Similar Cluster in New York and Toronto.

Coursera Final Project

Irvin Fernando Guzmán González.

IBM Data science

Applied Data Capstone

12/20/19

Contenido

Introduction

It is supposed to be a commerce located on a certain neighborhood on New York city, and the owner wants to open another one on the city of Toronto, Canada. What he wants to know is which section of Toronto is like the one on New York. Recognizing which areas are similar would give an idea where to search for a place to establish the new commerce on Toronto and expect a similar behavior on the sells since both cities are big cities.

The main problem is to identify which areas are similar between New York and Toronto.

Data

The data to be used are the geolocalization of postal codes in Toronto, Canada and the geolocalization of the Neighborhoods on New York city. It is not a directly comparison between postal codes and neighborhoods areas, but it still gives us and idea of some areas of both cities. The geolocalization will be used to obtain the venues around each localization from the Foursquare database. Then a process of clustering using the venues as indicators between New York and Toronto will give us clusters of neighborhoods and postal codes similar between each other. The result would be used to see which area from Toronto are like the one on New York city where the commerce is located.

Methodology

Toronto Data

The data from Toronto was obtained from Wikipedia [1]. It contained each neighborhood in Toronto, CA with their respective Borough and Postal code. The data was group by postal codes because the data the data given by the Coursera Project contained the latitude and longitude of each postal code. A map of Toronto with each mark of the postal code was created using the Folium package on python.

With the use of the Foursquare API, the venues around 500 meters of each postal code was obtained. The venues data was treated and prepared to be ready to use on a k-means clustering process. Using 5-means clustering other map was created coloring each mark of the postal code as one cluster.

New York Data

The data from New York was given by the Coursera Project. It contained each neighborhood in New York city, US with their respective Borough, latitude and longitude. A map of New York with each mark of the neighborhoods was created using the Folium package on python.

With the use of the Foursquare API, the venues around 500 meters of each neighborhood was obtained. The venues data was treated and prepared to be ready to use on a k-means clustering process. Using 5-means clustering other map was created coloring each mark of the neighborhood as one cluster.

Merging Data

Toronto and New York data were merged into 1 data frame. The process to obtain the venues data was performed on the same way as the New York data. With the use of the Foursquare API, the venues around 500 meters of each neighborhood was obtained. The venues data was treated and prepared to be ready to use on a k-means clustering process. Using 5-means clustering other map was created coloring each mark of the neighborhood as one cluster. Then the results were analyzed.

