1. Data Setup (Load and Preprocessing)

1.1 Yale Face Dataset B

Loop over the dataset, extract each image and save the image in array, also extract list of image class from the image name and file location to be used later for external evaluation.

Apply PCA to extract the features from the image with7 PCA components since most of the data variance found in the first 7. Reshape the data to be (number of images, (image Hight*7 PCA components)) to be suitable as input for cluster algorithm.

1.2Multi-Domain Sentiment Dataset

The code extracts each line as an observation and the review as a label for this line , and by applying CountVectorizer and TfidfTransformer we can get the final dataset which will be used later in different cluster algorithms.

The problem with this type of data is, it is very high in dimensions and will lead to a poor performance during clustering step , therefor the dimensions should be reduced, the PCA is not applicable in this case since the data is sparse data , therefore I used TruncatedSVD algorithm. However another problem will arise which is what is the suitable number of components, this is trade of between the computational cost of the clustering algorithm and the sum of variance components , the low number of components have a low sum of variance which is not a good representation of the data and the high number of components will be very computational expensive and will have also inefficient performance during the clustering algorithm .Therefore, I choose a 10 after testing different number to see their performance during the cluster step

2. K-means Clustering

2.1 Yale Face Dataset B

I had two array, one which contains the full dataset and the other one contains the 7 PCA components for every image.

I test two similarity measure namely the Euclidian distance and cosine similarity the reason to choose those measures is that ,the Euclidian distance is the default measures for kmean clustering algorithm because it is quit fast and usually yield a good result, and cosine similarity is a good measures since we are dealing with different observation represented as a vector. To find the best set of parameters which are (array with the PCA or without, similarity measure, K) I run a code to generate (Silhouette score vs K) graph for each case. And the best set of parameters will be the one with the highest Silhouette score.

Since the actual number of classes are 10 and the best clustering algorithm according to the (Silhouette score vs K) graph is different number then 10, so using the external evolution measure will be inefficient, therefore, I used only internal evolution measures.

So along with Silhouette score which is already used to find the best parameter, I used Calinski and Harabasz score which represent the ration between "within-cluster dispersion and the between-cluster dispersion " (sklearn documentation) and Davies-Bouldin score , and the reason to choose this two score is they are both internal and fast and easy interpret.so the higher Calinski and Harabasz score indicate a better performance and the lower Davies-Bouldin score indicate a better performance.

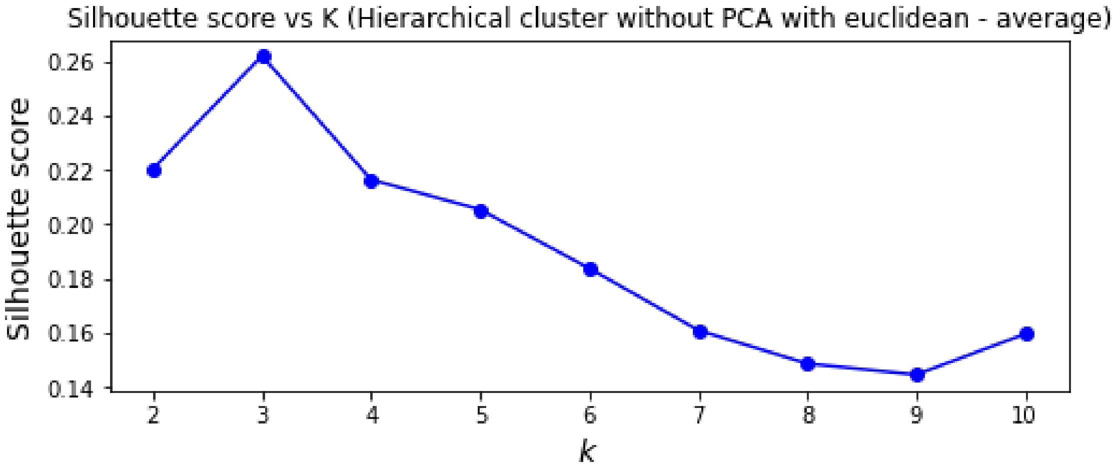
2.2 Multi-Domain Sentiment Dataset

The exact same concept had been applied to this dataset; however, I only used the array with the principle component analysis.

