7/25/2020

Where to Live

Prepared by Mohamed ElFouly

Contents

1. Introduction

1.1 Background

People traveling on business trips or people willing to move to other countries usually prefer to stay at locations similar to the ones they currently live in. Others may prefer to move to places that are similar to other places they would like to live in. For example, people who live in western-cultured neighborhoods may be inclined towards living in places with similar taste. On the other hand, they may wish to relocate to areas of a different flavor, a middle-eastern one per say. So, helping people decide the neighborhood they would live in based on places they prefer would be of great help.

This indeed would not be very different from recommender systems used in movies and so. Development in one field will surely affect others in the same domain.

1.2 Problem Statement

Recommending neighborhoods to stay in for travelers similar to the ones they prefer.

2. Data Acquisition and Cleaning

2.1 Data Sources

We will leverage the data obtained from the Foursquare API to gain insights about locations, and hence, decide on their similarity. We will focus on two areas, Toronto and New York. Also, we will obtain Toronto's dataset from Wikipedia and New York's dataset from "https://cocl.us/new_york_dataset"

2.2 Data Cleaning

2.2.1 New York's Dataset

New York's dataset was obtained in the form of a json file where the key we are interested in is "features". The "features" key referred to a list comprising data about specific neighborhoods, most importantly, id, coordinates and borough.

First, we extracted the data present in the "features" key; then, we created a pandas dataframe with "Borough", "Neighborhood", "Longitude", and "Latitude" as column labels. Finally, we filled the dataframe with the data contained within the "features" key accordingly. Also, the data was checked for missing values to ensure consistency.

