The Battle of Neighborhoods

New York City vs. Toronto

Vivian Lo

July 4, 2020

1. Introduction

A well-established Chinese restaurant group is expanding their branch restaurants. Two of the world's largest and exciting cities, New York City (US) and Toronto (Canada), are the two finalists on the agenda of the next board meeting. We are assigned a compare-and-contrast project of the two cities. We will utilize the Foursquare database and API to explore the neighborhoods in the two cities. We will use the explore function to get the most common venue categories in each neighborhood, and then use this feature to group the neighborhoods into clusters. We will use the k-means clustering algorithm to complete this task. Finally, we will use the Folium library to visualize the neighborhoods in the two cities and their emerging clusters. By presenting the similarities and dissimilarities of the two cities we hope to help the restaurant group to make a profitable investment decision.

2. Data Acquisition and Cleaning

2.1 New York City Data

New York City neighborhood has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood.

Here is the link to the dataset: https://geo.nyu.edu/catalog/nyu_2451_34572

We can also use geopy library to get the latitude and longitude values of an actual address.

2.2 Toronto Data

For the Toronto neighborhood data, a Wikipedia page exists that has all the information we need to explore and cluster the neighborhoods in Toronto:
https://en.wikipedia.org/wiki/list_of_postal_codes_of_Canada:_M

The dataframe consists of three columns: Postal Code, Borough, and Neighborhood. We will prepare the data as follows:

- Only process the cells that have an assigned borough. Ignore cells with a borough that is Not assigned.

- More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, you will notice that M5A is listed twice and has two neighborhoods: Harbourfront and Regent Park. These two rows will be combined into one row with the neighborhoods separated with a comma as shown in row 5 in the above table.

- If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough.

Now that we have built a dataframe of the postal code of each neighborhood along with the borough name and neighborhood name, in order to utilize the Foursquare location data, we need to get the latitude and the longitude coordinates of each neighborhood.

Here is a link to a csv file that has the geographical coordinates of each postal code: http://cocl.us/Geospatial_data

3. Methodology

Foursquare is a technology company that has built a comprehensive and accurate location database. We can make calls to the Foursquare API for different purposes. We can construct a URL to send a request to the API to search for a specific type of venues, to explore a particular venue, to explore a Foursquare user, to explore a geographical location, and to get trending venues around a location. Also, we can use the visualization library, Folium, to visualize the results.

Next, we utilized the Foursquare API to explore the neighborhoods and segment them. The restaurant group would like us to focus on the neighborhoods in Manhattan borough in New York City data, and Toronto boroughs with “Toronto” in their names (we will name it “Toronto” borough in this report) which includes Downtown Toronto, East Toronto, West Toronto and Central Toronto boroughs in the Toronto data. So, we sliced the original dataframes and created new dataframes of the interested subsets of data, Manhattan and Toronto.

3.1 Exploratory Data Analysis

For each neighborhood we requested the first 100 venues within the radius of 500 meters of the neighborhood’s latitude and longitude. We retrieved each venue’s name, category, latitude and longitude.

For the first neighborhood in the Manhattan dataframe, Marble Hill, we got 24 venues. Venues included Pizza Place, Yoga Studio, Diner, Coffee Shop, Donut Shop, etc. And no Chinese venues were found nearby.

For the entire Manhattan dataframe we got 3160 venues with 7 variables: Neighborhood name, Neighborhood latitude, Neighborhood longitude, Venue name, Venue latitude, Venue longitude and Venue category.

There were 327 unique categories. For each unique category we applied one hot encoding, resulting in a dataframe of 3145x328. Then we grouped rows by neighborhood and by taking the mean of the frequency of occurrence of each category, resulting in a dataframe of 40x328 with one row representing each neighborhood. After sorting the Venue category in descending order for each neighborhood we created the new dataframe to display the top 10 venues for each neighborhood.

We did the same in Toronto dataframe for the neighborhoods in Toronto borough.

The first neighborhood in Toronto is Regent Park, Harbourfront. The top 100 venues within 500 meters of the latitude and longitude requested to Foursquare API returned 44 venues. And the entire Toronto dataframe returned a total of 1614 venues. There were 233 unique venues. After one hot encoding and grouping by neighborhood we got a dataframe of 39x234.

It’s ready for clustering analysis.

3.2 Clustering

We ran k-means to cluster the neighborhoods based on the most common venues in each neighborhood. By trial and error, we clustered Manhattan data into 5 clusters, and Toronto data into 3 clusters.

After examining each cluster and its most common venues we determined the discriminating venue categories that distinguished each cluster.

4. Result

New York City had 5 boroughs and 306 neighborhoods. Toronto had 10 boroughs and 103 neighborhoods.

We focused in details on Manhattan borough in New York City dataframe and Toronto borough in Toronto dataframe. Manhattan borough had 3160 venues with 327 unique venues in its 40 neighborhoods. And Toronto borough had 1614 venues with 233 unique venues in its 39 neighborhoods. From the clustering of Manhattan and Toronto data we distinguished the neighborhoods features as below:

Manhattan borough neighborhoods has the following 5 clusters:

Italian Restaurant/ Coffee Shop

Hotel and Plaza

Sandwich Place

Park

Other Ethnic Restaurant

Toronto borough neighborhoods are similar in venues. Most of the neighborhoods belong to one cluster with Coffee Shops and other eateries. Other clusters include Park and Pool.

5. Discussion

Manhattan borough had 40 neighborhoods with 3160 venues in 327 categories, while Toronto borough had 39 neighborhoods with 1614 venues in 234 categories. Manhattan had more venues and, in more varieties, than Toronto. By clustering analysis, we also revealed that Manhattan had a lot of Italian restaurants, coffee shops and other eateries. Manhattan also had other types of venues like hotels, plazas and parks. In contrast Toronto borough was predominantly coffee shops and other eateries. Based on this information we would recommend to open a new Chinese restaurant in Manhattan for the volumes of potential customers and for merging into the similar neighborhoods.

6. Conclusion

In this project we utilized the Foursquare database and API to explore New York City, with the main focus on Manhattan borough, and Toronto, with the main focus on the four boroughs including Downtown Toronto, East Toronto, West Toronto and Central Toronto. Manhattan borough had 40 neighborhoods with 3160 venues in 327 categories, while Toronto borough had 39 neighborhoods with 1614 venues in 234 categories. With lots of restaurants, hotels, and plazas we concluded that Manhattan had more commercial activities, which could bring more potential customers if we open our new Chinese restaurant there. In future studies we would further subset venues to Chinese venues, and segment the two cities based on their Chinese venues respectively. We would also study the trending and tips of venues in the focused neighborhoods. These results would provide the valuable information to the restaurant group in their decision process.