CS 57300 Data Mining Assignment 5

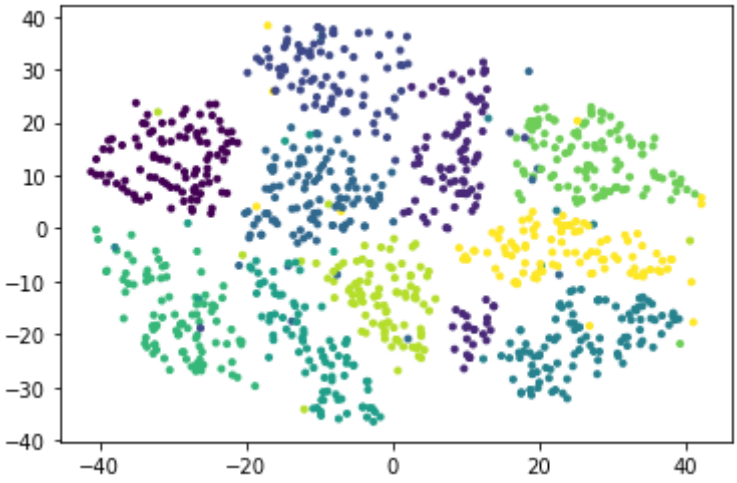
Shuang Wu (wu1716@purdue.edu)

Environment

See `README.md`. Generally, the Python we use with Poetry is ~3.9 for better type hinting.

1. Exploration

In [ ]:
```
%run -i exploration.py
```



2. K-Means Clustering

In [ ]:
```
DATA_FILENAME = 'digits-embedding.csv'
CLUSTER_COUNT = 10
```

2.1 Code

In [ ]:
```
%run -i kmeans.py {DATA_FILENAME} {CLUSTER_COUNT}
```

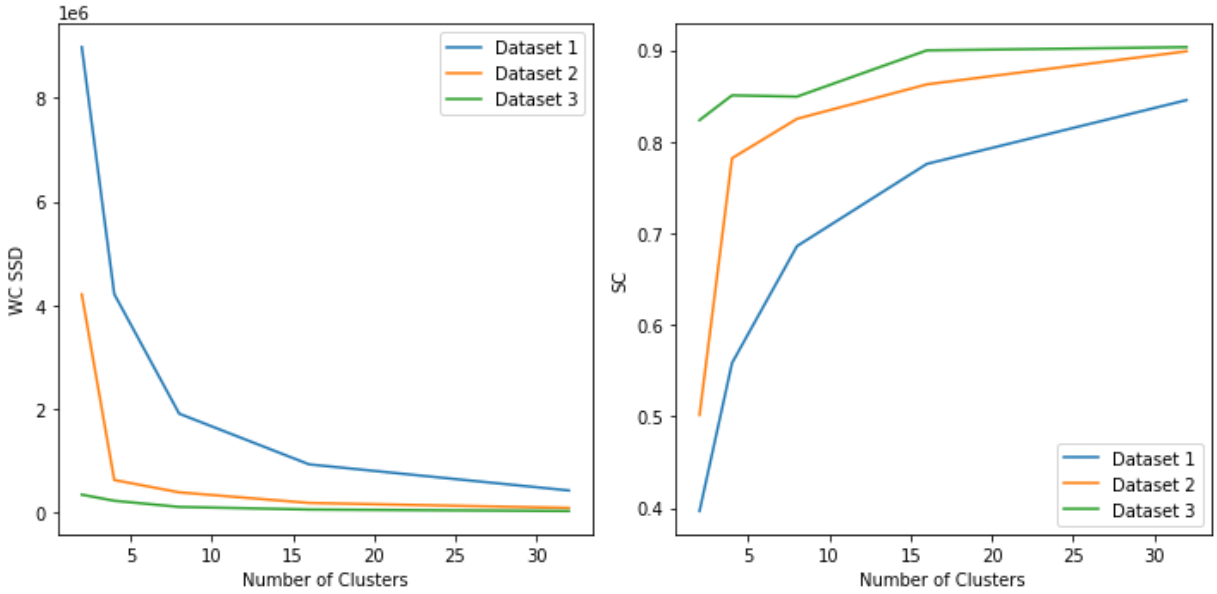
```
WC-SSD: 1460716.0038661587
SC: 0.7212120088665859
NMI: 0.384923508559809
```