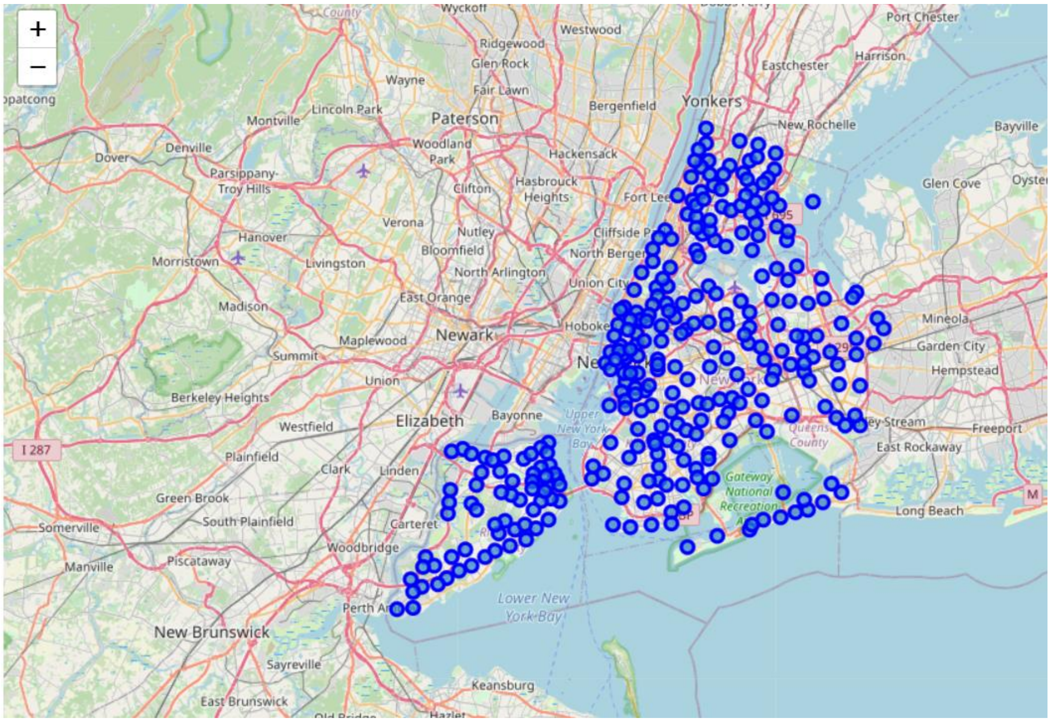
2.2.2 Toronto's Dataset

Toronto's dataset was parsed directly from the Wikipedia page into a pandas dataframe using pandas library. The dataframe comprised the following columns: Neighbourhood, Borough, and Postal Code. The Neighbourhood column was renamed to Neighborhood to be consistent with New York's data frame. We removed boroughs that were not assigned; then, we named the neighborhoods that were not assigned to their corresponding boroughs. Again, we made sure that there were no missing data. We read the co-ordinates, latitude and longitude, from a csv file into a dataframe; consequently, we merged the two aforementioned dataframes to obtain a complete dataframe with "Borough", "Neighborhood", "Longitude", and "Latitude" as column labels. After that, we dropped any rows that were not related to Toronto.

3. Exploratory Data Analysis

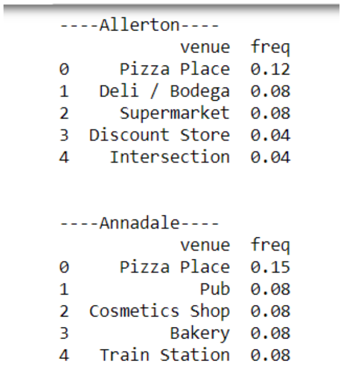
3.1 New York

By examining New York, we can find that it has 5 boroughs and 306 neighborhoods. Here's how they are superimposed on top of a map



Then, we can examine each neighborhood, characterized by its top venues with data obtained via the Foursquare API. For instance, if we look closer at the neighborhood "Wakefield", we find that within a radius of 500 meters, 9 venues were returned via the Foursquare API.

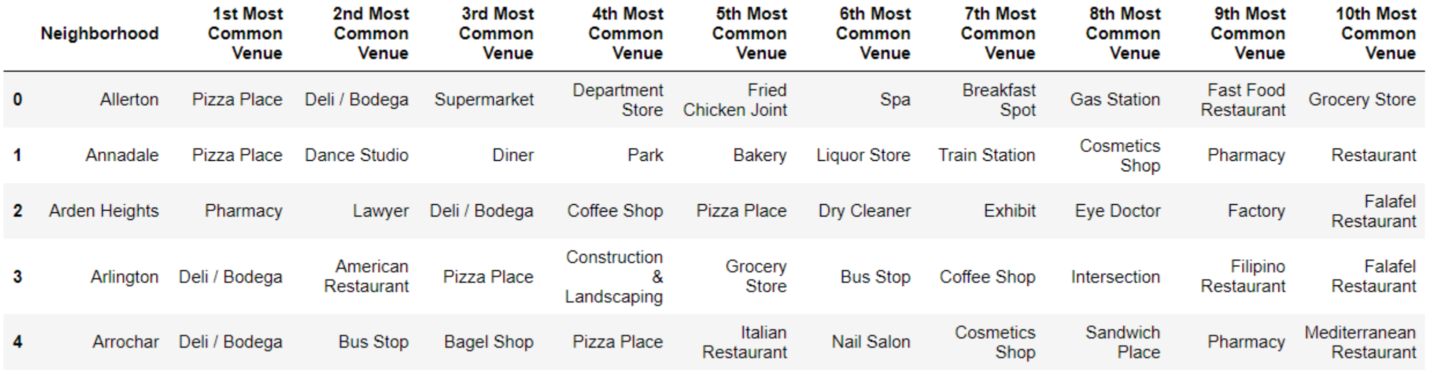
By repeating the same process for all neighborhoods, we find that there is a total of 10,115 venues in New York, with 425 unique categories. To be able to deal with the unique categories, we one-hot encoded them. We then calculated the contribution of each category of venues to each neighborhood. Here is an example of 2 neighborhood with their top 5 venues frequencies:



```
----Allerton----
            venue  freq
0     Pizza Place  0.12
1    Deli / Bodega  0.08
2     Supermarket  0.08
3  Discount Store  0.04
4    Intersection  0.04


----Annadale----
            venue  freq
0     Pizza Place  0.15
1             Pub  0.08
2  Cosmetics Shop  0.08
3          Bakery  0.08
4   Train Station  0.08
```

