FINE TUNING OF PARAMETERS FOR IMPROVED CLUSTERING OF NEIGHBORHOODS USING KMEANS

Capstone Project

Dr. José Antonio Díaz Ávila

INTRODUCTION/BUSINESS PROBLEM

- In week 3, a project about clustering was assigned.
- The results were not satisfactory:
  - The distribution of the clusters was not even.
  - Clusters 1 and 3 occupied different areas in the map.
  - Cluster 5 covered almost the entire area of the map.

INTRODUCTION/BUSINESS PROBLEM (cont'd)

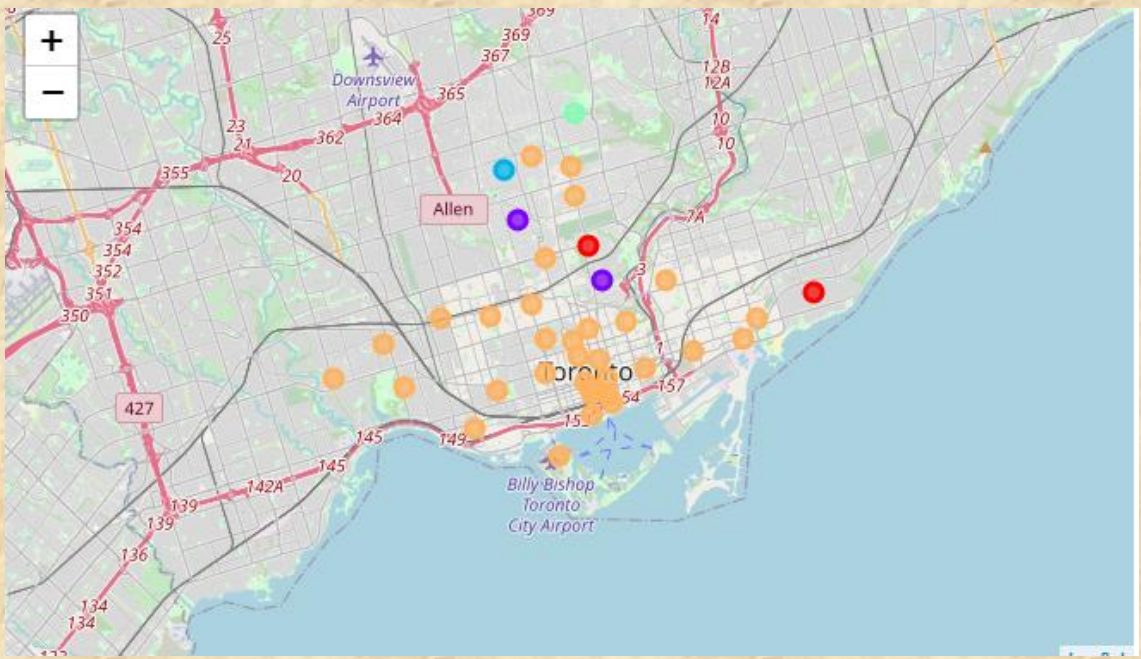

Figure 1. Clustering of 39 Toronto neighborhoods using a radius of 500 m for exploration.

Cluster 1: 2.
Cluster 2: 2.
Cluster 3: 1.
Cluster 4: 1.
Cluster 5: 34.

Table 1. Number of neighborhoods per cluster in figure 1.

DATA

- The data that were used to solve the problem were the following:
  - Toronto neighborhood information. This information was downloaded from the url: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.
  - CSV file with location coordinates for the Toronto neighborhoods. Downloaded from the url: http://cocl.us/Geospatial_data.
  - Map of Toronto provided by Folium.
  - Toronto venue information provided by Foursquare.

METHODOLOGY
Addition of the location information to the toronto_grouped dataframe

