IBM-COURSERA :
DATA SCIENCE CAPSTONE PROJECT

Canadian Cities Comparison

Abstract

In this final submission of the Data Science Capstone Project, a comparison between the neighborhoods of the Canadian cities of Montreal and Vancouver is presented. In order to carry out the proposed analysis, it was necessary to build the required data sets, use Foursquare's API and use the K-Means clustering algorithm to segment the neighborhoods based on their venues similarity.

Hugo Rocha De A.

Introduction

Moving to a different city can be quite a challenge as you require to take multiple decisions before you arrive to the final destination. Finding a flat or house located in an optimal location, is optimal to reduce the stress from moving to a new country/city and to ensure the proximity to basic venues (grocery stores, markets, bus stops, metro stations, touristic venues, etc.). Therefore, the importance of knowing the venues present in the neighborhoods that are being considered. The same situation happens when a user of a mobile application, such as Airbnb, has to choose a place to stay during a holiday, knowing the proximity to touristic venues can improve the host's experience.

This project is focused in the scenario where a student has been accepted to McGill University in Montreal and University of British Columbia in Vancouver, and wants to compare the neighborhoods close to both universities. In order to carry out this task, the neighborhoods will be compared based on the venues present in them. In this point of the analysis, K-Means clustering is used to segment the neighborhoods closer to the universities.

Data Description

Two data sets were bilt in order to carry out the proposed analysis. The two data sets contain information on the neighborhoods, boroughs, zip codes and geographical coordinates for both cities. The Montreal's data set contains 106 data points and the Vancouver's data set contains 66 data points. Using this data, it was possible to display on each city's map the geographical location of the neighborhoods. After this visualization, it was possible to select the neighborhoods closer to McGill University and University of British Columbia.

Montreal's Neighborhoods:

- Ville-Marie
- Le Plateau-Mont-Royal
- Outremont
- Le Sud-Ouest
- Cote-des-Neiges-Notre-Dame-de-Grace

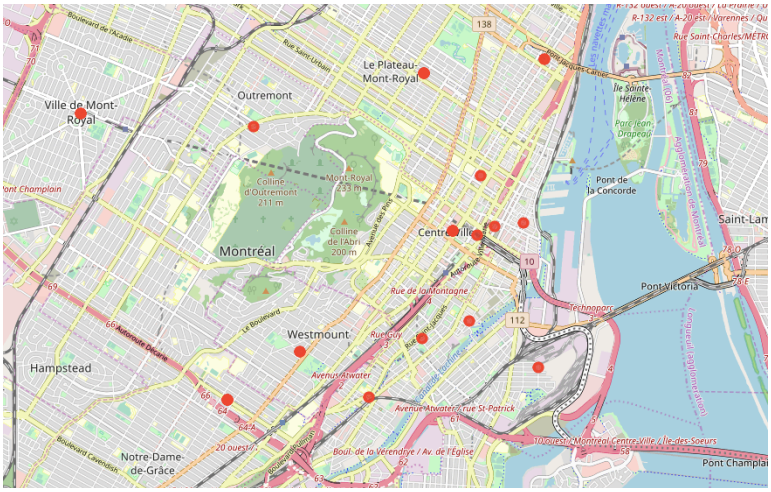

Figure 1 Montreal's selected venues map.

Vancouver's Neighborhoods:

- University Endowment Lands
- West Point Grey
- Dunbar-Southlands
- Kitsilando
- Arbutus Ridge
- Kerrisdale
- Shaughnessy
- Fairview
- South Cambie
- Oakridge
- Marpole

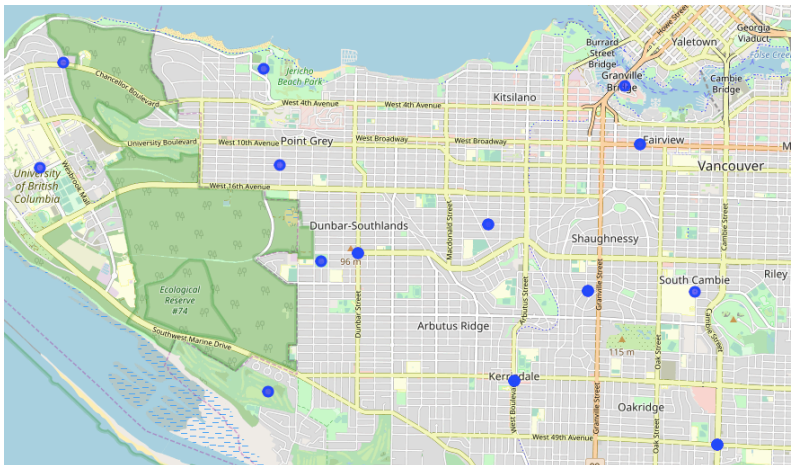

Figure 2 Vancouver's selected venues map.