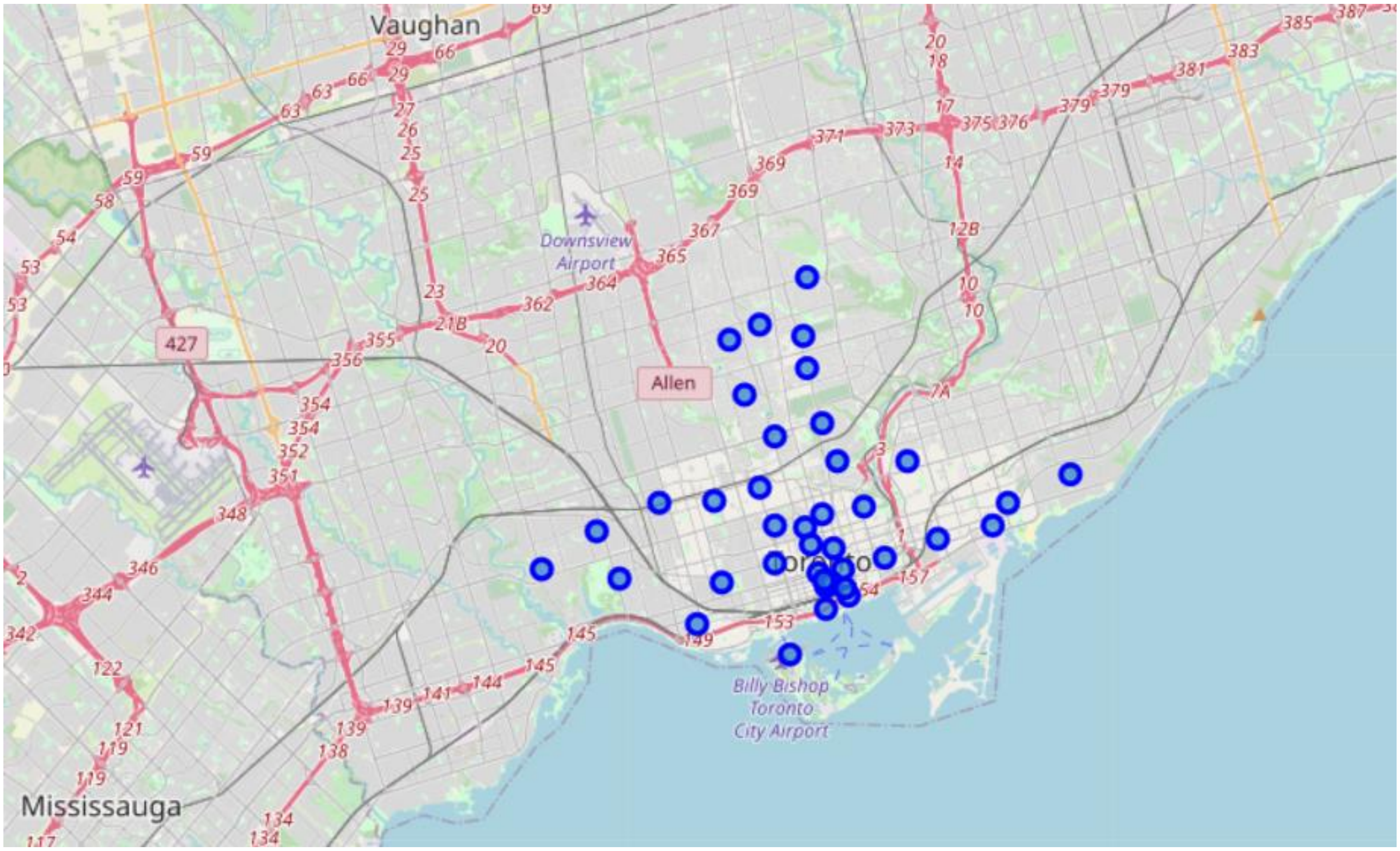
We then sorted the frequencies obtaining a data frame where each row represents a neighborhood, and columns represent the top 10 venues in that neighborhood ranked from first to tenth in order of frequency. Here is an example of how the dataframe looks like:



