K-Means Clustering Analysis in New York and Toronto

Applied Data Science Capstone Project by M Indra Wicaksono

Introduction

The project main focus is to see what kind of information we can gain from the produced clusters and analyse it. The clusters should give us insight on how the city venues are growing on current trends and how it is gonna benefit entrepreneurs,investors and stakeholders to build a business in the city.

Data

First geospatial data is Names of borough, neighborhood, latitude and longitude for each city. New York data is provided from Week 3 Lab hands-on. For Toronto, data is extracted from wikipedia and geospatial Data for Toronto provided by Week 3 Assignment. Second geospatial data is venue data from the neighborhood of each city provided by the foursquare API by using the **Explore** endpoints.

Methodology

From existing latitude and longitude data, each city venue is extracted from Foursquare API for each neighborhood. The API is set to look for venues in a radius of 1500 meters from the point given and is going to return up to 100 venues from the radius. The result of API calls is JSON data containing each venue name and location coordinates. Subsequently the JSON data is transformed into a dataframe containing name, borough, neighborhood latitude, longitude and top 3 most common venues. The dataframe then is going to be modeled using K-Means Clustering with K=3. From the finished clustering results, we are going to analyse how the venues are formed in each city.

Results

The huge number of Coffee shops and Cafe really stands out more than other Food and Beverage (FnB) businesses around Toronto. Number of other kinds of FnB like Bar, Pub or ethnic-themed restaurant are abysmal compared to Coffee Shops. This trend may also be part of the third wave coffee movement that is emerging all around the world.

Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Park(6), Food and Drink Shop(1), Piano Bar(1)	Park(3), Bakery(1), Convenience Store(1)	Donut Shop(2), Pool(2), Yoga Studio(2)
1	Coffee Shop(20), Grocery Store(8), Cafe(5)	Cafe(11), Coffee Shop(9), Park(8)	Italian Restaurant(5), Electronics Store(4), Park(4)
2	Baseball Field(2)	Yoga Studio(2)	Drug Store(2)

Results

New York clusters also not evenly divided between clusters with 257 data points are labeled as the first cluster (label 0). New York Clusters are mainly dominated by Italian-themed businesses like Pizza Place, Italian Restaurant, and Deli/Bodega. As a city with a long history of immigration, the existence of a variety of ethnic-themed

Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Pizza Place(33), Italian Restaurant(26), Bar(17)	Pizza Place(19), Coffee Shop(17), Italian Restaurant(12)	Pizza Place(19), Bakery(14), Coffee Shop(13)
1	Deli/Bodega(12), *Beach(6),*Bus Stop(6), Italian Restaurant(5),*Bus Station(2), Carribean Restaurant(2)	Deli/Bodega(13), Chinese Restaurant(3), Grocery Store(3)	Deli/Bodega(6), American Restaurant(3), Bakery(3)
2	Park(3), Pool(1)	Yoga Studio(3), South American Restaurant(1)	Fish & Chips Shop(2), Fish Market(1), Grocery Store(1)

**From the 2nd cluster (Label 1) we can see the 1st most common venue is Beach and Bus Stop, which is peculiar for a venue. Because we focused on FnB venues I added two venues ranked right 4th and 6th from the data to make the analysis clearer.*

Discussion

Clustering using K-means with a lot of dimension is not really giving me a significant statistical inference from the findings. The imbalanced size of clusters happens in every K i tried, also with various number of common venues combination. I think there are better ways to cluster the venues instead using K-means.

From the findings I see that some foursquare API returns non-man made venues such as beaches but also *less important* venues like bus stops to their database. As a user-driven app, I believe that the number and type of venue data is biased from city to city. I suggest that there is a need to manually verify the location and type of venues from the internal division of foursquare to lessen the bias of data itself. Internal verification of datas would benefit foursquare and the customer much more because it adds accuracy and definitely trust to the company itself.

Conclusion

Toronto venues is packed by either Coffee Shop or Cafe. The number of coffee-focused venues is way dominant than other FnB venues such as restaurant and pub.The demand of coffee shops maybe is part of the emerging trend of Third wave coffee movement that taking place all over the world.

On The other hand, Manhattan has quite diverse collection of FnB venues with various Italian-themed venues in the top spot, specifically **Pizza Place**, **Italian Restaurant** and **Deli/Bodega**. There is also a big number of Chinese and Japanese Restaurant in the area. However from the clusters we can see that Coffee shop and Cafe also taking the top number of venues in the area, like we see in Toronto.

From those neighborhoods we can conclude that there are **similar** growing demand of **coffee shops** and **cafe** in both citiess.