Coursera Capstone

IBM Applied Data Science Capstone

Vancouver or Toronto

By: Chenhan Zhao

March 2020

Vs.

Introduction

For many people, they work their whole life to pursue their dream of working for a big company at a big city. They can have convenient shopping experience, world-class medical care and excellent school district for their kids. However, there are so many big cities which one do you want to go become a problem for many young people. Vancouver and Toronto are two of the biggest cities in Canada which contains 20% population in Canada, and they offer completely different experience for their citizens. One is on the Pacific coast surrounded by mountains and islands, the other is a major Great Lake city. As a result, property developers also taking advantage of their geological environment and built different kinds of venues around the city. The entertainment life contributes a lot to our overall happiness which ultimately affect our productivity. Therefore, choosing the right city that consistent with your lifestyle is crucial to lead a healthy and happy life.

In this report, we aim to use cluster method and venue information extract from Foursquare API to analyze the lifestyle of Vancouver and Toronto. By the analyzed results, we can provide city recommendation for people who are worried which city they should go.

Data

To solve the above problem, the following data are obtained:

- List of postal codes, borough and neighborhoods for both Vancouver and Toronto. These data confines that this project to the city of Vancouver and Toronto.
- Latitude and longitude for each neighborhood. This information is used to obtain the exact location of each neighborhood in order to plot the maps and get venue data.
- Venue data, especially data relate to lifestyles like restaurants, gyms, trails, etc. These data will be used to analyze the lifestyle of each city and each neighborhood.

Data source and extraction method

- List of neighborhoods are extracted from the Wikipedia page for Toronto (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) and Vancouver (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_V) . Web scraping methods are used to extract the tables from web page with Python package *pandas*. In order to only contain the central areas of each city, we only extract neighborhoods whose borough contains the word “Vancouver” or “Toronto”.
- Geological information is obtained from GeoNames website (http://download.geonames.org/export/zip/). The package contains all postal codes in Canada and their latitude, longitude is downloaded. This information is imported as *pandas* DataFrame and assigned to corresponding neighborhood.
- Venue information is imported through Foursquare API and after data wrangling, K-means clustering method will be used to analyze each neighborhood and their similarities.

Methodology

First, we need to obtain the list of postal codes and corresponding neighborhood for city of Vancouver and Toronto. This information is available on Wikipedia page Vancouver and Toronto. We use web scraping method with *Python* and *Pandas* library to extract the list of postal codes and neighborhood information for both cities. Because the page of Vancouver contains all cities in British Columbia province in Canada, we need to pre-process the data. Here, we only select postal codes with Vancouver as borough. For Toronto, we only select postal codes which has the word Toronto in borough. The next step is to obtain geographical coordinates for each neighborhood. Here we use the package on GeoNames website which has coordinates for all neighborhoods in Canada. We download the package as zip file, extract the data in csv format and import it to Jupyter Notebook as data frame in *pandas* library. Then we map the coordinates to each neighborhood with the unique postal code.

