FINE TUNING OF PARAMETERS FOR IMPROVED CLUSTERING OF NEIGHBORHOODS USING KMEANS

Introduction/Business Problem

In week 3, a project of clustering the Toronto neighborhoods was assigned. As a result of the project, Foursquare was used to obtain venue data, and Kmeans was used to cluster the Toronto neighborhoods. Figure 1 shows the results of clustering the Toronto neighborhoods in 5 groups using a radius of 500 m to explore the neighborhoods.

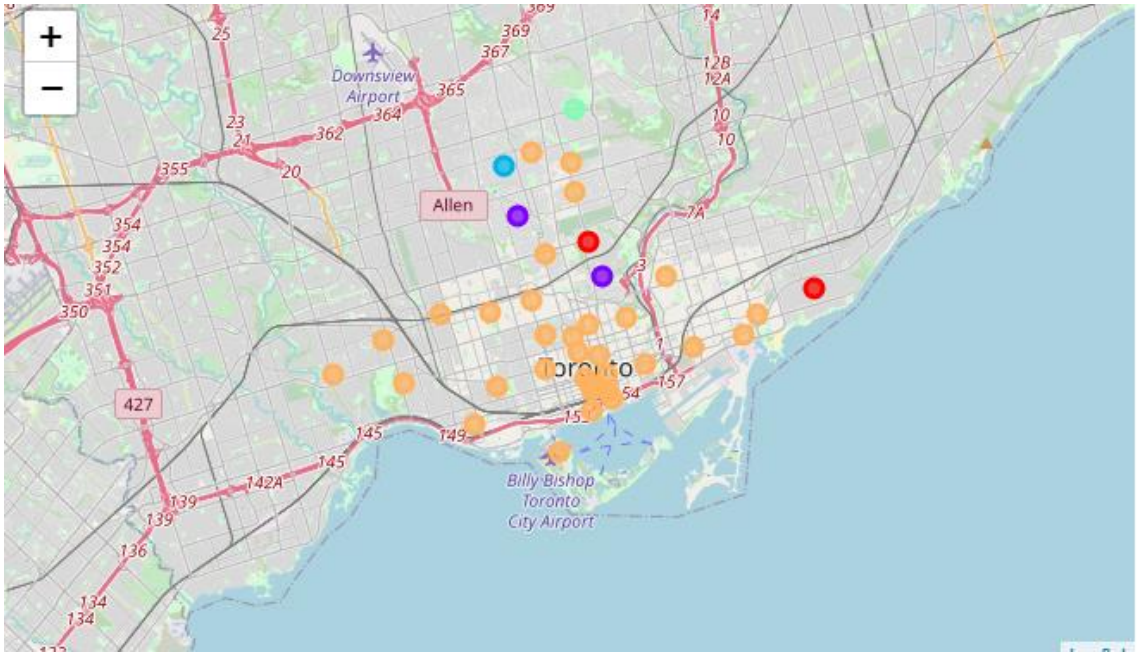

Figure 1. Clustering of 39 Toronto neighborhoods using a radius of 500 m for exploration.

Here, we observe several problems:

1- The number of neighborhoods is not evenly distributed, this is, the number of neighborhoods in each cluster is the following:
   a. Cluster 1: 2.
   b. Cluster 2: 2.
   c. Cluster 3: 1.
   d. Cluster 4: 1.
   e. Cluster 5: 34.
2- Clusters 0 and 1 with 2 neighborhoods occupy different areas in the map.
3- Cluster 4 covers almost the entire area of analysis. Only cluster 3 in the northern part of the Toronto, appears outside the area covered by cluster 4.

In order to solve this problem, the radius of the exploration data was changed from 500 m to 1000 m, yielding results shown in figure 2, which is a much better result.