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Allerton	Pizza Place	Deli / Bodega	Supermarket	Department Store	Fried Chicken Joint	Spa	Breakfast Spot	Gas Station	Fast Food Restaurant	Grocery Store
1	Annadale	Pizza Place	Dance Studio	Diner	Park	Bakery	Liquor Store	Train Station	Cosmetics Shop	Pharmacy	Restaurant
2	Arden Heights	Pharmacy	Lawyer	Deli / Bodega	Coffee Shop	Pizza Place	Dry Cleaner	Exhibit	Eye Doctor	Factory	Falafel Restaurant
3	Arlington	Deli / Bodega	American Restaurant	Pizza Place	Construction & Landscaping	Grocery Store	Bus Stop	Coffee Shop	Intersection	Filipino Restaurant	Falafel Restaurant
4	Arrochar	Deli / Bodega	Bus Stop	Bagel Shop	Pizza Place	Italian Restaurant	Nail Salon	Cosmetics Shop	Sandwich Place	Pharmacy	Mediterranean Restaurant

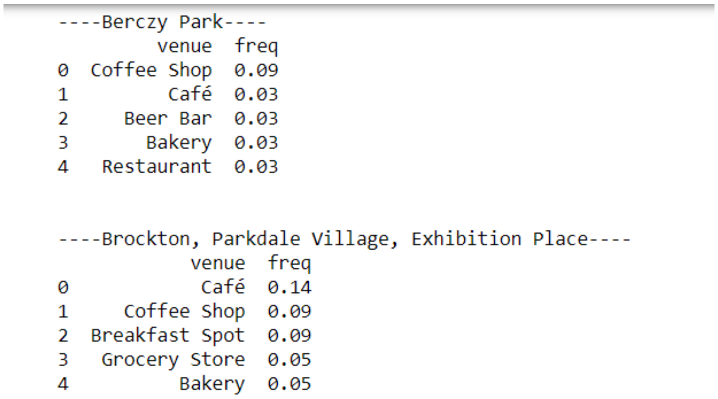
3.2 Toronto

A closer look at Toronto will reveal that it has 4 boroughs and 39 neighborhoods, which is nearly an eighth of the number of neighborhoods in New York. Here we can see them on the map:



And by examining each neighborhood according to the venues returned by the Foursquare API, we can find that, for instance, the number of venues in Regent Park, Harbourfront within a 500-meter radius is 45.