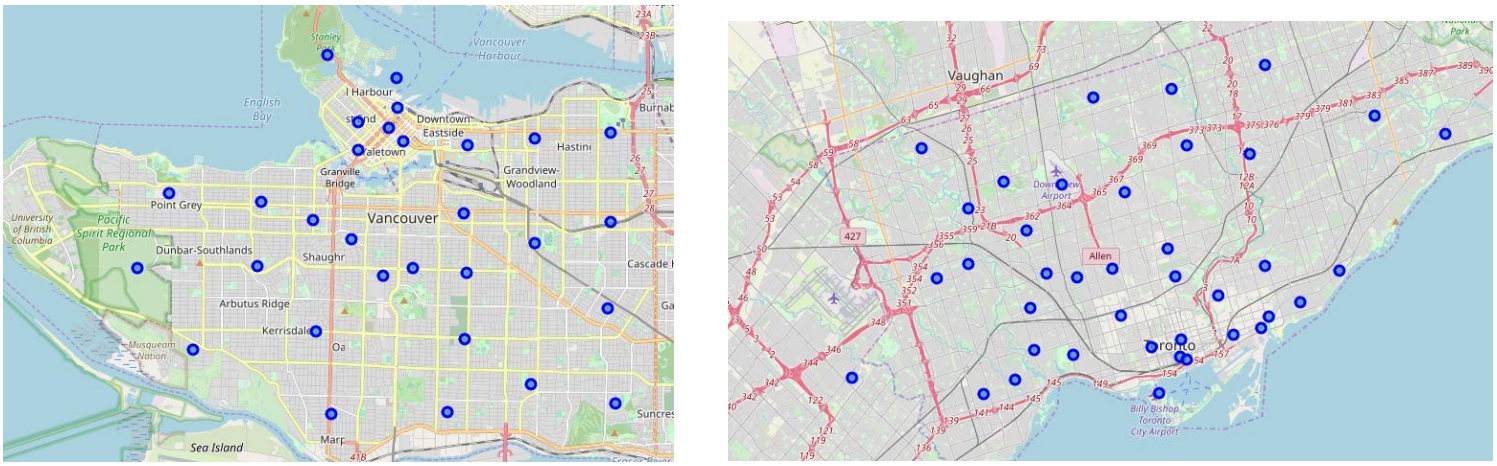
The next step is to acquire venue information for each city. Here, we use the Foursquare API to obtain the top 100 venues that are within a radius of 100 meters. The request is made through a Python function that passes the geographical coordinates of each neighborhood and receives the venue information. The venue information contains the venue location, venue category, etc. With the venue data, we can analyze the lifestyle for people living in these two cities. Next, we dig deep to each neighborhood. We assign each venue to their corresponding neighborhood and analyze each neighborhood. We group the data by neighborhood and calculate the frequency for each venue category. By doing so, we have prepared the data for clustering.

Last, we performed data clustering by k-means clustering method. We set the number of centroids (k) for each city as k = 5 and allocate every neighborhood to its nearest neighborhood. The result will show the similarity between each city and give suggestion for people with different lifestyles.

Result

Neighborhood Distribution in Vancouver and Toronto

By web scarping the information from Wikipedia and process the data, we obtained 31 neighborhoods in Vancouver and 39 neighborhoods in Toronto, respectively. As shown in Fig. 1, Vancouver has neighborhoods distributed more evenly on the map. However, Toronto has more neighborhoods around the Lake Ontario and less in the north side of city.