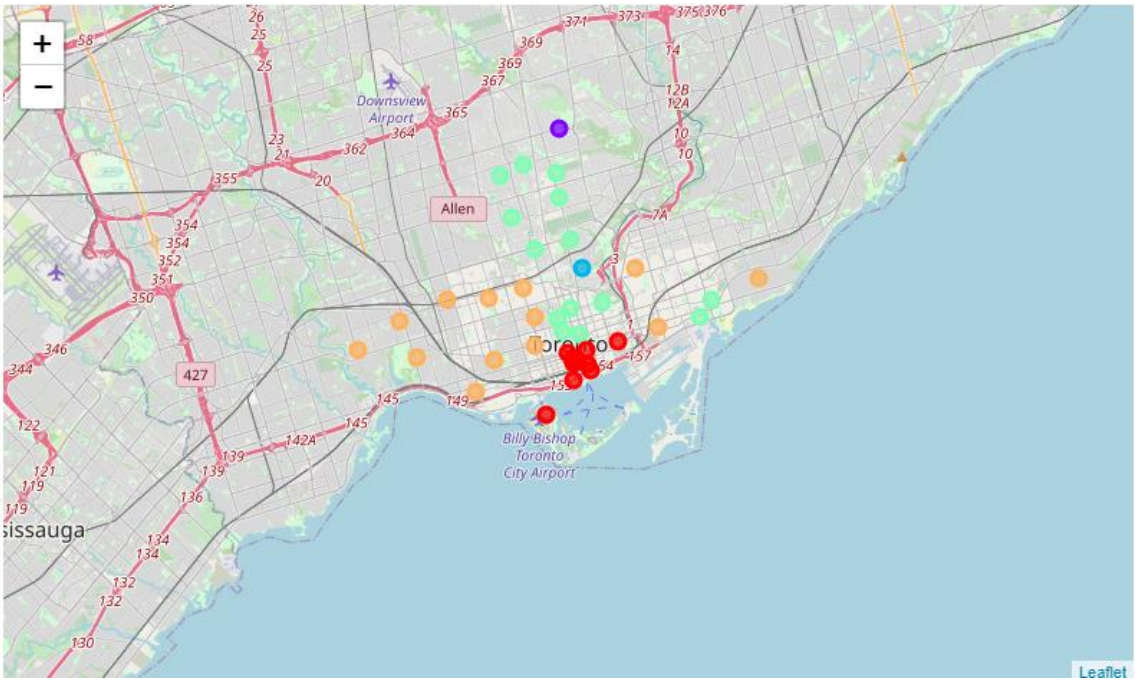

Figure 2. Clustering of 39 Toronto neighborhoods using a radius of 1,000 m for exploration.

From figure 2 we can make several observations:

1- The distribution of the clusters is more even. The new distribution is as follows:
   a. Cluster 1: 10 neighborhoods.
   b. Cluster 2: 1 neighborhood.
   c. Cluster 3: 1 neighborhood.
   d. Cluster 4: 14 neighborhoods.
   e. Cluster 5: 13 neighborhoods.
2- Clusters 0 and 3 appear as to be “clustered” in single regions, while cluster 4 appears to be split in two regions.

Even though the result in figure 2 is much better than the one in figure 1, one can ask why we have this issue. The main reason is that we did not provide the location information to the Kmeans clustering algorithm, therefore, the algorithm does not know where the neighborhoods are located, and makes the clustering based on the similarities among venues alone. Hence, it should not surprise us that the clusters are split in several regions as we are trying to display them in a map without providing the neighborhood location. This situation might bewilder the user as it happen to me. Hence, one of the objectives of this project is to improve the results of the kmeans algorithm by including the location data among the parameters to be passed to the algorithm, in order to make the clusters appear more concentrated in the map.

In order to test the previous hypothesis, an experiment was designed. In this experiment, only the location of the 39 Toronto neighborhoods shown in figures 1 and 2 were passed to the Kmeans algorithm, to perform the analysis. The results are shown in figure 3.