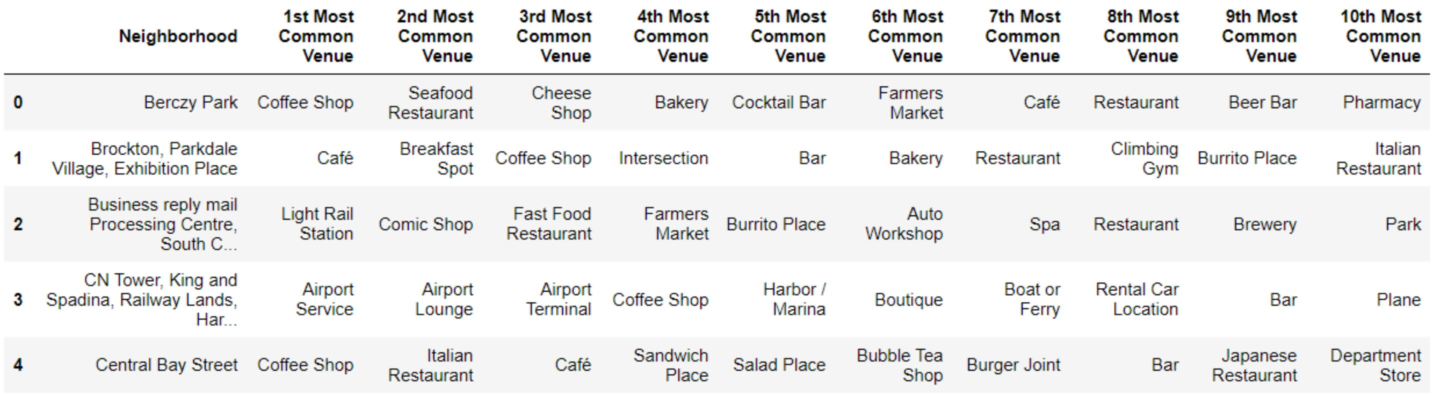
Again, by repeating the same process, we can find that there is a total of 1637 venues, partitioned in 231 unique categories, which we will have to manipulate by one-hot encoding. We can then find out each category's contribution to a certain neighborhood. Here is a sample of what would that look like:



```
----Berczy Park----
         venue  freq
0  Coffee Shop  0.09
1         Café  0.03
2     Beer Bar  0.03
3       Bakery  0.03
4   Restaurant  0.03


----Brockton, Parkdale Village, Exhibition Place----
            venue  freq
0            Café  0.14
1     Coffee Shop  0.09
2  Breakfast Spot  0.09
3   Grocery Store  0.05
4          Bakery  0.05
```

Similar to what we did with New York, the frequencies are sorted and displayed in a data frame, where each row represents a distinct neighborhood, and each column represents the top-venue categories ranked from one to 10, in accordance with their frequencies. And that is how it looks like in a table:



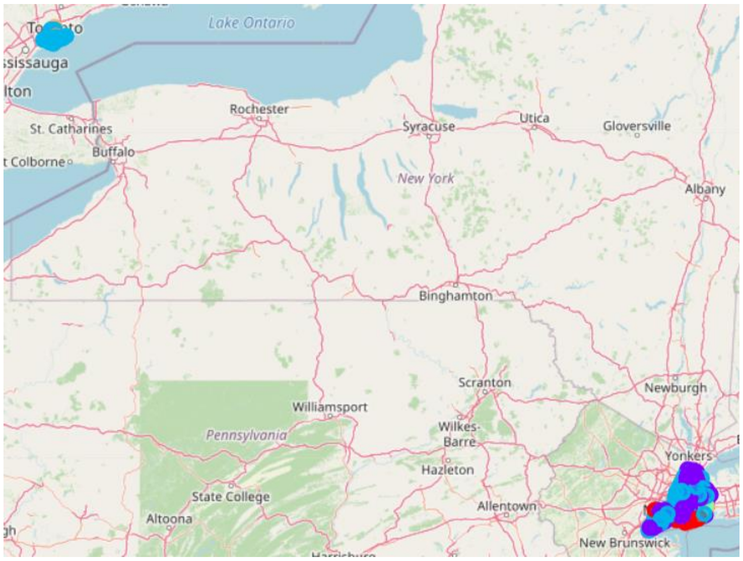
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Berczy Park	Coffee Shop	Seafood Restaurant	Cheese Shop	Bakery	Cocktail Bar	Farmers Market	Café	Restaurant	Beer Bar	Pharmacy
1	Brockton, Parkdale Village, Exhibition Place	Café	Breakfast Spot	Coffee Shop	Intersection	Bar	Bakery	Restaurant	Climbing Gym	Burrito Place	Italian Restaurant
2	Business reply mail Processing Centre, South C...	Light Rail Station	Comic Shop	Fast Food Restaurant	Farmers Market	Burrito Place	Auto Workshop	Spa	Restaurant	Brewery	Park
3	CN Tower, King and Spadina, Railway Lands, Har...	Airport Service	Airport Lounge	Airport Terminal	Coffee Shop	Harbor / Marina	Boutique	Boat or Ferry	Rental Car Location	Bar	Plane
4	Central Bay Street	Coffee Shop	Italian Restaurant	Café	Sandwich Place	Salad Place	Bubble Tea Shop	Burger Joint	Bar	Japanese Restaurant	Department Store

