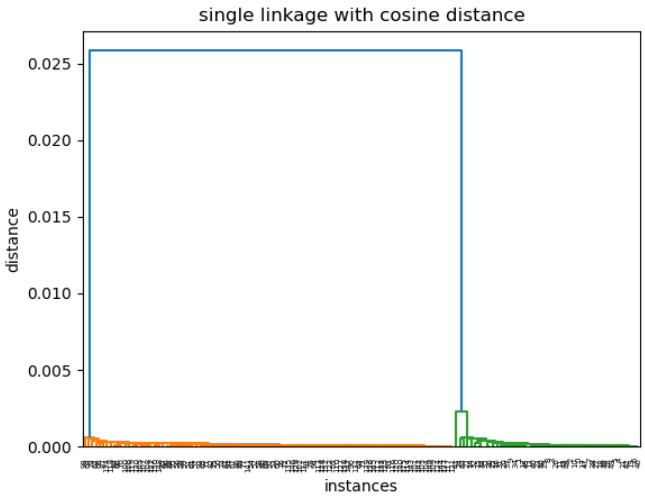
Derya Tınmaz 2380947

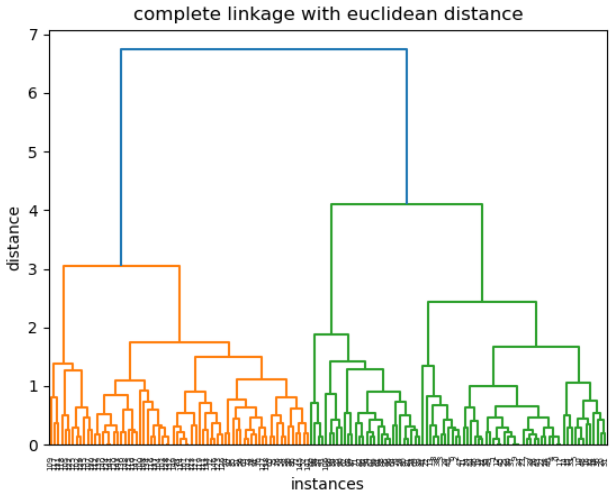
Part 3 - Hierarchical Agglomerative Clustering

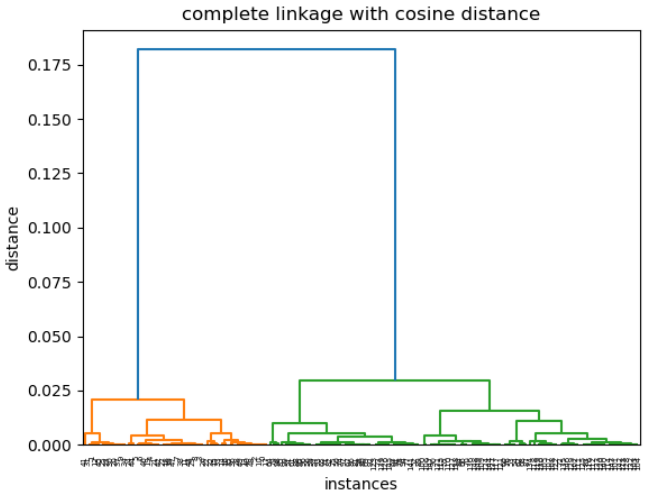
In this part we were expected to implement HAC algorithm and grid search on hyper parameters of HAC which are linkage criterion and distance/similarity measure. Also, for each hyper parameter configurations we were expected to perform silhouette analysis on given dataset for different K values to select most suitable K value(number of clusters).

