**Image Clustering**

Name: Sahil Shrikrishna zele

Net id: szele

Miner username: Sahilzele

G no: G01465963

Approaches. And methodologies:

1. Data processing: As the dataset had many white pixels (0 values), Tried working with PCA, TSNE, UMAP for dimensionality reduction. TSNE and UMAP both used better on nonlinear datasets and provided better results than PCA for this dataset. Tried gaussian filter to blur the UMAP components and smoothen the clustering process. Initially used sigma = 1, which smoothened a lot and therefore distorted the result. Reduced the value to 0.2 which gave a little but non-negligible difference.
2. K-means Model:
3. Initializing the centroids: Initially assigned random values to the centroids and calculated the difference between data and randomly initialized centroids and update the centroids accordingly.

Pseudo code:

*Assign random values to centroids.*

*Loop from 1 to k centroids*

*Calculate distance from data to centroids.*

*Compare min (distance calculated above and initial value and update the centroid with new value*

*add updated centroids to the list.*

*return centroids.*

1. Assigning the data to centroids: calculated the distance between the data points and the centroids. Assigned the datapoint to the centroid having the minimum distance, updating the labels for the datapoints.
2. Updating the Centroids: updating the centroid values by taking the mean of datapoints assigned to the centroid and returning the updated the centroid values.
3. Calculating the silhouette scores: Silhouette score provides a good cluster metric using which we can determine how well the clusters are formed.
4. Selecting the best: Updating the centroids for certain number of iterations (max\_iters = 300 in our case) or until convergence. Running the process multiple times ( num\_runs = 20 in our case) , calculate the silhouette score for each run and update the values of ‘best\_labels’, ‘best\_centroids’ and ‘best\_centroid\_score’ when the silhouette score is better from previous run. Finally return the ‘best\_labels’ which will have our final labels.

Below Plot for showing silhouette scores calculated for value of k from 2 to 20.