2.2 Analysis

In [ ]:
```
PLOT_DIFFERENT_K = 0
PLOT_BATCH_DIFFERENT_K = 1
PLOT_CLUSTERS = 2
```

2.2.1

In [ ]:
```
%run -i kmeans-analysis.py {PLOT_DIFFERENT_K}
```

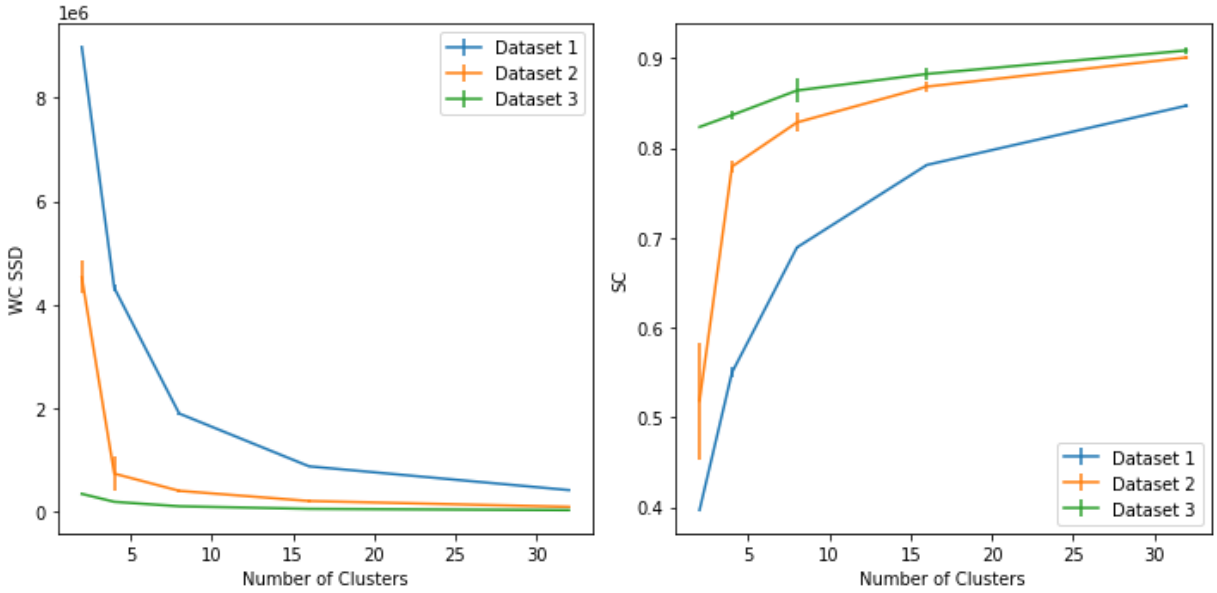


2.2.2

- For dataset 1, as there is no peak on the SC plot, we look for the "elbow" in the WC SSD plot and believe `k=16` is the one.
- For dataset 2, as there is no peak on the SC plot, we look for the "elbow" in the WC SSD plot and believe `k=4` is the one.
- For dataset 3, as there is no peak on the SC plot, we look for the "elbow" in the WC SSD plot and believe `k=8` is the one.