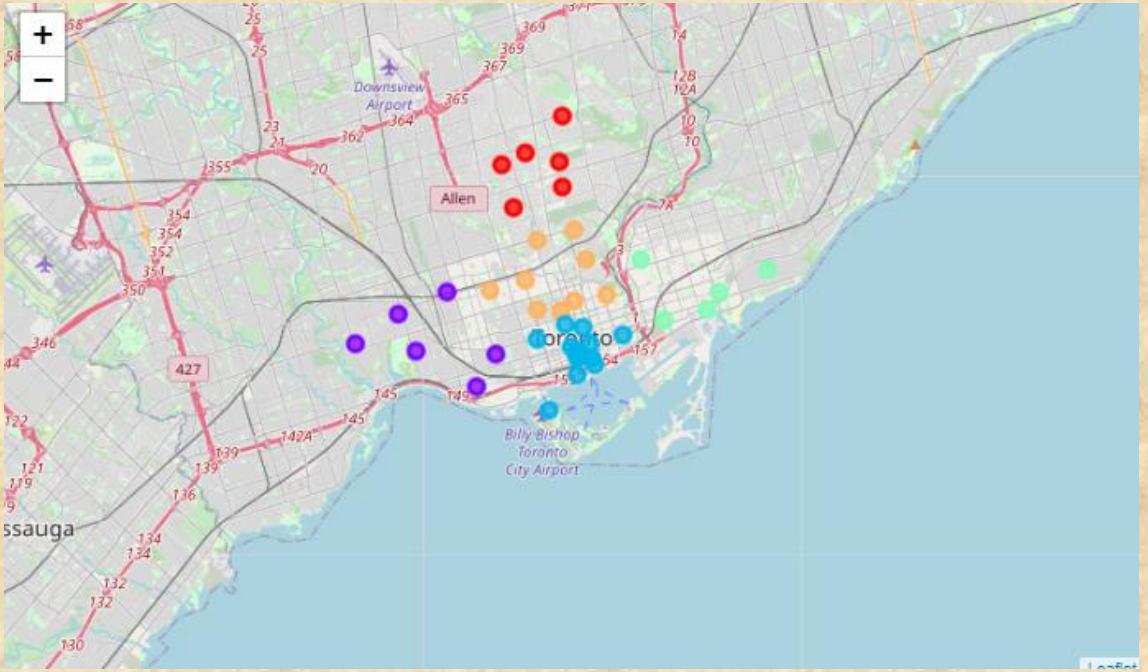

Figure 2. Effect of adding the location information to the Toronto_grouped dataframe.

[70]:

Cluster Labels	Neighborhood
0	6
1	6
2	13
3	5
4	9

Table 2. Distribution of the clusters in figure 5.

METHODOLOGY (cont'd)
Processing of Results of the One Hot Coding Technique

Figure 3. Distribution of the Toronto neighborhoods taking into account the number of venues per neighborhood.

Cluster Labels	Neighborhood
0	3
1	10
2	21
3	4
4	1

Table 3. Distribution of the neighborhoods in figure 2.

METHODOLOGY (cont'd)
Regrouping the venue categories

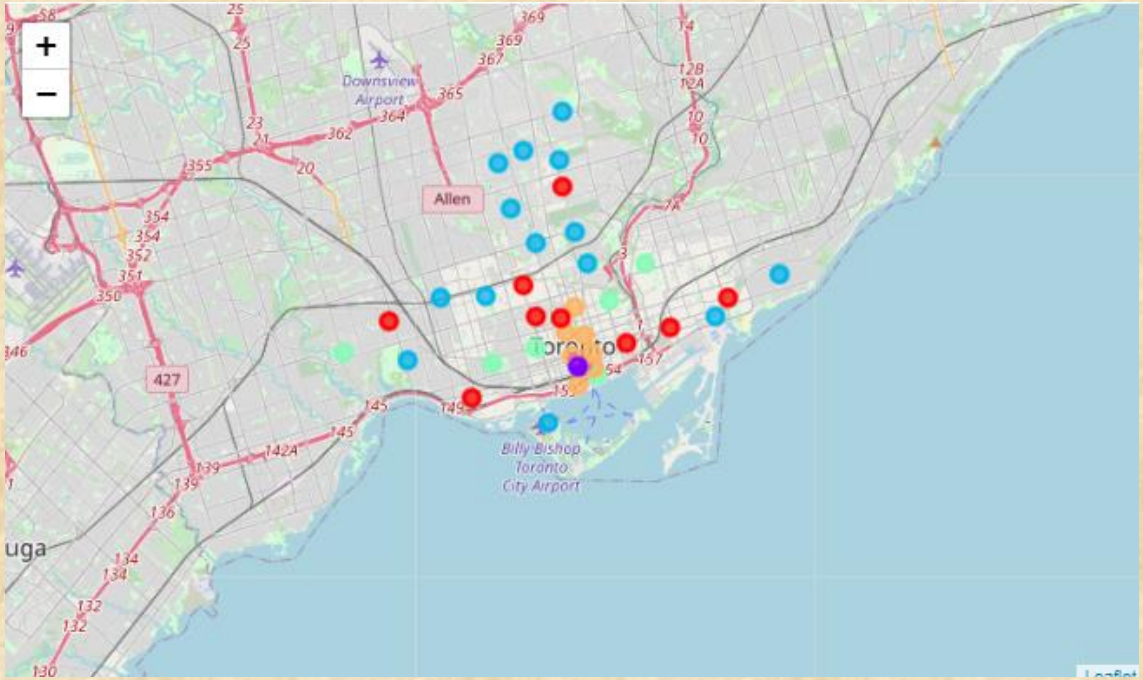

Figure 4. Distribution of the neighborhoods using 10 main venue categories.