3. Hierarchical Clustering

3.1 Yale Face Dataset B

For this part I choose Agglomerative clustering algorithm to create the hierarchical clustering, however there are two mean parameters should be consider for this type of clustering, the similarity measure which is called the affinity in sklearn library ,and the link type which is called the linkage in sklearn . so, I chose three different affinity namely Euclidean, cosine, Manhattan and three type of linkage, to choose the set of parameter in including the optimal K and the two arrays with the PCA or without it, I run a code which will generate a graph titled with set of parameter the cluster is using against the different k values as in this graph.



And I ran the code for each array to choose the best set of parameters available.
The evaluation for the cluster performance had been done also using the Calinski and Harabasz and Davies-Bouldin score for the same reason.

3.2 Multi-Domain Sentiment Dataset

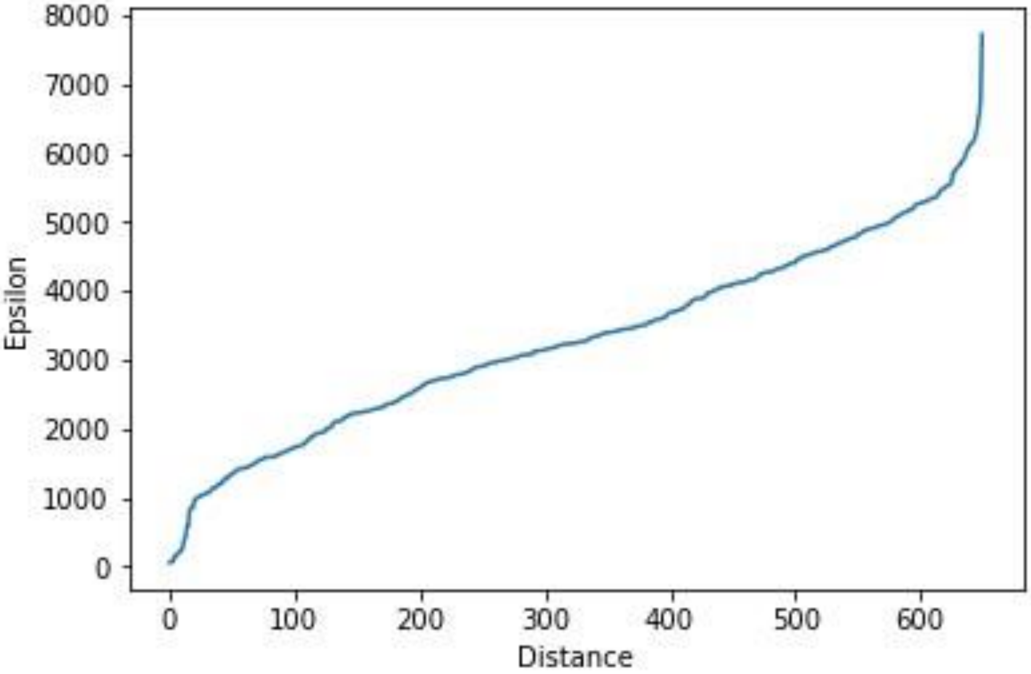
The exact same concept had been applied to this dataset; however, I only used the array with the PCA. On set of parameter shows a good result with number of cluster =5 , therefore I tried to test the cluster performance using two external evolution method namely (adjusted_rand_score and homogeneity_completeness_v_measure) however the performance of the cluster was very bad

4. Other Clustering Method (

4.1 Yale Face Dataset B

For this part I used DBSCAN, without having a visualization on how the dataset is represented, it is quit hard to choose the best cluster algorithm, so by assuming that the DBSCAN on this dataset may be applicable, I work with DBSCAN which is required three different parameter , the most important one is the epsilon value , to choose it , a graph between the epsilon vs the distance the seventh- NN (as the one below) can be used to see where is the sharp change in the line occurred which will be corresponding to best epsilon value.