Dendrogram Plots





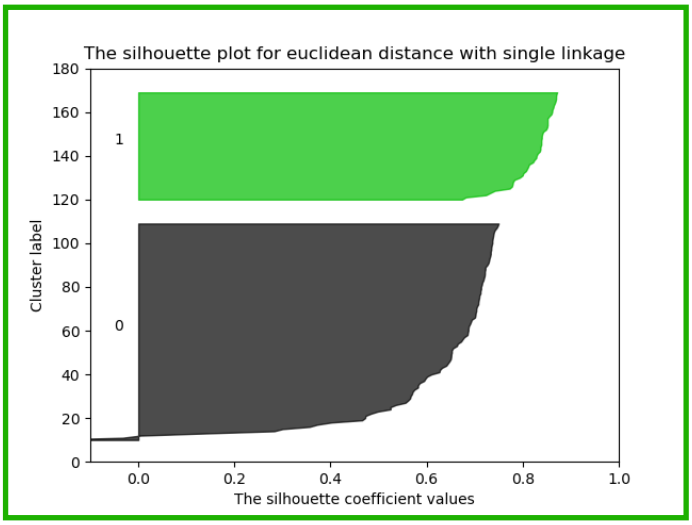




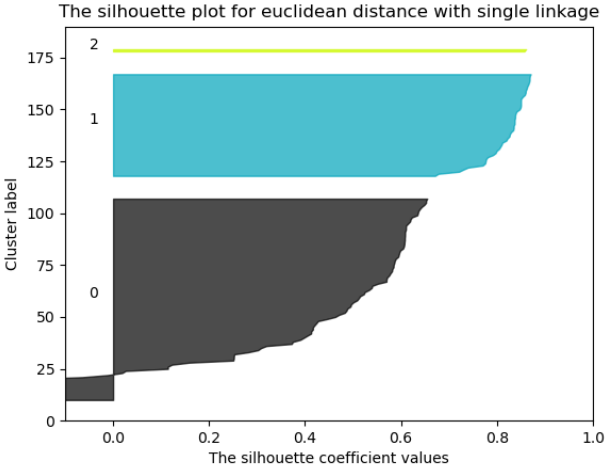
Derya Tınmaz 2380947

Silhouette Value Plots

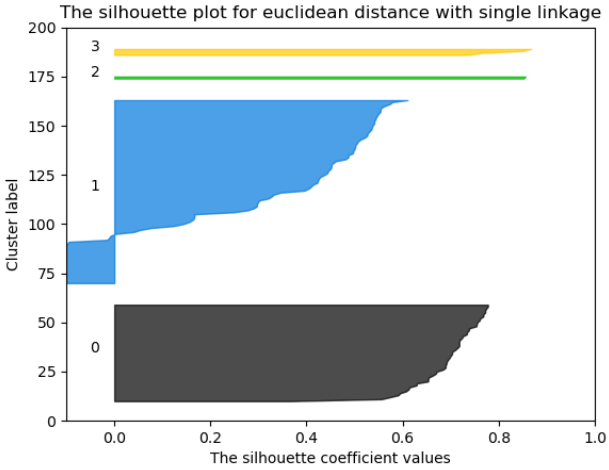
a) single linkage with euclidean distance



