The Battle of Neighbourhood's

Identifying the similarity (or dissimilarity) between Neighbourhoods, and predict any business opportunities amongst themselves

Koushik Karmakar

November 28, 2019

1. Introduction

1.1. Background

- Nowadays, cities across the world (let's say financial capitals of the countries) present an array of job, business and tourism opportunities, to attract people from different parts of the globe (especially from financial capital of a competitive country).
- At the same time, people (individuals, entrepreneurs, etc) in this era of globalization, are willing to travel across the world and grab these opportunities.
- While opportunities, play a major role in decision making (to travel across cities), another important factor that one lingers to, is Neighbourhood/locality. People can definitely let go of opportunities, if the destination cities doesn't present them with the choice of their preferred Neighbourhood/locality.
- A comparative study between one's current neighbourhood, and the neighbourhoods of the destination city will be of great help in the decision-making process.

1.2. Problem

This project aims to compare the neighbourhoods of few major financial capitals of the world, on the basis of the venues (e.g. Restaurants, Parks, Museums, Hotels, Stores, etc) present in the Neighbourhood, and present two pieces of information:

i. How similar or dissimilar are the neighbourhoods of one city, compared to another

ii. What are the business opportunities that the neighbourhoods present, in terms of their similarity with another neighbourhoods (within same city or different city)

1.3. Interest

- Tourists, who want to travel these cities. They can select the Neighbourhood to live, depending on what the Neighbourhood has to present to them or as per their own taste of Neighbourhood.
- People, who are willing to relocate across different cities of the world in search of better job opportunities.
- Entrepreneurs, who are willing to expand their business (overseas or within the same city). Using this report, they can identify locations, which has appetite for their business.

2. Data Acquisition and cleaning

2.1. Data Sources

There are two pieces of information that are required here, namely:

1. Neighbourhood information: Name/details of Borough, Neighbourhood, along with its longitude and latitude
2. All different kind of venues (Restaurants, Parks, Museums, Hotels, Stores, etc) present in the neighbourhood.

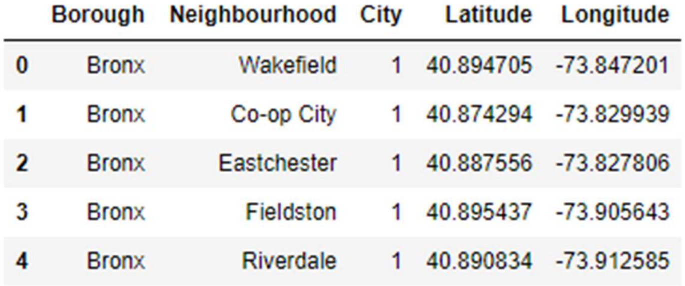
As a part of this report, we will compare New York, Toronto & Paris. Thus, we needed the above information for the three cities.

1. New York: The information with neighbourhoods of New York (along with latitude and longitude), exists as a json file on the web.
2. Toronto: The information with neighbourhood information of the city of Toronto is present on a Wikipedia page. This information didn’t have the latitude and longitude details. The latitude and longitude details are present in the web.
3. Paris: The neighbourhood information of Paris (known as Arrondissement & Quartiers) is present on a Wikipedia page. The latitude and longitude details were not present readily, thus were extracted one by one from google.

The venue information for each neighbourhood, was extracted using the **explore location** option of the **PLACES** API, provided by **FOURSQUARE**.

2.2. Data Processing/Cleaning and Feature Selection

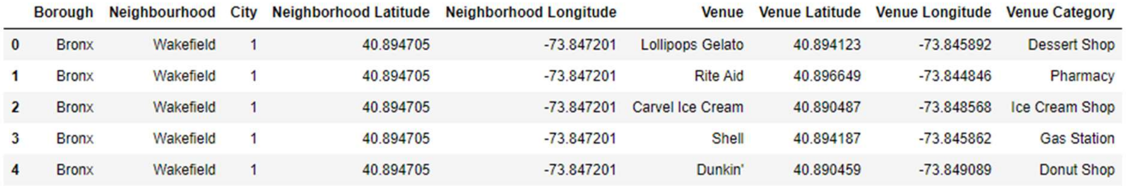
- The Neighbourhood information (Neighbourhood details Name, Postcode, etc & location information latitude, longitude, etc) were available in a different manner for different cities. The task here is to:
  a. Extract the information from different sources (for e.g. json file for New York, wiki page & excel sheet for Toronto, & wiki page & google for Paris)
  b. Standardize the column information (for e.g. Neighbourhoods in New York and Toronto are segregated as Neighbourhood & Borough, whereas Neighbourhoods in Paris are segregated as Arrondissement (Districts) & Quartiers).
  c. Label Neighbourhoods from different cities before they can be merged into one frame.
  d. Finally, the datasets are merged to create one dataset, and the features are Borough, Neighbourhood, City, Latitude & Longitude. The total number of Neighbourhoods are 489, and city numbers are New York 306, Toronto 103 and Paris 80.
  The Final dataset, after completing the above process looks something like:



	Borough	Neighbourhood	City	Latitude	Longitude
0	Bronx	Wakefield	1	40.894705	-73.847201
1	Bronx	Co-op City	1	40.874294	-73.829939
2	Bronx	Eastchester	1	40.887556	-73.827806
3	Bronx	Fieldston	1	40.895437	-73.905643
4	Bronx	Riverdale	1	40.890834	-73.912585

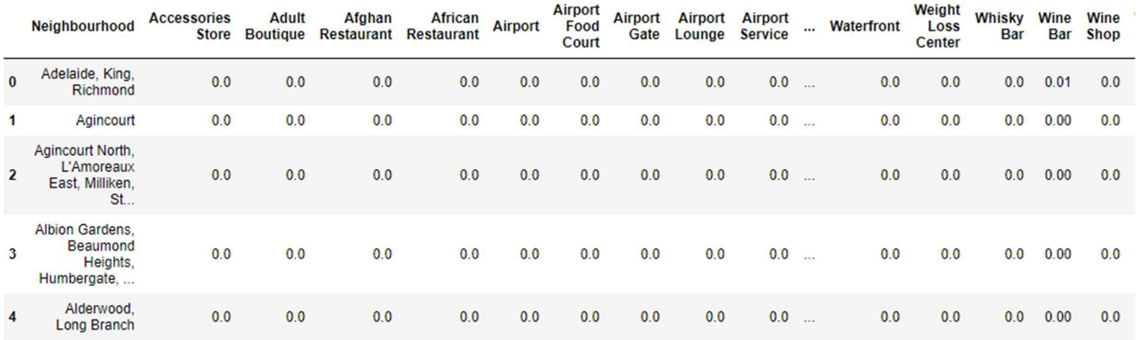
- **Explore location** option of the **PLACES API**, provided by **FOURSQUARE** was used to extract venues for within 1000 metre radius of each Neighbourhood/Location.
  - ✓ The API returned a total of 33,428 venues for the 489 neighbourhoods.
  - ✓ The total number of unique categories of venues returned, were 547.

    The returned information looked something like:



	Borough	Neighbourhood	City	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bronx	Wakefield	1	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Bronx	Wakefield	1	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
2	Bronx	Wakefield	1	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
3	Bronx	Wakefield	1	40.894705	-73.847201	Shell	40.894187	-73.845862	Gas Station
4	Bronx	Wakefield	1	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

- There were certain venue categories, which were general as well as widely classified. For e.g.
  - ✓ There are restaurant & food venues are very well classified. Restaurant as Caribbean, Indian, French, Japanese, Brazilian, etc (total 8742 venues). Food as Fast Food, Comfort Food, Food Trucks, etc (total 363). But there are 413 venues, which are just marked as “Restaurant”, and 59 which are just marked as “Food”. We will rename the “Restaurant” & “Food” categories as “General Eating Venue”.
  - ✓ There are other categories where general and classified categories co-exist (for e.g. Pub, Gastropub, Irish Pub or Market, Christmas Market, Flea Market, Fish Market, Super Market), but the total venues for these categories are very less. Thus, we will keep them as it is.
- The dataset that we now have (as seen in the above screenshot) is like, one row for each venue in a neighbourhood. Thus, we have 33,428 rows in our dataset. As our aim is to compare neighbourhoods, we need to compress this information to neighbourhood level. Thus, we:
  a. Use one hot encoding, on feature “Venue Category”. This creates 493 columns (number of unique category) for each row.
  b. Then we group the information, on basis of ‘Neighbourhood’ and find the mean for each “Venue Category”.
  c. The dataset now, looks something like this:



	Neighbourhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	...	Waterfront	Weight Loss Center	Whisky Bar	Wine Bar	Wine Shop
0	Adelaide, King, Richmond	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.01	0.0
1	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.00	0.0
2	Agincourt North, L'Amoreaux East, Milliken, St...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.00	0.0
3	Albion Gardens, Beaumond Heights, Humbergate, ...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.00	0.0
4	Alderwood, Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.00	0.0

- This is the dataset we will use for Comparing the Neighbourhoods. This dataset consists of 482 rows (neighbourhoods) & 548 features (‘Neighbourhood’ name, and 547 unique categories).
- <u>**Few Assumptions**</u>:
  - ✓ There are certain Neighbourhoods (17 in total) where the sum total of Venue of various Venue Categories, is less than 10. We are not going to drop these neighbourhoods from comparison process, as lesser number of Venue would represent absence of Venues in these neighbourhoods.