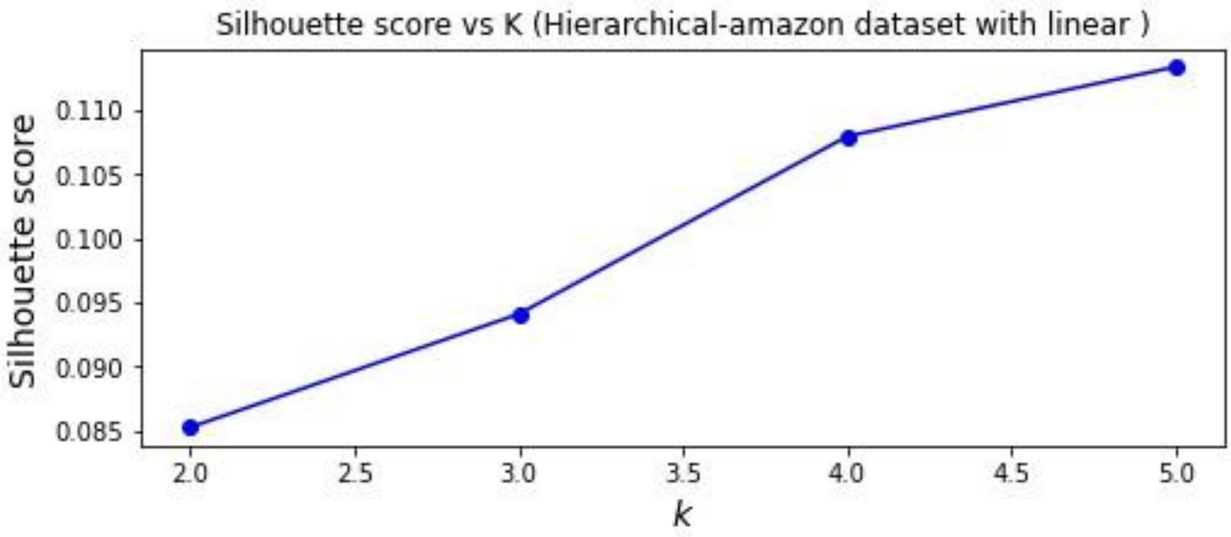
The second parameter which is number sample I chose it by trial and error method, and by applying three different similarity measure we can evaluate the best set of parameters.



I chose the best parameters by checking Silhouette score, however the problem with DBSCAN is that, the label with different set of parameters may be having one class because of how the DBSCAN work, which make Silhouette score not aapplicable in this case. the Calinski and Harabasz and Davies-Bouldin score also are used but the overall performance was poor for this cluster

4.2 Multi-Domain Sentiment Dataset

Due to the poor performance for DBSCAN for yale dataset and after I test it on this dataset and get a poor performance, I chose the Spectral Clustering for the sentiment dataset as the Optional Clustering Method. There are two mean parameter for this algorithm effect the it performance significantly , the number of k and the affinity, for the affinity a pairwise kernels algorithm can be used to test for the similarity , therefore I chose this set of pairwise kernels
(linear, nearest neighbors, RBF, Laplacian, Sigmoid, cosine) and by plotting Silhouette score vs K for each pairwise kernel, the best set of parameters can be found.