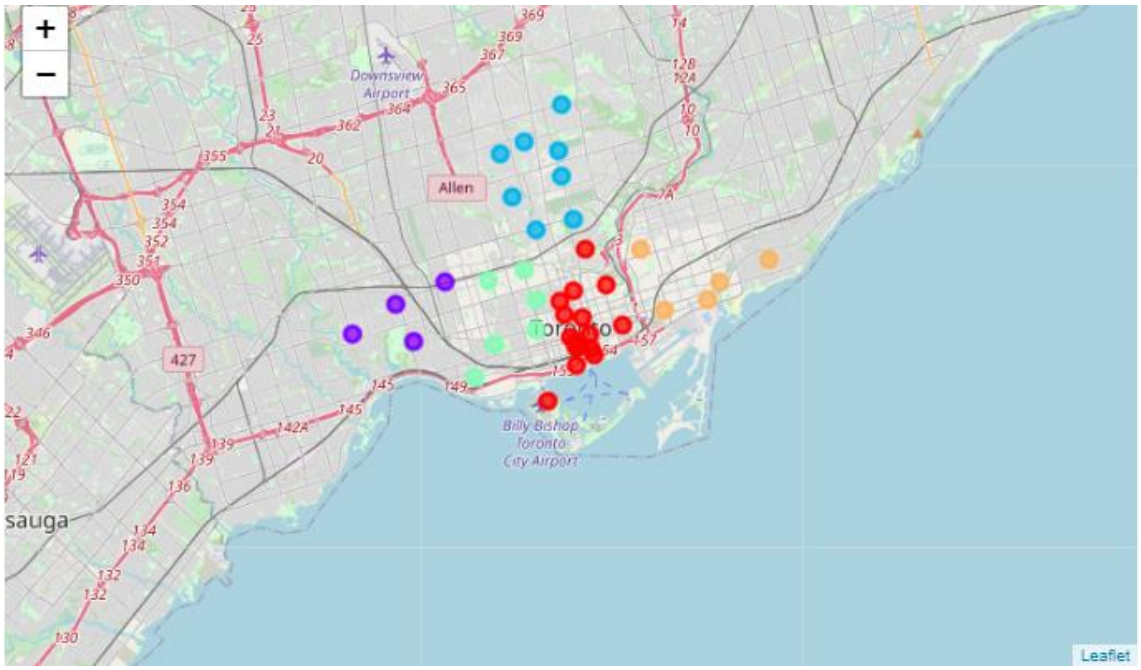

Figure 3. Clustering of 39 Toronto neighborhoods using location data only.

By analyzing figure 3, several observations can be made:

1- The clusters are more evenly distributed on the map, this is, the clusters appear to occupy similar size areas.
2- The number of clusters is more evenly distributed, having the following number of neighborhoods:
    a. Cluster 1: 16.
    b. Cluster 2: 4.
    c. Cluster 3: 8.
    d. Cluster 4: 6.
    e. Cluster 5: 5.

    Now, we do not have clusters with one neighborhood and the values among them do vary as much as in the previous two cases.
3- Each cluster occupies a single area, this is, the clusters are not split in two or more regions, which is what we would expect from a clustering algorithm

The results of this experiment shows that the location information can be used in combination with the venue data to fine tune the parameters of the kmeans algorithm. The issue/problem is that the results of the Kmeans algorithm as shown in figure 2, using only the venue data, do not provide a good clustering of the neighborhoods, even though a good amount of time and

effort was spent adjusting the parameters to provide the best possible clustering of the neighborhoods.

Another issue with the values passed to the Kmeans algorithm is the density of the venues in each neighborhood. The venue information passed to the Kmeans algorithm is in the Toronto_grouped dataframe. An analysis of this dataframe reveals that the sum of all the venues in one neighborhood equals 1. The problem is that each neighborhood has a different number of venues and the sum of all the venues for each neighborhood is equal to one. Therefore, a neighborhood with two restaurants and two parks and another with one restaurant and one park will appear with the same values in the Toronto_grouped dataframe and therefore will appear equal to the Kmeans algorithm, while the reality is that one of them has double the amount of venues, this is what I call the density issue. I think this problem might be solved by not normalizing the sum of all venues in one row to one, which will allow us to make a better clustering of the neighborhoods.