Now we will move on with clustering the neighborhoods of both Toronto and New York, along with the results of that clustering

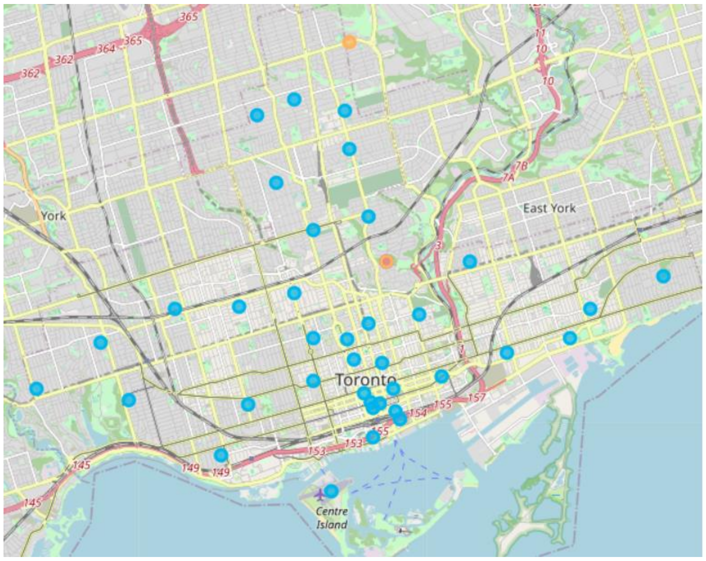
4. Clustering Results

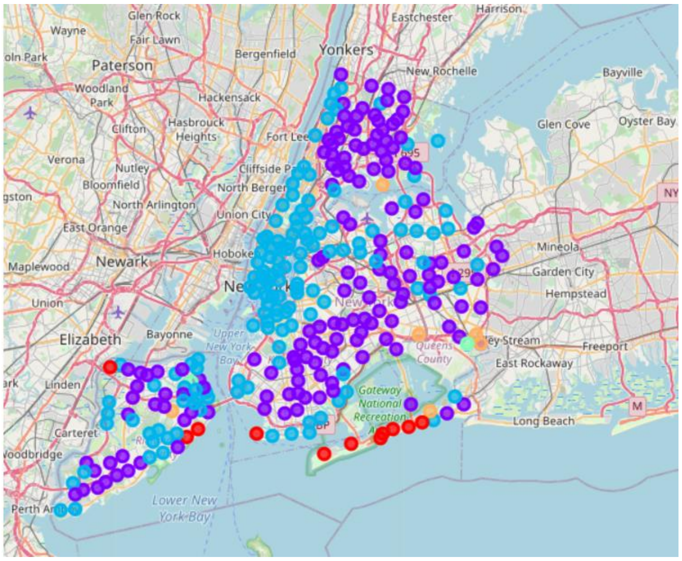
First, we will have to combine both dataframes, Toronto and New York; then, we can run our k-means clustering algorithm. Five clusters will be pre-determined, to which neighborhoods will relate.

Upon fitting the algorithms to our data, we obtained the following map of clusters:



Evidently, the map doesn't reveal much due to the distance between the two countries. After some image manipulation, and ignoring distance proportionality, we can get the following image:





5. Discussion, Conclusion and Future Work

It is clear that New York (the image on the right) is much more diverse than Toronto, as Toronto nearly falls in just one cluster. This technique has great potential, yet it may have not quite shined here. Future work may include trying different number of clusters, or trying to refine the dataset.

Moreover, other methods could be tried and implemented while comparing the results. For instance, we might cluster the neighborhoods in a city using k-means clustering algorithm, then we can classify neighborhoods of other cities using k-nearest neighbors classification algorithm; then, results can be compared and contrasted.