This decision reduced the size of both data sets to 33 and 27 data points respectively. The developers edition of the Foursquare's API can only handle a limited number of calls per day, therefore, the importance of limiting the number of neighborhoods to analyze.

Methodology

The next step on the analysis consisted in using the Foursquare's API to obtain the 50 most visited venues within a radius of 1 km of each neighborhood coordinate. These produced two data frames with the venues data for both cities. Next, the data was grouped within the data frames based on the zip code, borough and neighborhood. This was done in order to check the number of venues found per zip code, this process was done for both cities venues data frames (figures 3 and 4). Finally, the next step consisted on detecting the number of unique type of venues stored within each city's data frame. The Montreal's venues data frame contained 162 unique types of venues and the Vancouver's venues data frame contained 129 unique types of venues.

PostalCode	Borough	Neighborhood	
H2H	Le Plateau-Mont-Royal	Plateau Mont-Royal North	50
H2J	Le Plateau-Mont-Royal	Plateau Mont-Royal North Central	50
H2K	Ville-Marie	Centre-Sud North	50
H2L	Ville-Marie	Centre-Sud South	50
H2T	Le Plateau-Mont-Royal	Plateau Mont-Royal West	50
H2V	Outremont	Outremont	47
H2W	Le Plateau-Mont-Royal	Plateau Mont-Royal South Central	50
H2X	Le Plateau-Mont-Royal	Plateau Mont-Royal Southeast	50
H2Y	Ville-Marie	Old Montreal	50
H2Z	Ville-Marie	Downtown Montreal Northeast	50
H3A	Ville-Marie	Downtown Montreal North	50
H3B	Ville-Marie	Downtown Montreal East	50
H3C	Le Sud-Ouest	Griffintown	50
H3G	Ville-Marie	Downtown Montreal Southeast	50
H3H	Ville-Marie	Downtown Montreal Southwest	50
H3J	Le Sud-Ouest	Petite-Bourgogne	50
H3K	Le Sud-Ouest	Pointe-Saint-Charles	26
H3P	Le Plateau-Mont-Royal	Mount Royal North	21
H3R	Le Plateau-Mont-Royal	Mount Royal Central	21
H3S	Cote-des-Neiges-Notre-Dame-de-Grace	Cote-des-Neiges North	50
H3T	Cote-des-Neiges-Notre-Dame-de-Grace	Cote-des-Neiges Northeast	50
H3V	Cote-des-Neiges-Notre-Dame-de-Grace	Cote-des-Neiges East	50
H3W	Cote-des-Neiges-Notre-Dame-de-Grace	Cote-des-Neiges Southwest	50
H3Y	Cote-des-Neiges-Notre-Dame-de-Grace	Westmount North	50
H3Z	Cote-des-Neiges-Notre-Dame-de-Grace	Westmount South	50
H4A	Cote-des-Neiges-Notre-Dame-de-Grace	Notre-Dame-de-Grace Northeast	50
H4B	Cote-des-Neiges-Notre-Dame-de-Grace	Notre-Dame-de-Grace Southwest	50
H4C	Le Sud-Ouest	Saint-Henri	50
H4E	Le Sud-Ouest	Ville Amard	29
H4P	Le Plateau-Mont-Royal	Mount Royal South	21
H4Z	Ville-Marie	Tour de la Bourse	50
H5A	Ville-Marie	Place Bonaventure	50
H5B	Ville-Marie	Place Desjardins	50

Figure 3 Montreal's venues count per zip code.

PostalCode	Borough	Neighborhood	
V5W	Oakridge	NE Oakridge	50
V5X	Marpole	East Marpole	50
	Oakridge	SE Oakridge	50
V5Z	Fairview	East Fairview	50
	South Cambie	South Cambie	47
V6H	Fairview	Granville Island	50
		West Fairview	50
	Shaughnessy	NE Shaughnessy	11
V6J	Shaughnessy	NW Shaughnessy	11
V6L	Arbutus Ridge	NW Arbutus Ridge	30
	Dunbar-Southlands	NE Dunbar-Southlands	17
V6M	Arbutus Ridge	SE Arbutus Ridge	30
	Kerrisdale	NE Kerrisdale	44
	Oakridge	NW Oakridge	50
	Shaughnessy	South Shaughnessy	11
V6N	Arbutus Ridge	Musqueam	3
	Dunbar-Southlands	South Dunbar-Southlands	17
	Kerrisdale	West Kerrisdale	44
V6P	Kerrisdale	SE Kerrisdale	44
	Marpole	West Marpole	50
	Oakridge	SW Oakridge	50
V6R	West Point Grey	Jericho	17
		West Point Grey	21
V6S	Dunbar-Southlands	Chaldecott	20
		NW Dunbar-Southlands	17
	University Endowment Lands	South University Endowment Lands	37
V6T	University Endowment Lands	UBC	50

Figure 4 Vancouver's venues count per zip code.

