Introduction

The development of technology, high-speed modes of transport in the 21st century makes it possible not to live permanently in one city and travel often. People travel, learn new cultures, move to other countries. In a trip, the similarity of his home with his places of residence does not matter for a person. However, if a person is going to move to another city or another country, it is very important that the area in which he lives is similar to the area from which he moved (especially if it is a long business trip). You can segment various places in the neighborhood according to the category of the object, and then group neighborhoods that include similar view of the surroundings. In this project, areas similar to each other in the cities of New York and Toronto will be found.

Problem

Finding identical neighborhoods in different cities to help ensure quick adaptation to neighborhood to make a decision choosing an area that is far away but still feels at home.

Data Sources

In the project used datasets: New York’s different neighborhoods and their respective geometric coordinates, Toronto’s different borough and their respective postcodes. Links for datasets:

<https://cocl.us/new_york_dataset>

<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

Geographic coordinates in the Foursquare platform will be added to the data from the sources.

Methodology

The goal of this project is to group together the similar neighborhoods in the city of New York and Toronto. We will use k-means because our data is unsupervised. k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

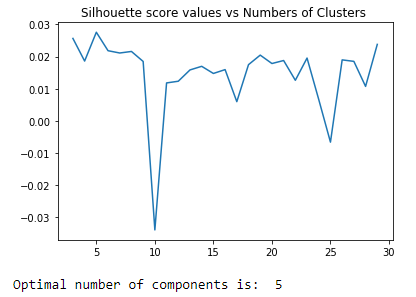
Results

**Silhouette Score:** This is a better measure to decide the number of clusters to be formulated from the data. It is calculated for each instance and the formula goes like this:

**Silhouette Coefficient = (x-y)/ max(x,y)**

where, **y** is the mean intra cluster distance: mean distance to the other instances in the same cluster. **x** depicts mean nearest cluster distance i.e. mean distance to the instances of the next closest cluster.

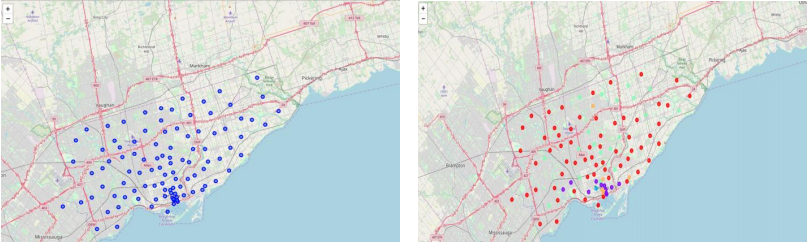
The coefficient varies between -1 and 1. A value close to 1 implies that the instance is close to its cluster is a part of the right cluster. Whereas, a value close to -1 means that the value is assigned to the wrong cluster.

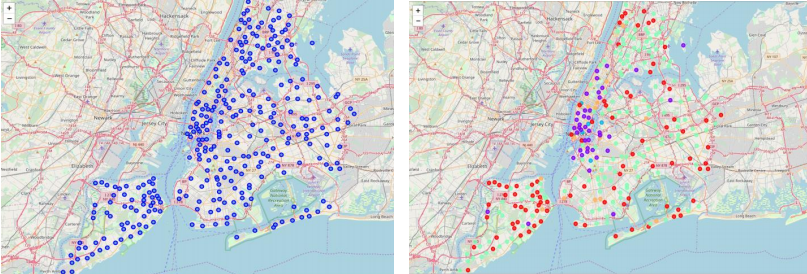


As per this method k=5 should be chosen for the number of clusters. This method is better as it makes the decision regarding the optimal number of clusters more meaningful and clear. But this metric is computation expensive as the coefficient is calculated for every instance. Therefore, decision regarding the optimal metric to be chosen for the number of cluster decision is to be made according to the needs of the product.

Visualizing the Clusters on the Map

Created folium maps to help obtain a visual perception of how the very different clusters look on the map when plotted on the map of the city of New York and Toronto.

On the left are places secured in Toronto before clustering, on the right are places secured in Toronto after clustering.



On the left are places secured in New York before clustering, on the right are places secured in New York after clustering.

Discussion

This project shows the operation of the machine learning method without a k-man teacher. Using the zip codes of New York and Toronto, Foursquare, we were able to obtain the geographical coordinates of the areas and conduct an analysis. In the course of work, 150 features were used and the optimal number of clusters was chosen - 5. Having more samples may result in a better clustering.

Conclusion

People are frequently moving into new cities. And in this ever growing world filled with technology, having a neighborhood recommendation based on location data is something to be considered basic now-a-days. And the application of neighborhood segmentation lies beyond this application too. This can serve to be an impressive tool to better organize a city resources. Furthermore, it can be used as a tool for security measurement if combined with crime data.