Finding Similar Neighborhoods between Toronto, ON and Queens, New York

Collins Opoku-Baah

1. Introduction

1.1 Background

The world has become one giant global village and thus, each and every minute, people travel from one place to another. The purpose for traveling could be temporary that is for vacation, business, visits etc. or could be permanent e.g. school, work etc. When people live in a particular region for a long time, they tend to embrace the cultures (e.g. food, clothing, language etc.) of that region, making it very difficult to transition into other neighborhoods. For example, a person who loves seafood and attend yoga classes will want to move to a new place with such venues in order to continue having their pleasant life experience.

1.2 Problem

Finding a place that share similar venue as your current neighborhood can be burdensome considering how developed most cities in the world are currently. Hence, this project aims to find neighborhoods between two big cities namely Toronto, ON and Queens, New York that are similar with respect to venues. To do this, I will employ machine learning approaches and other techniques to segment and cluster neighborhoods in these two cities.

2. Data Acquisition and Cleaning

2.1 Data Sources

The data for this project will comprise the venue locations that are within a defined radius of the neighborhoods in the two big cities, which are Toronto, ON, Canada and Queens, New York, USA. First, we will obtain the various neighborhoods for the Toronto and Queens, New York from the web sources. The neighborhoods for Toronto, ON will be obtained by scraping the website, 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M' and that for Queens, New York will be obtained by downloading json file from https://cocl.us/new_york_dataset.

2.2 Data Cleaning

Both datasets contain more information than needed for this project. While the Toronto data contains information about all the Boroughs and Neighborhoods in Canada, the Queens data contains all the Boroughs and Neighborhoods in New York. With regards to the data for Toronto, we will create a dataframe containing only the neighborhoods under Boroughs named Toronto. Likewise, we will create another dataframe containing only the neighborhoods in Queens Borough. Both dataframes will then be combined into a single data frame containing the neighborhoods in both cities.

2.3 Feature Selection

The features for this project will be constructed from the distinct categories of venue locations in the neighborhoods of the two cities. However, in order to get these venues, we will have to obtain the latitude

and longitude coordinates for each of the neighborhoods. While the dataset for Queens already come with these coordinates, that of Toronto doesn't. We will employ the geocoder library to obtain the coordinates for each of the neighborhoods in our combined datasets.

After successfully obtaining these coordinates, we will utilize the Foursquare API to obtain location venues such as yoga places, restaurants, etc. that are within a defined radius around these neighborhoods. One-hot encoding will be employed to create features based on the distinct venue categories. The resulting dataframe will be grouped by neighborhoods to obtain a dataframe with neighborhoods in the rows and distinct venue categories in the columns. During the grouping, the number of venue under each category for a particular neighborhood will be averaged by the total number of venues for that neighborhood. This will give a fair representation of the proportion of each venue category for that neighborhood. See figure 1 for an example of the resulting dataframe.

Using K Means clustering approach, we will segment and cluster these neighborhoods to determine which ones between the two cities are similar.

	Neighborhood	Borough	Accessories Store	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	...	Vietnamese Restaurant	Warehouse Store	Weight Loss Center	Whisky Bar	Wine Bar	Wine Shop
0	Adelaide	Downtown Toronto	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.00000	0.0	0.010000	0.000000
1	Arverne	Queens	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.00000	0.0	0.000000	0.062500
2	Astoria	Queens	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.00000	0.0	0.000000	0.010000
3	Astoria Heights	Queens	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.00000	0.0	0.000000	0.000000
4	Auburndale	Queens	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.00000	0.0	0.000000	0.000000
5	Bathurst Quay	Downtown Toronto	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.038462	...	0.000000	0.0	0.00000	0.0	0.000000	0.000000
6	Bay Terrace	Queens	0.02381	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.02381	0.0	0.000000	0.000000
7	Bayside	Queens	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.00000	0.0	0.014925	0.000000
8	Bayswater	Queens	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.00000	0.0	0.000000	0.000000
9	Beechhurst	Queens	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.00000	0.0	0.000000	0.000000
10	Bellaire	Queens	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.00000	0.0	0.000000	0.000000
11	Belle Harbor	Queens	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.00000	0.0	0.000000	0.000000
12	Bellerose	Queens	0.00000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.00000	0.0	0.000000	0.043478

Figure 1. An example of the data for the project.