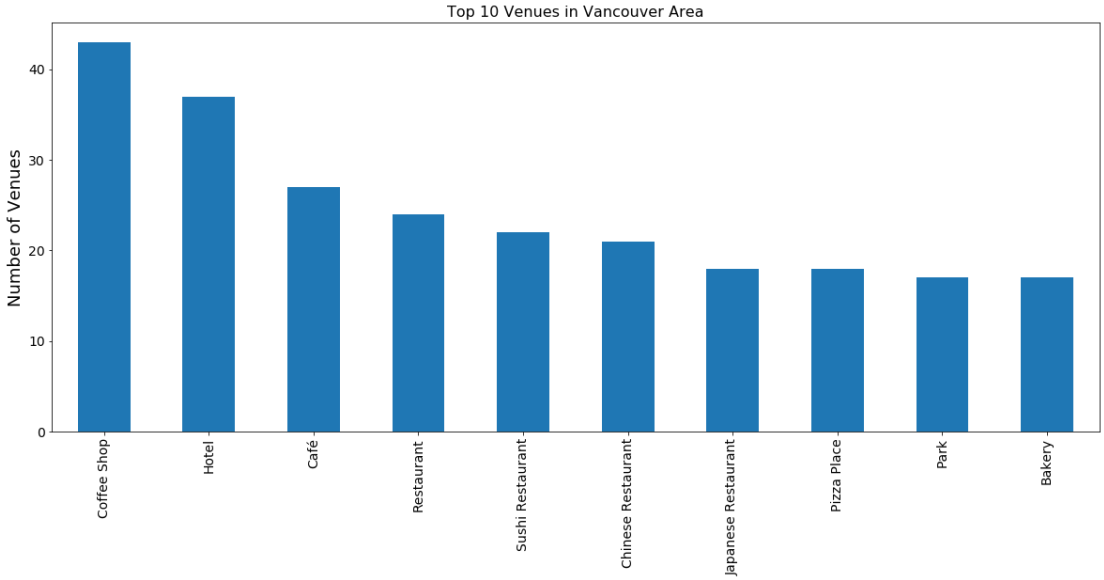


(A) **(B)**

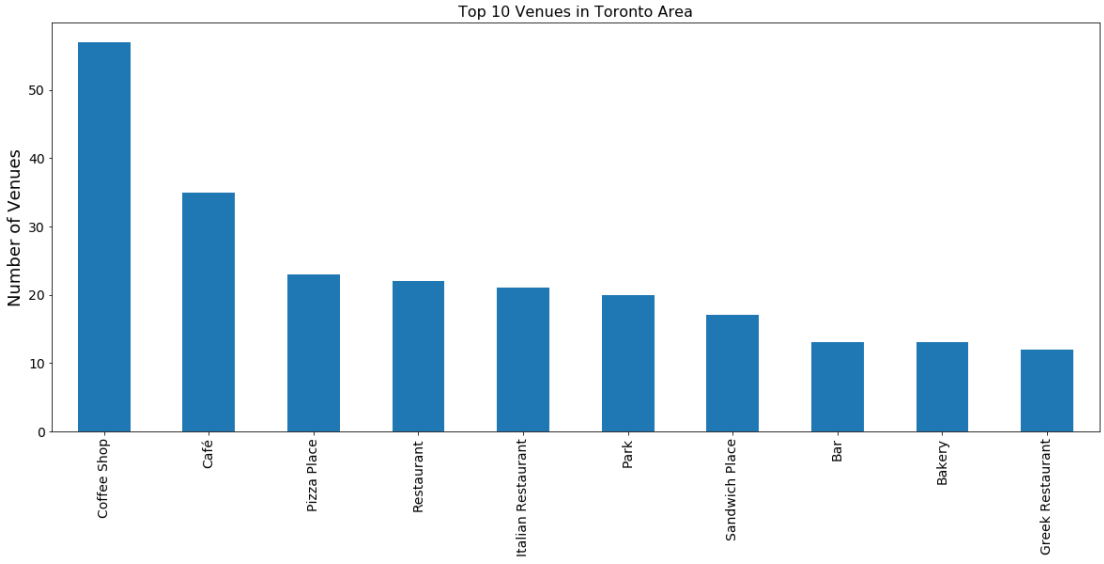
Fig. 1: Neighborhood distribution for Vancouver and Toronto. (A) Vancouver has neighborhoods all over the city. (B) Toronto's neighborhoods are mostly distributed in the south side of the city.

Venues in Vancouver and Toronto

Foursquare API is used to obtain the venue list for each city. According to the request list, Vancouver has 175 kinds of unique venue categories and Toronto has 201 kinds. As shown in Fig. 2, the most dominate category of venue in are coffee shops. And café in both cities are in top 3 kinds venues. However, the difference between the two cities is that hotels is second most kind of venue in Vancouver, but Toronto has pizza places in the top 3 list.