Another issue that was observed while developing the assignment for week 3 is that the one hot coding technique assigns a number 1 to a venue, if present in that neighborhood, no matter what the type of venue is. For example, if a restaurant exists, a number one will be assigned. The problem is that the restaurant category has been subdivided in many different subcategories. As a result, we may have 20 types of restaurants, and 1 type of park, this causes the results of the Kmeans algorithm to be biased towards restaurants. For example, a neighborhood with 10 different types of restaurants and one bank will appear very similar to a neighborhood having 10 different types of restaurant and bank, while it may be fairer to say that neighborhood one has restaurants and banks and neighborhood two has restaurants and no banks. Let us look at table 1 to illustrate this point.

28	Runnymede, Swansea	Coffee Shop	Café	Sushi Restaurant	Italian Restaurant	Pub	Pizza Place	Yoga Studio	Smoothie Shop	Bookstore	Sandwich Place
29	Moore Park, Summerhill East	Restaurant	Gym	Trail	Summer Camp	Department Store	Event Space	Ethiopian Restaurant	Electronics Store	Eastern European Restaurant	Donut Shop
30	Kensington Market, Chinatown, Grange Park	Café	Coffee Shop	Bakery	Mexican Restaurant	Dessert Shop	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Bar	Gaming Cafe	Park
31	Summerhill West, Rathnelly, South Hill, Forest...	Coffee Shop	Pub	Liquor Store	Supermarket	Sushi Restaurant	Vietnamese Restaurant	Bagel Shop	Light Rail Station	American Restaurant	Pizza Place
32	CN Tower, King and Spadina, Railway Lands, Har...	Airport Service	Airport Lounge	Airport Terminal	Harbor / Marina	Bar	Plane	Rental Car Location	Sculpture Garden	Boat or Ferry	Boutique
34	Stn A PO Boxes	Coffee Shop	Café	Restaurant	Cocktail Bar	Japanese Restaurant	Italian Restaurant	Beer Bar	Seafood Restaurant	Park	Bakery
35	St. James Town, Cabbagetown	Coffee Shop	Café	Restaurant	Italian Restaurant	Pub	Bakery	Pizza Place	General Entertainment	Sandwich Place	Butcher
36	First Canadian Place, Underground city	Coffee Shop	Café	Hotel	Gym	Restaurant	Japanese Restaurant	Asian Restaurant	Salad Place	Seafood Restaurant	Deli / Bodega
37	Church and Wellesley	Coffee Shop	Sushi Restaurant	Japanese Restaurant	Restaurant	Gay Bar	Café	Pub	Men's Store	Mediterranean Restaurant	Hotel
38	Business reply mail Processing Centre, South C...	Light Rail Station	Yoga Studio	Garden	Skate Park	Burrito Place	Farmers Market	Spa	Fast Food Restaurant	Butcher	Restaurant

Table 1. Top 10 most common venues for the first cluster of neighborhoods in the Toronto area (radius = 500 m).

If we look at neighborhoods 32 (CN Tower King and Spadina…) and 34 (Stn A PO Boxes), we notice that their most common venues are completely different. For neighborhood 32, the

most common venues are related to transportation services, while for neighborhood 34, the most common venues are restaurants and a park. This is, neighborhood 32 seems to be a commercial area, and neighborhood 32 is very likely a residential area. The question is, why are they in the same cluster? Why do the other clusters have only 2, 1, 1, and 1 neighborhood and the first cluster has 34 neighborhoods? The answer might be in the large number of subcategories in some categories (like restaurants), that seems to bias the results of the Kmeans algorithm. While I am not 100% positive that this is the reason why neighbohoods 32 and 34 are in the same cluster, I am confident that we are losing information when we split a category into many different subcategories, and distort the results of the Kmeans algorithm.

Finally, another aspect to be considered is the radius of the neighborhoods for venue data collection. This parameter had a large impact in the results of the Kmeans algorithm. I believe that some study needs to be done to provide guidance on how to choose this parameter. For example, at first, a radius of 500 m was chosen and Kmeans returned clusters with 34, 2, 1, 1 and 1 neighborhood. When using a radius of 1000 m the numbers were 10, 1, 1, 14 and 13 neighborhoods, a much better result. This value should not be chosen too small because we would not get enough venue information, and should not be too large, because we might have some overlap with other neighborhoods. From this reasoning, we conclude that the neighborhood location and separation between neighborhoods is of importance and should be taken into account in order to make a good selection of the radius.