2.2.3

In [ ]:
```
%run -i kmeans-analysis.py {PLOT_BATCH_DIFFERENT_K}
```

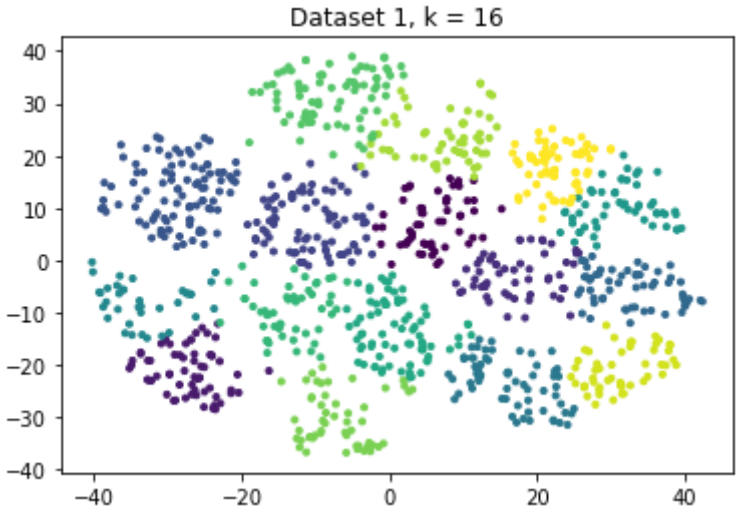


We find that k-means algorithm is sensitive to size of dataset. When the size of dataset increases, the variance of both indices decreases. For example, the variance of both WC SSD and SC with dataset 2 and 3 are larger than those with dataset 1. Apart from that, it is also sensitive to the number of clusters, `k`. When the `k` increases, the variance of both indices also decreases.

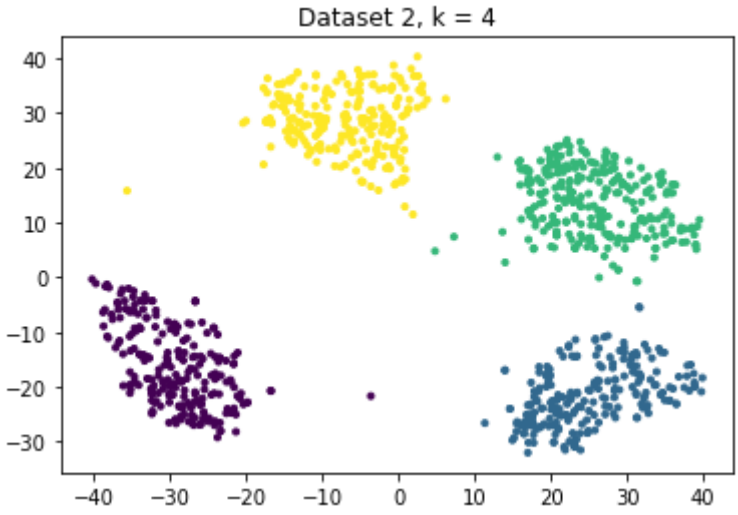
2.2.4

In [ ]:
```
%run -i kmeans-analysis.py {PLOT_CLUSTERS}
```

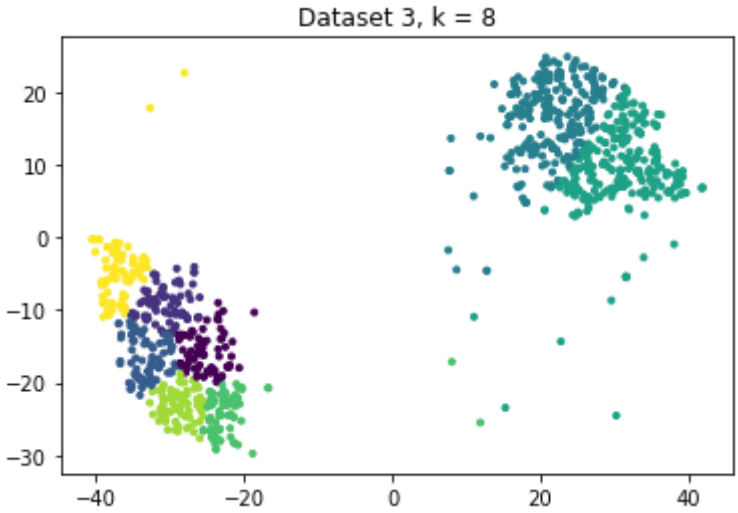
```
Dataset 1 NMI: 0.37681900542139835
```




```
Dataset 2 NMI: 0.45465341281006216
```



