IT350 Assignment 4

NAME: SUYASH CHINTAWAR
ROLL NO.: 191IT109
TOPIC: CLUSTERING

Note:
1) The colab link has been attached below. After opening the link, if it opens in drive, click on “Open with Google Colaboratory” to view the complete code.
2) Only output screenshots have been attached along with the explanation. Code for the same can be found in the colab notebook.

Colab notebook link:
https://colab.research.google.com/drive/19_NKoh0maXYpYmCpK2EADL0OpUTsd3lk

Q1. Find the clusters in the given dataset based on the content similarity and image similarity using k-means clustering and hierarchical clustering methods. (Batch- 1 ---> Use First Question as a Dataset)

Pytesseract has been used for OCR of all the answer sheets. Clusters have been formed for the clusters given in the dataset.
TF-IDF vectorizer has been used to convert the obtained text from image to vectors which are then passed on to the KMeans clustering model for clustering.

The cluster folders that have been used for clustering are,

- Cluster_1
- Cluster_3
- Cluster_5
- Cluster_6
- Cluster_7

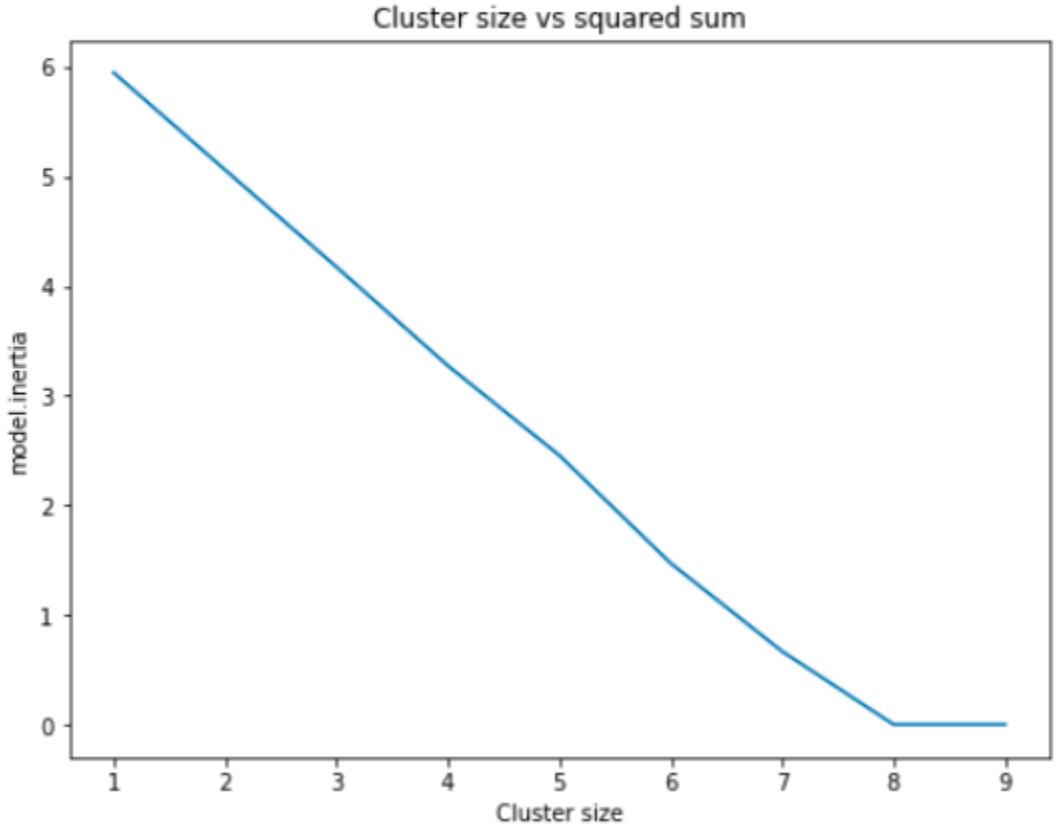
Outputs of one of the clusters (‘Cluster_1’) has been included in this report. The outputs for other clusters can be found in the provided colab notebook link.

The Elbow method has been used to find the optimal cluster size value (‘k’) by applying k-means over all the possible cluster sizes.

Output of Elbow method:

```
elbow_method('Cluster_1')
```

```
Cluster_1
```



We can see that for the first cluster the appropriate k is 8. So we run k-,eans with cluster size equal to 8.