(A)

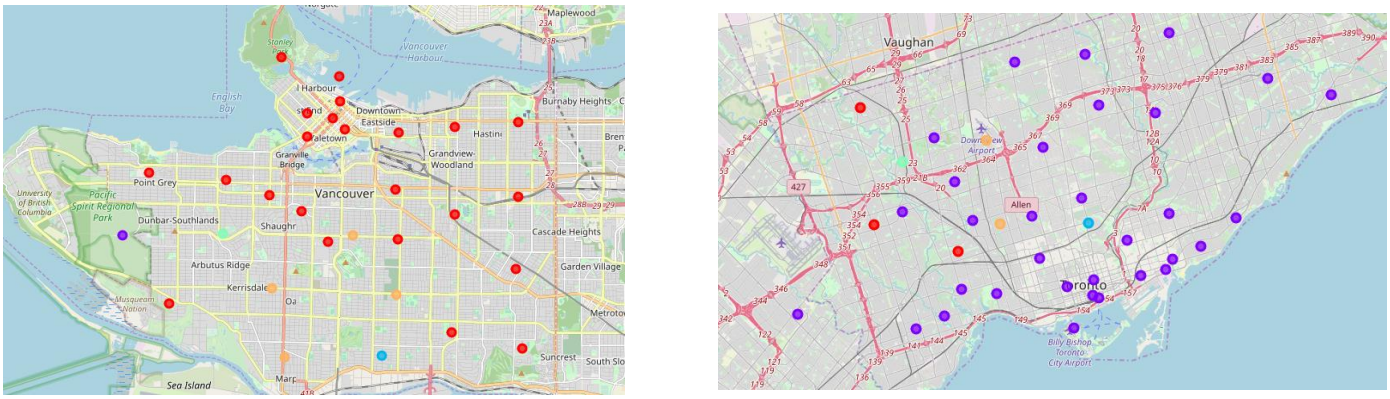


(B)

Fig. 2: Top 10 categories of venues in Vancouver and Toronto. (A) Vancouver has 2 non-food/beverage venues in top 10 and Asian food are very common. (B) Toronto's top 10 categories of venues are dominated by food/beverage related venues.

Neighborhood Cluster for Vancouver and Toronto

K-means clustering method is applied to data of both cities. The best number of centroids for Vancouver is 4 and Toronto is 5. Fig. 3 Shown the cluster result for both cities.



(A) **(B)**

Fig. 3: Neighborhoods after clustering for Vancouver and Toronto. (A) Vancouver's north neighborhoods share a same cluster, different neighborhoods mainly on the south side. (B) Toronto's neighborhoods gather at south side and different neighborhoods are around east area.

Discussion

Data Analysis

According to the result, Toronto has more choice of neighborhoods. Toronto has more kinds of venues than Vancouver. However, Vancouver has 25.51 venues per neighborhood and Toronto 19.89 venues per neighborhood. Each city has Coffee Shop as most common venue. Vancouver has 1.38 Coffee Shop each neighborhood and Toronto has 1.46. Vancouver has more hotels around neighborhoods and Toronto has more pizza place around. There are 3 Asian food restaurants in top 10 venues in Vancouver and Toronto has none. Vancouver has 0.54 park around each neighborhood and Toronto has 0.51. Vancouver has 0.16 gym around each neighborhood and Toronto has 0.31.

Recommendations

From the analyzed data, we give city recommendations based on people's lifestyle.

You should choose Vancouver if you like one or more following lifestyles:

- Don't care about venue type, just want to have a venue close to your neighborhood.
- Have visitors very often and want them to stay closer to your neighborhood.
- Like Asian food a lot and want to eat Asian food frequently.
- Like to go to parks that are close to your neighborhood.

You should choose Toronto if you like one or more following lifestyles:

- Don't want to go same type of venue every time, what to have more diverse experiences.
- Like western food than Asian food, especially like pizza.
- Want to have a Coffee Shop close to your neighborhood.
- Like to go to gyms that are close to your neighborhood

Conclusion

In this project, we analyzed the lifestyle of two of the biggest cities in Canada and give advice for people who are deciding which city they should go. Specifically, we extracted neighborhood information for both cities form Wikipedia using web scraping method. Obtained venue information around neighborhoods using Foursquare API. By calculating the frequency of each venue category and clustering the neighborhoods by most common venues, we conclude the lifestyle for each city. The findings of this project will help relevant stakeholder choosing their ideal work location while having a comfortable lifestyle.