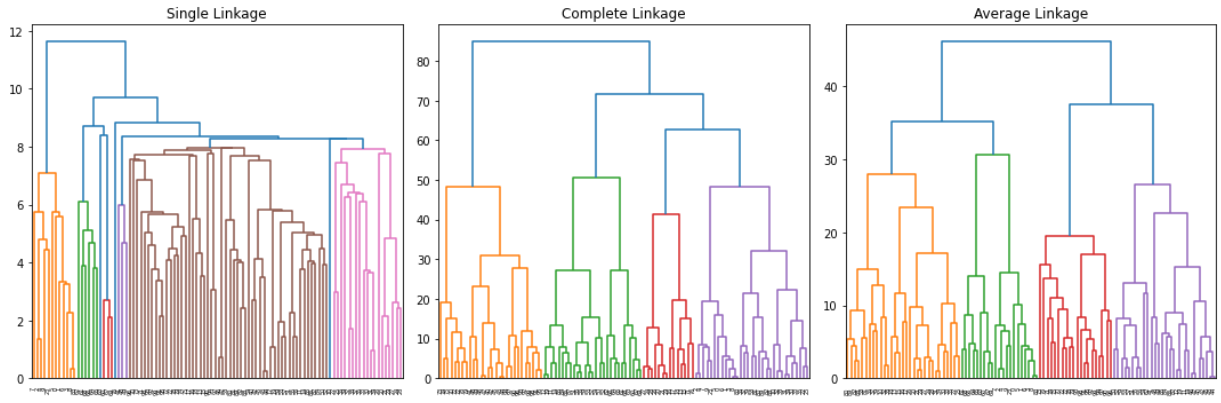

```
Dataset 3 NMI: 0.26015317053944664
```

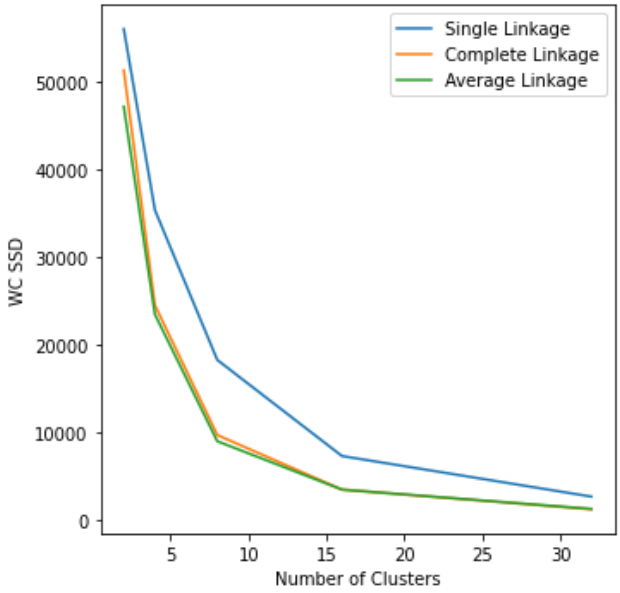


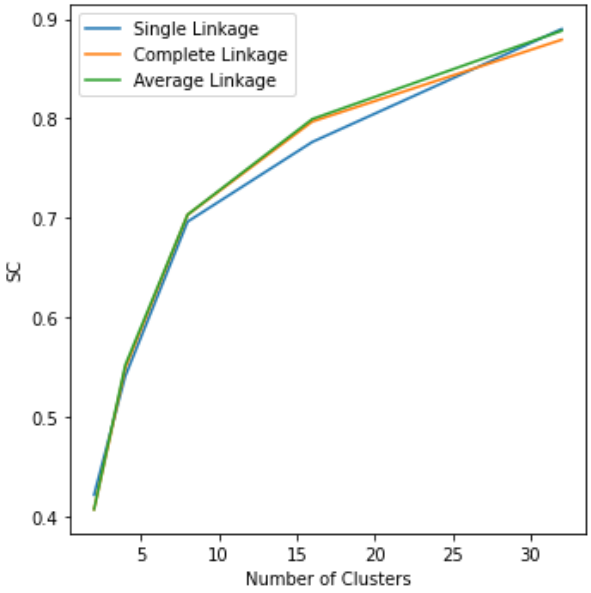
The dataset 2 with `k=4` performs the best among these three sets, which is corresponding to the highest NMI score. In comparison, the dataset 3 with `k=8` results the worst, which is corresponding to the lowest NMI score.