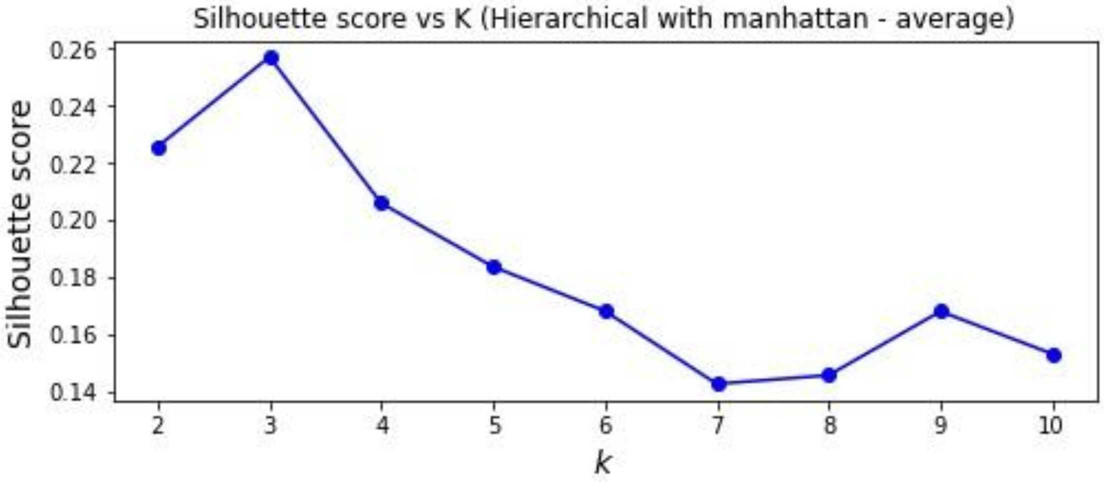


The Calinski and Harabasz and Davies-Bouldin score also are used to evaluating the performance of the best set.

5. The Best Model

5.1 Yale Face Dataset B

The best model was Hierarchical Cluster Number clusters=3 affinity=Manhattan linkage=average (on the original dataset without PCA)



Silhouette score=.27 Davies-Bouldin=1.2

Calinski and Harabasz= 262.03

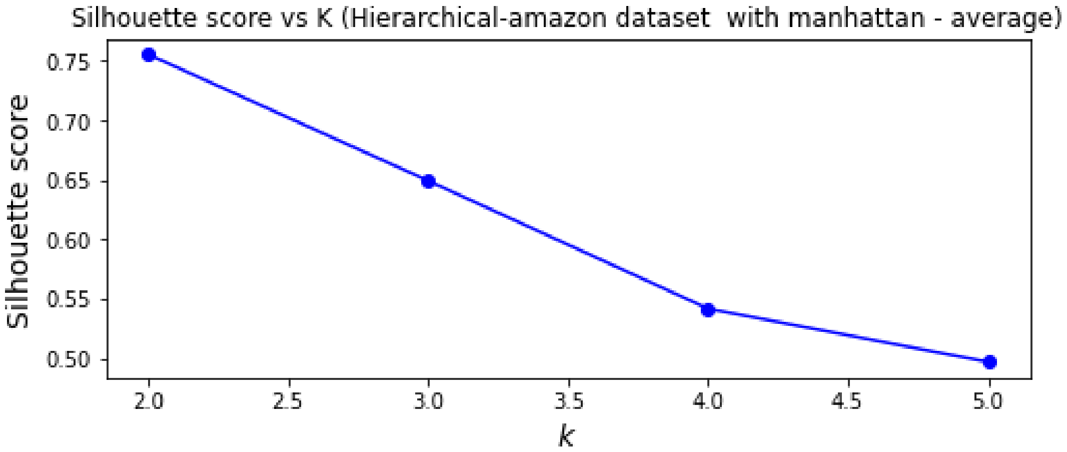
Also, the kmean with the Euclidian distance yield a good result.

5.2. Multi-Domain Sentiment Dataset The best model for this dataset was

The best model was Hierarchical Cluster

Number clusters=2 affinity=Manhattan

linkage=average (on the dataset PCA applied)



Silhouette score=.75

Davies-Bouldin=.41

Calinski and Harabasz= 272.11

Also Spectral Clustering with n_clusters=2,affinity='nearest_neighbors' yield a good result