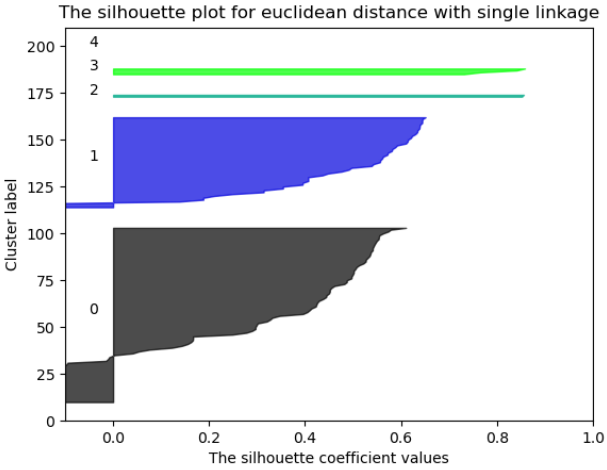
k=2 avg silhouette score: 0.68810517



k=3 avg silhouette score: 0.53133893



k=4 avg silhouette score: 0.39444217

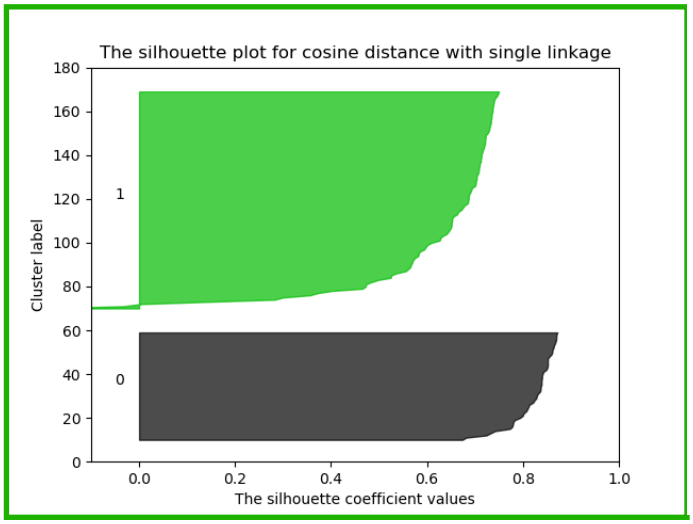


k=5 avg silhouette score: 0.31220323

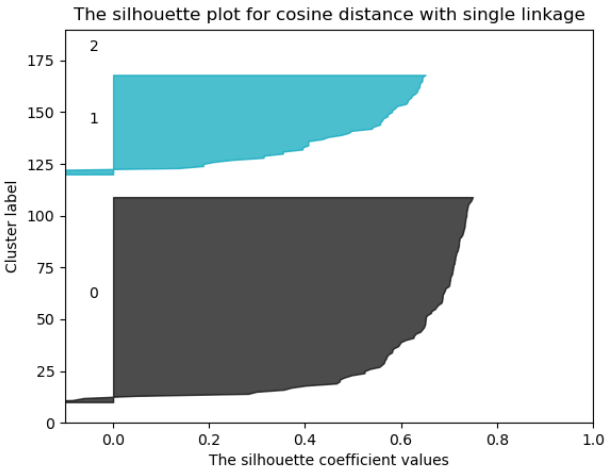
For single linkage with euclidean distance the clusters are formed best when k=2. We can conclude it from their average silhouette score (highest one) and from the graphs. Silhouette values of data points of all clusters are closer to average silhouette score when k=2, and less number of silhouette values are less than 0.

Derya Tınmaz 2380947

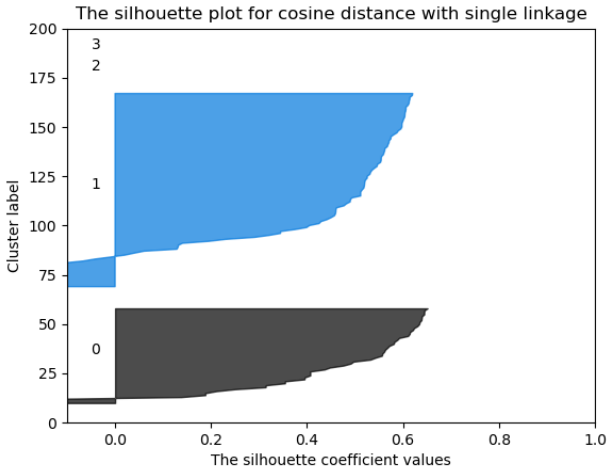
b)single linkage with cosine distance



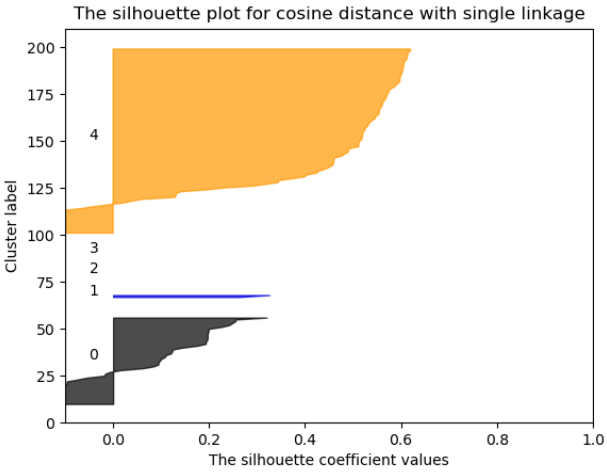
k=2 avg silhouette score: 0.68810517



k=3 avg silhouette score: 0.56061506



k=4 avg silhouette score: 0.376524



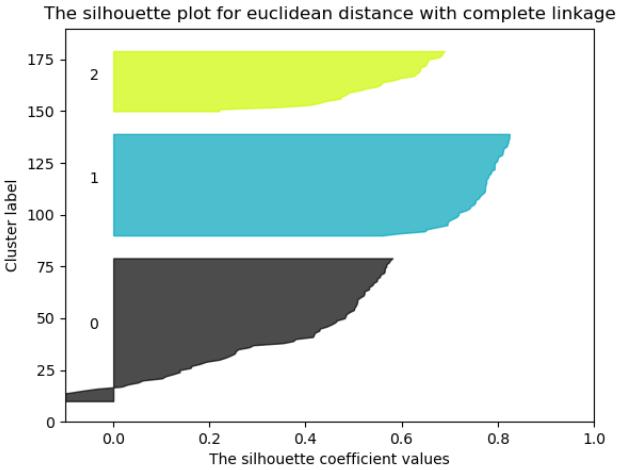
k=5 avg silhouette score: 0.2478568

For single linkage with cosine distance the clusters are formed best when k=2. We can conclude it from their average silhouette score (highest one) and from the graphs. Silhouette values of data points of all clusters are closer to average silhouette score when k=2, and less number of silhouette values are less than 0.