Results

Maps of 5-means clustering on both cities separated

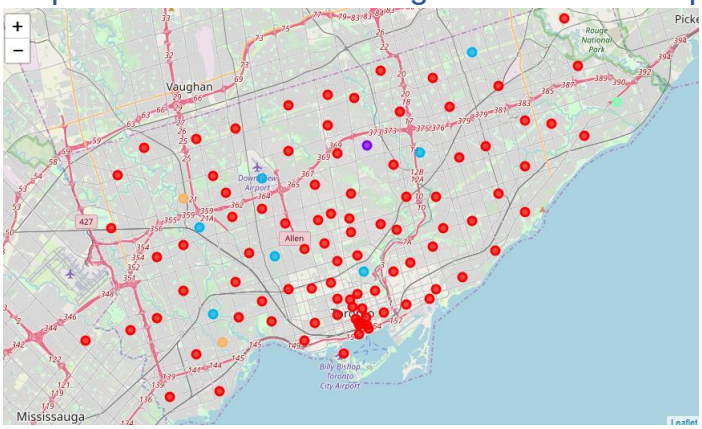

Figure 1. Toronto

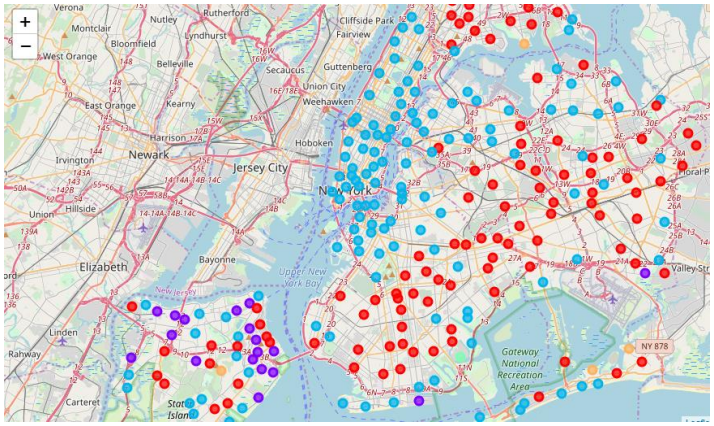

Figure 2. New York

Maps of both cities with 5-means clustering at same time

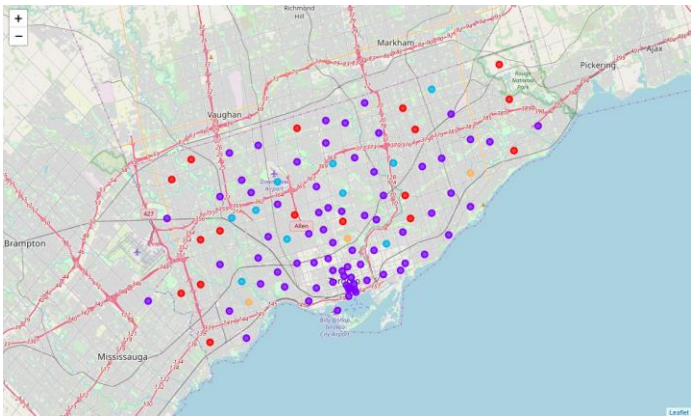

Figure 3. Toronto

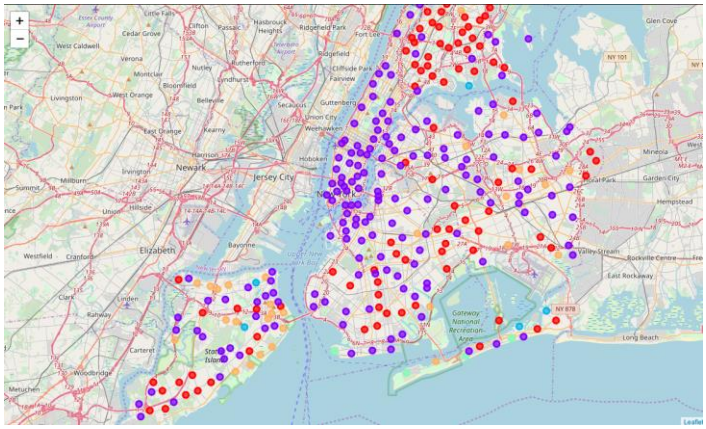

Figure 4. New York

Discussion

As it is observed on the Figure 1, Toronto has an evenly distributed cluster where de red ones are the most prominent and then the city has another sky-blue cluster evenly distributed. Also, there are 2 postal codes on orange color, 1 postal code on purple and another light green on the far northeast.

On the city of New York, Figure 2, it's very impressive how the cluster almost delimited the real boroughs of New York. The borough of Manhattan is totally cover with the sky-blue cluster. The borough of Bronx, Queens and Brooklyn are mostly covered in red with the nearest parts to Manhattan on sky blue and the borough of Staten Island has its own purple cluster.

It must be remarked that the clusters on Toronto are distributed across all the city, unlike New York where it seems to be more region localized.

When the process of clustering was performed with the data of both cities at the same time it is observed on the Figure 3 and 4 that both cities Toronto and New York almost remains equally distributed as it was before merge the data of both cities with a little differences between them. On the images can be observed that practically the clusters before and after the union of the cities are the same, but now the

color between the cities match. The purple cluster are who have the most quantity of venues and variety mostly represented by restaurants of different countries. The red cluster is more represented by fast food restaurants and pizza places. The orange cluster is mainly represented by the presence of deli / bodega, convenience store and playgrounds. The skyblue cluster mostly represented by parks and the green cluster was localized on the beach.

With this data can be observed that firstly the downtown of Toronto and all the nearby area of Manhattan are very similar, so if the commerce of the owner was localized on Manhattan would be good to locate the next commerce on downtown Toronto

Conclusion

In conclusion, with the obtained results, can be observed which similarities are between different area of Toronto and New York. It could help to decide where could be a good area to locate a new commerce taking reference from another one that is already establish and was successful. Therefore, if it is assigned a commerce on New York where is successful commerce just left to choose a region on Toronto that its assigned cluster match the color with the New York one.

To obtain more reliable result is recommendable to obtain another source of data for the venues on Toronto due to a difference on quantity of venues compared to New York. It is thought that is because a lack of information on the Foursquare database about Toronto, and not by an actual considerably difference of venues. Also is desired to design an algorithm or manual process to identify similar types of venues since there are venues on both cities that are very alike but are named differently.

References

[1] Wikipedia, "Wikipedia," 20 12 2019. [Online]. Available: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.