The next sept consisted on doing one hot encoding on both data frames and then find the 10 most visited venues in each area for each city. This was an essential task to apply K-Means clustering in the data, in order to obtain the neighborhood segmentation. This machine learning algorithm was selected, as we are dealing with an unsupervised task. Also, it has the benefit that is possible to manually adjust the number of desired clusters. The obtained results are presented in the following section.

Results

After applying K-means clustering, it was possible to identify 3 clusters per city. The profiles of each cluster are presented bellow:

Montreal's clusters:

- Cluster 0 (red): The venues part of this cluster are mainly Asian restaurants, bakeries, coffee shops, hotels and bars.
- Cluster 1 (purple): It is the largest clusters and the venues part of this cluster are a wide variety of restaurants, bakeries, coffee shops, tea shops, parks, banks, hotels and gyms.
- Cluster 2 (green): The venues part of this cluster are mainly coffee shops, train stations, pharmacies, bakeries, parks, gyms, banks and some restaurants. Probably the neighborhoods part of this cluster, are living areas.

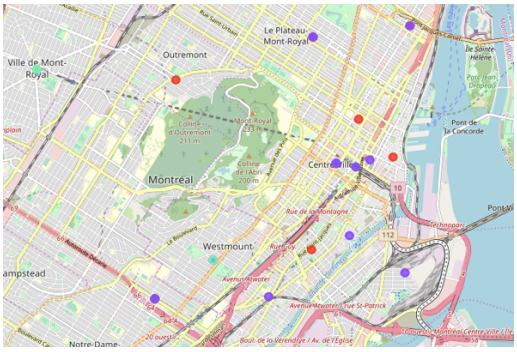

Figure 5 Montreal's selected neighborhoods venue segmentation.

Vancouver's clusters:

- Cluster 0 (red): It is the largest cluster and the venues part of this cluster are restaurants, coffee shops, parks, tea shops, markets, grocery store, malls and sport centers.
- Cluster 1 (purple): The venues part of this cluster are restaurants, coffee shops, parks, a golf course, gyms, tea shop, pharmacies and bars.
- Cluster 2 (green): The venues part of this cluster are coffee shops, gardens, parks, restaurants, malls, bakeries, bus stops, furniture stores, art/crafts stores and event spaces. The neighborhoods in this cluster, are probably mainly living areas.

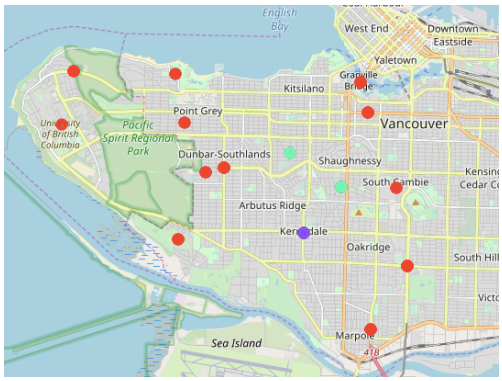

Figure 6 Vancouver's selected neighborhoods venue segmentation.

From the exploratory data analysis and the clustering steps, it was possible to observe that each city is very different. All of Vancouver's neighborhoods have parks, gardens or sport venues within them. In contrast, in Montreal, only the purple and green clusters have parks within them. Another interesting difference between these cities, is that only in the red Vancouver's cluster is possible to observe the presence of markets and grocery stores. In terms of restaurants, coffee shops and tea shops, both cities have a wide variety of this type of venues.

Discussion

After completing the proposed task, there are several considerations that can be taken to improve the obtained results. In order to build both cities data sets, I had to manually search for the necessary data. I couldn't find any web page that had information in a structured enough format, in order to carry out text scrapping. I tried to use several geocoder APIs to obtain the geographical coordinates of each zip code, but couldn't obtain the required data. I would suggest to explore another geocoder to try to optimize this step of my project. Also, I consider a huge disadvantage the limited number of calls to the Foursquare's API that can be made per day.

As we are dealing with an unsupervised task, it is not possible to determine the optimum number of clusters. If this project is deployed into production, it would be important to give the user the option of defining the number of clusters and the number of top venues to use. Perhaps, it would be interesting to display the list of top venues and the cluster map side by side.

Conclusion

After carrying out the data analysis process, it was possible to confirm that two cities can be compared based on the obtained profiles by K-Means clustering. An application based on this project can help a user to identify a flat/house within a neighborhood that matches the desired profile. Also, if it would be used to complement an application such as Airbnb, it would help clients to consider available rentals based on the surrounding venues and not only on price differentiation.