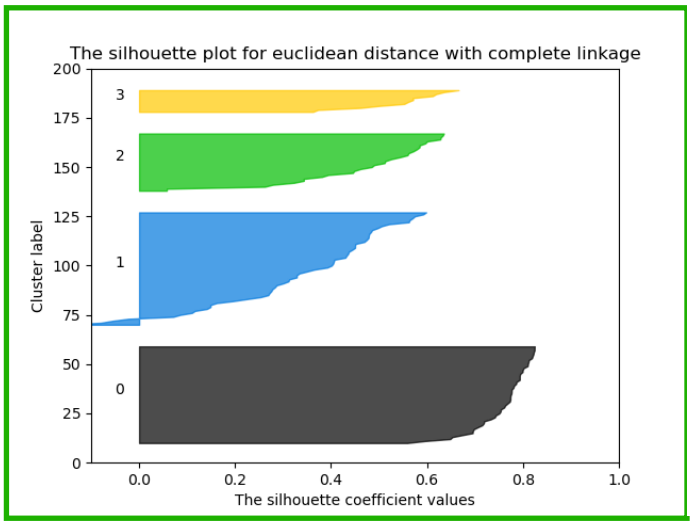
Derya Tınmaz 2380947

c) complete linkage with euclidean distance

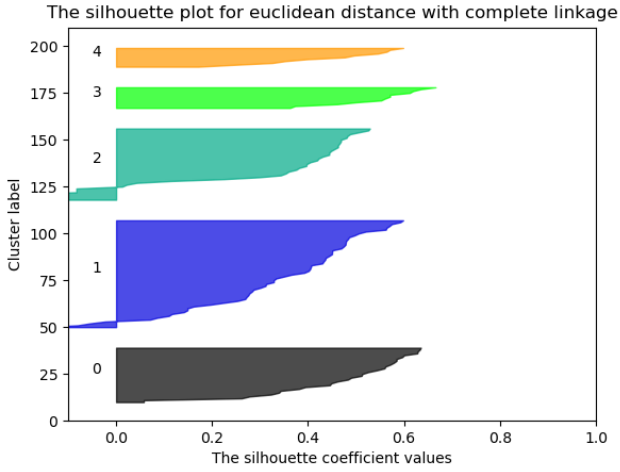
k=2 avg silhouette score: 0.5023174



k=3 avg silhouette score: 0.5191346



k=4 avg silhouette score: 0.51993114

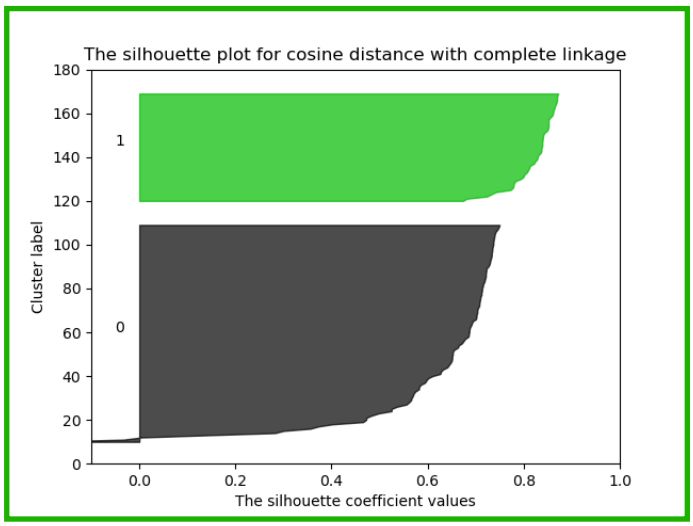


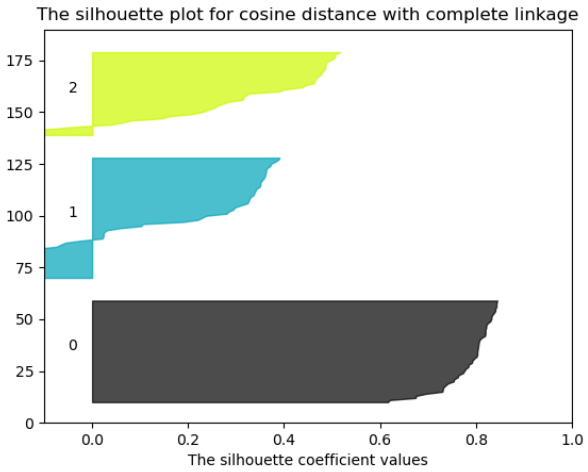
k=5 avg silhouette score: 0.3702072

For complete linkage with euclidean distance the clusters are formed best when k=4. We can conclude it from their average silhouette score (highest one) and from the graphs. Silhouette values of data points of all clusters are closer to average silhouette score when k=4, and less number of silhouette values are less than 0.