3. Hierarchical Clustering

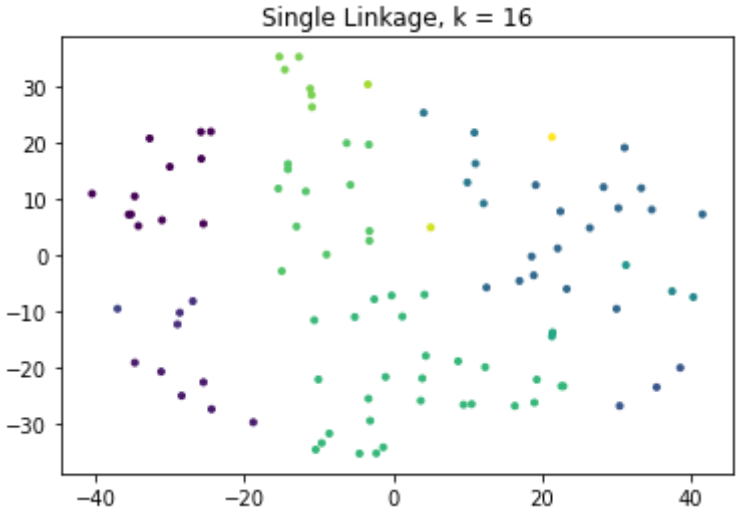
In [ ]:
```
%run -i hierarchical.py
```



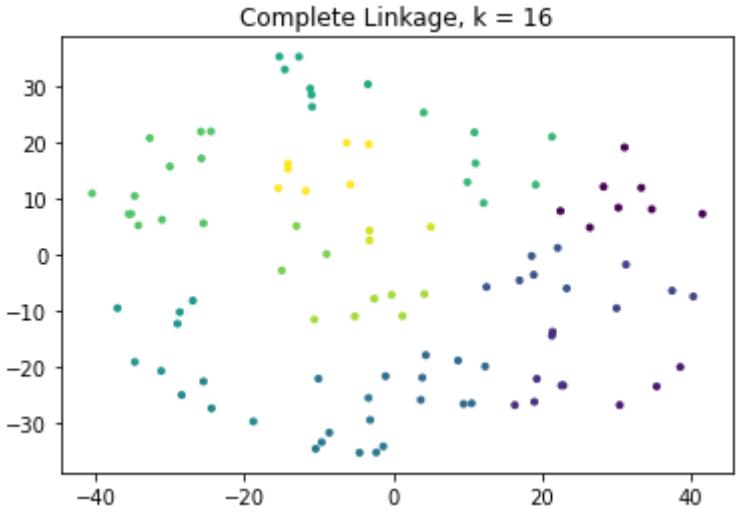




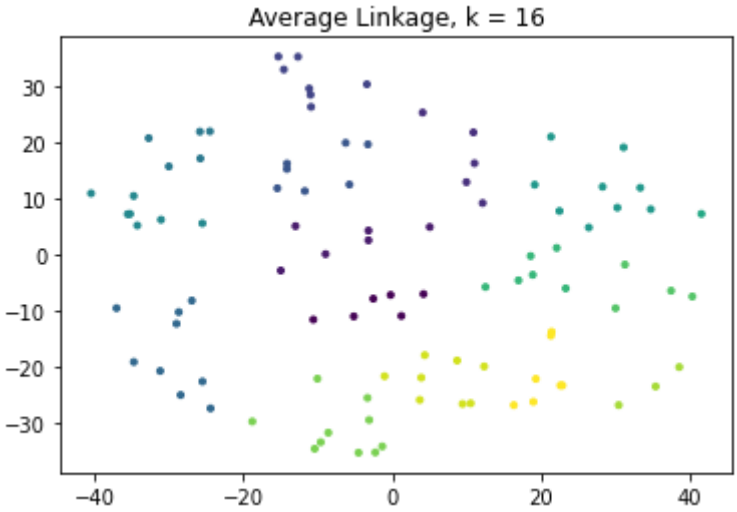
Single Linkage NMI: 0.36162963648828295



Complete Linkage NMI: 0.4097030095789396



Average Linkage NMI: 0.41163603264689924



According to the WC-SSD and SC plots, we would choose `k=16` for all three linkages as the "elbows" on the WC-SSD plots for them are at 16 without peak on the SC plots. Three hierarchical results with different linkage methods share similarity with the result from k-means analysis with dataset 1. We choose 16 as the number of clusters for all four models. Besides, in comparison to the k-means analysis with dataset 1, all results from hierarchical clustering are slightly worse, according to the NMI.