Cluster Labels	Neighborhood
0	9
1	3
2	14
3	6
4	7

Table 4. Number of neighborhoods per cluster in figure 4.

METHODOLOGY (cont'd)
Calculation of the distance between neighborhoods

- The Harvesine formula was implemented.
- The minimum distance between two neighborhoods is 150 m.
- The average distance between two neighborhoods is 4.94 Km.
- Only five neighborhoods are separated by less than 500 m.

RESULTS

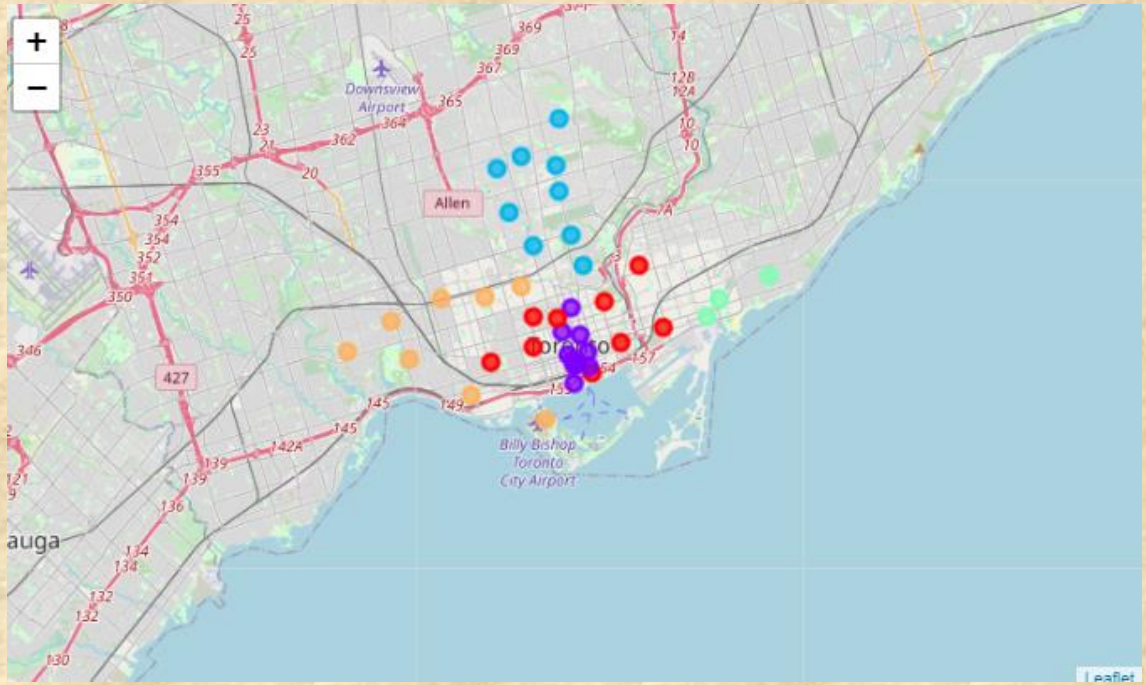

Figure 5. Clustering of the Toronto neighborhoods using a scale factor of 50.

Cluster Labels	Neighborhood
0	9
1	10
2	9
3	3
4	8

Table 5. Distribution of the neighborhoods in figure 5.

Cluster 0 (red): Central zone, shopping-recreational area.
Cluster 1 (violet): Downtown, business-recreational area.
Cluster 2 (blue): Northern zone, residence-recreational area.
Cluster 3 (green): Eastern zone, recreational area.
Cluster 4 (orange): Western zone, residential area.

Table 6. Labeling of the clusters

CONCLUSIONS

- The use of the location data helped cluster the neighborhhods in a single area.
- The number of venues per neighborhood was taken into account in the clustering process.
- The 237 venue categories were regrouped into 10 main venue categories.
- The distance between different neighborhoods was used to determine the radius.
- The neighborhoods were grouped into meaningful clusters.