Derya Tınmaz 2380947

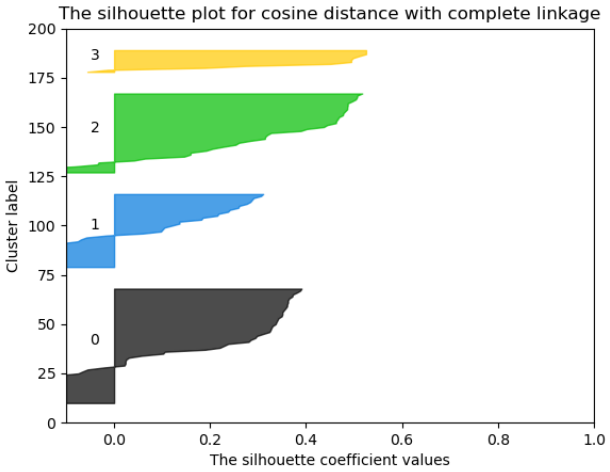
d) complete linkage with cosine distance

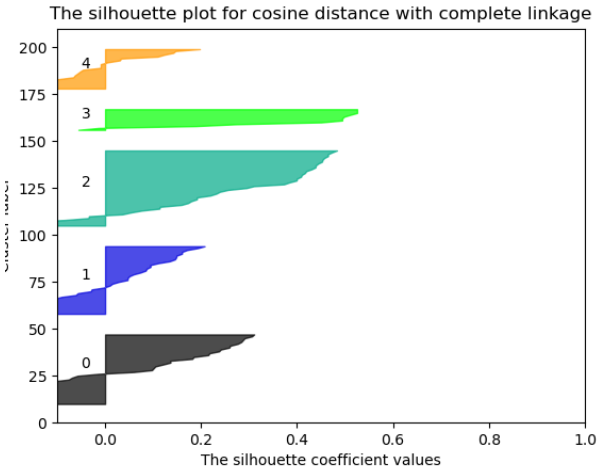




k=2 avg silhouette score: 0.68810517

k=3 avg silhouette score: 0.3834758





k=4 avg silhouette score: 0.15615867

k=5 avg silhouette score: 0.09422624

For complete linkage with cosine distance the clusters are formed best when k=2. We can conclude it from their average silhouette score (highest one) and from the graphs. Silhouette values of data points of all clusters are closer to average silhouette score when k=2, and less number of silhouette values are less than 0.

Derya Tınmaz 2380947

Silhouette Analysis

Linkage	Distance Metric	K	Average Silhouette Score
Single	euclidean	2	0.68810517
Single	euclidean	3	0.53133893
Single	euclidean	4	0.39444217
Single	euclidean	5	0.31220323
Single	cosine	2	0.68810517
Single	cosine	3	0.56061506
Single	cosine	4	0.376524
Single	cosine	5	0.2478568
Complete	euclidean	2	0.5023174
Complete	euclidean	3	0.5191346
Complete	euclidean	4	0.51993114
Complete	euclidean	5	0.3702072
Complete	cosine	2	0.68810517
Complete	cosine	3	0.3834758
Complete	cosine	4	0.15615867
Complete	cosine	5	0.09422624

Among these configurations three of them have the highest average silhouette score. When k is equal to 2 single linkage with euclidean distance, single linkage with cosine distance, and complete linkage with cosine distance. Thus, these three configurations can be used to cluster our dataset well.

Worst-case running time analysis for HAC:

N: the number of data points , **D:** dimension

Since we've to perform **N** iterations (worst case) and in each iteration, we need to update the similarity matrix **(N*D)** and restore the matrix (**N**) in big O notation -> **O(D*N^3)**

- This is huge compered to K Means algorithm complexity for big datasets, that is why I would choose to implement K Means clustering algorithm with a dataset consisting of 1 million data points each of which has a dimension of 120000