The battle of the neighborhoods – Relocation between Toronto and New York City

by Ricardo F Reategui

Introduction

Thousands of skilled workers, businessmen, science and technology professionals, entrepreneurs, move from one city to another anywhere in the world every year seeking better opportunities for themselves and their families.

Imagine a relocation company which offers services to these clients in a unique way. This particular company could leverage the Foursquare location data with the needs and expectations of its clients in terms of housing, recreation, diversity, transportation, and other characteristics in order to tailor the needs of their clients who want to relocate to a neighborhood as similar as possible to their original place.

For this project I will use as a relocation model a company with a client who wants to move from Toronto to New York City or vice versa.

Data

- Datasets of Toronto and New York
- Foursquare API
- Folium library
- k-means algorithm

The dataset for New York City is available on the web (https://geo.nyu.edu/catalog/nyu_2451_34572). It can be retrieved as a JSON file and transformed into a dataframe. A portion of this dataframe is displayed Table 1. Information about the number of boroughs and neighborhoods can be obtained from this dataframe.

Table 1. First 5 rows of a pandas dataframe with the coordinates of New York City neighborhoods

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

The dataset for Toronto was obtained by reading a Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) directly into a pandas dataframe. The Wikipedia page contained the postal codes and the boroughs of the Toronto neighborhoods. A Geospatial data of the Toronto postal codes and their coordinates were already available as a CSV file (http://cocl.us/Geospatial_data). This file was read into a pandas dataframe.

Joining the Wikipedia page and the Geospatial data of Toronto generates the Toronto dataframe that will be used for this project. Table 2 shows the first 5 rows of this dataframe.

Information about the number of boroughs and neighborhoods can be obtained from this dataframe.

Table 2. First 5 rows of a pandas dataframe with the coordinates of Toronto neighborhoods

	Postal code	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park / Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor / Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park / Ontario Provincial Government	43.662301	-79.389494

The Foursquare API will be used with Toronto and New York dataframes to explore their respective neighborhoods and segment them. In order to use Foursquare, credentials such as Client ID, Client Secret, and Version of Foursquare will be defined.

A new dataframe will be generated for each city, Toronto and New York. Each dataframe will contain the venues that Foursquare encounters for each neighborhood. Venues will be grouped by neighborhood.

Clustering of the venues will be subsequently performed using the *k*-means clustering algorithm.

The Folium library will be used to visualize the clusters in the maps of New York City and Toronto.

Each cluster of New York City and Toronto will be analyzed in order to find similarities in venues. The objective will be to select the clusters that are similar the most.