New restaurant in Toronto

Wojciech Stasiuk

August, 2020

Abstract

Owner of restaurants in Bronx, New York needs to find which neighborhood in Toronto is most appropriate to start new restaurant. Using k-mean clustering and info about venues from Foursquare API, most similar Toronto's neighborhood will be choosed.

1 Business Problem

Owner of many successful restaurants in New York's city borough Bronx, wants to start new restaurant in Toronto. Stakeholder needs to know in which neighborhood of Toronto he should settle his first restaurant, that's why some method to comparison those two cities are needed.
My approach is to characterize each neighborhood of New York by most common venues type in each of them. Then to cluster the neighborhoods with greatest similarity using k-mean clustering , and then see which of them occurs the most in Bronx borough.
Next step will be to describe Toronto's neighborhoods in the same manner. Comparing clusters between two cities will tell neighborhoods of which cluster are the most appropriate to start new restaurant.

2 Data

There are two sources of data, first is names of neighborhoods, that will be taken from Wikipedia. Then, using Google geocoder API coordinates will be provided. Following step is to get 100 venues for each neighborhood using Foursquare API and then, to see which venue category occurs the most often. This data will be used to prepare clustering.

3 Methodology

First step is to assign 100 venues in radius of 500 meters from each neighborhood coordinates. Next step is to OneHot encode each venue based on it's category, after that summing how many many times it occurs in data category

we got frequency of appearance. Data reprocessed that way are now able to by sorted by k-mean clustering.
Score used to define which number of clusters will be the best is Silhouette.The silhouette value is a measure of how similar an object is to its own cluster compared to other clusters. The silhouette ranges from 1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

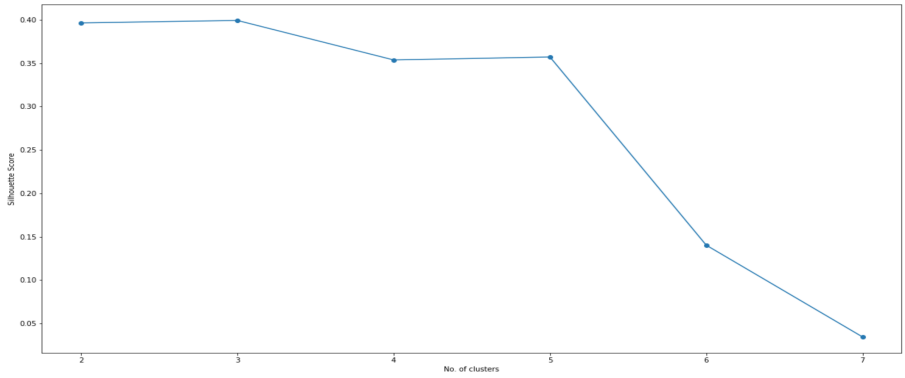

Figure 1: Silhouette score for different amount of clusters of New York's neighborhoods.

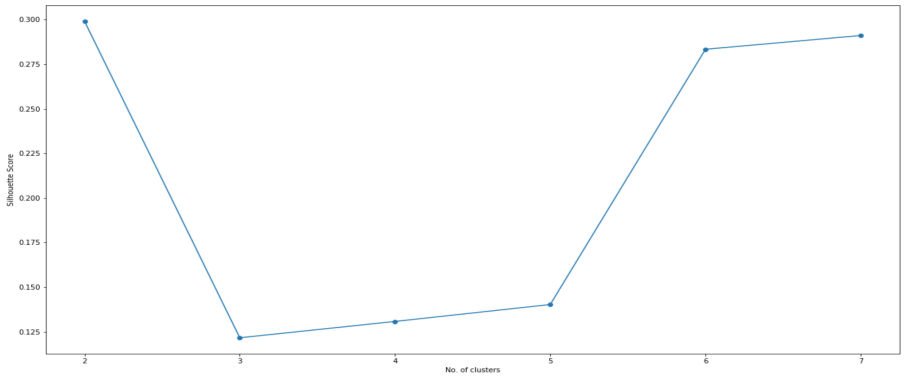

Figure 2: Silhouette score for different amount of clusters of Toronto's neighborhoods.

Figures above show that the best value of clusters of NY clusters is 3 and for Toronto is 2. After this step we would like to know which cluster occurs the most in Bronx borough, results are shown in table 1.

Now we can see that most of neighborhoods falls into cluster 1, that's why we will define cluster 1 by most frequent occurrences venues, data shows that most common venues are pizza places and coffee shops and now we will be looking for neighborhood's clusters in Toronto with the same pattern. Doing the same process for Toronto' neighborhoods we are able to see that cluster 1 has the same pattern.

Cluster Labels	Occurrence
1	51
0	1

Table 1: Clusters in Bronx borough.

4 Results

Comparison of neighborhoods shown that cluster 1(purple) shown in Figure 3 is most similar with cluster 1 from New York data. And that's the neighborhoods I would recommend to start new restaurant.

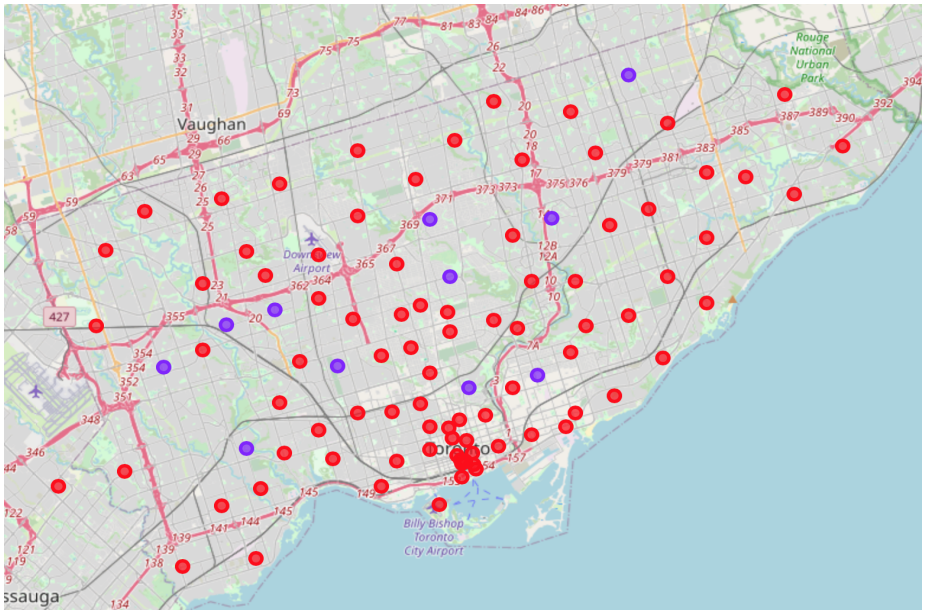

Figure 3: Map of clusters in Toronto neighbor hood cluster 0 (red) and cluster 1(purple).

5 Discussion

Results shows that it's plenty of possible neighborhoods to start new restaurant, similarity guarantees that environment is close to Bronx, thanks for that

project is more likely to succeed.

6 Conclusion

Greatest uncertainty lies in lack of data, because relying only on data of venues in neighborhood doesn't guarantee that neighborhoods are exactly the same, there are many other factors like for example income of citizens and so on.