Output of k-means:

continued…

```
[ ]  k_means('Cluster_1',8)
```

```
Cluster_1
***Obtained clusters are:***
cluster 0 :
 191IT110.1.jpeg
cluster 1 :
 191IT104.1.jpg
cluster 2 :
 191IT221.1.jpeg
cluster 3 :
 191IT141.1.png
cluster 4 :
 191IT220.1.jpeg
 191IT149.1.jpg
cluster 5 :
 191IT240.1.jpg
cluster 6 :
 191IT135.1.png
cluster 7 :
 191IT101.1.png
```

Similarly, run hierarchical clustering (agglomerative clustering in this case) with cluster size=8,

Output of hierarchical clustering:

```
[ ]  heirarchical_clustering('Cluster_1',8)
```

```
Cluster_1
***Obtained clusters are:***
cluster 0 :
 191IT220.1.jpeg
 191IT149.1.jpg
cluster 1 :
 191IT110.1.jpeg
cluster 2 :
 191IT135.1.png
cluster 3 :
 191IT221.1.jpeg
cluster 4 :
 191IT141.1.png
cluster 5 :
 191IT104.1.jpg
cluster 6 :
 191IT240.1.jpg
cluster 7 :
 191IT101.1.png
```

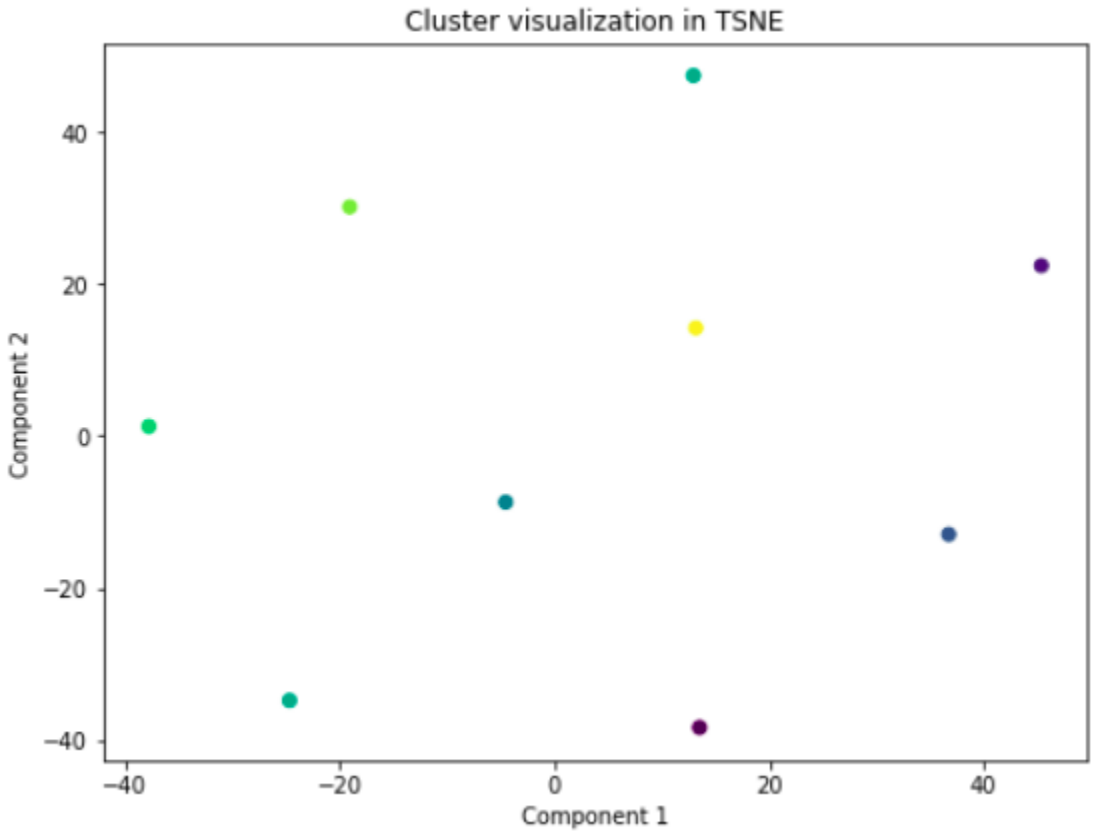
We can see that the same clusters are obtained in this particular case from k-means as well as from hierarchical clustering.

Q2. Plot t-SNE visualization for derived clusters.

For t-SNE, sklearn library has been used for the same derived clusters as earlier.

Output of t-SNE visualization:

```
Cluster_1
```



Outputs of other clusters can be found in the colab notebook link.

Q3. Evaluate the clusters that are obtained using appropriate methods.

The number of clusters obtained for each of the listed cluster is,

Cluster no.	No. of data points	No. of obtained clusters
1	9	8
3	13	13
5	38	29
6	7	7
7	16	11

Ideally, the number of clusters must be high signifying less similarity between answer sheets. Almost similar results have been obtained as shown in the above table.

THANK YOU