IBM Capstone – Week 4

Wenyu Xin

11/16/2020

Problem

For those who live in the United States, we may have encountered many news headlines that say, “Moving to Canada search spiked in Google Trends”. In fact, if someone look into the Google Trends for search term “move to Canada” and its searching index for the past 12 months, one will find this search scored 100 out of 100 and spiked on the Google Trends search index during November 1st and 11th [1].

With the current political and social unrests in the United States, it is understandable for Americans living in the United States to think about moving to Canada. On the day of writing this report, American reported over 130 thousand new Covid-19 cases, total of over 11 million cases, over 247 thousand death, and infection rates are on the rise [2]. Moreover, the question of who the next president is still in the air after almost 2 weeks after the election day.

Interest

The closest country that are friendly to and welcome Americans is Canada. It is the nation where the daily new Covid-19 cases number is only in the 4 thousand and total reported cases is just a little bit over 300 thousand [2]. With many similarities and less political and social unrests than the United States, Canada is the best choice for someone to leave the United States.

Data

Data Sources

Neighborhood data for Toronto is from Wikipedia, https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.

Geospatial data for both Toronto, https://cocl.us/Geospatial_data, and Manhattan and neighborhood data for Manhattan, https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json, are provided by the Coursera.

Venue data are pulled from Foursquare API, https://developer.foursquare.com/.

Data Cleaning

Neighborhood data for Toronto from Wikipedia was scraped from the website and transformed into a table. Geospatial data is read into a table as well. The first table did not have geospatial data like the latitude and longitude and the second table did not have the borough and neighborhood information. Thus, two tables were merged by the postal index. Furthermore, this merged dataset contained other boroughs other than Downtown Toronto. Thus, further filtering was done to reach the desired dataset.

Neighborhood information for Manhattan was in JSON format. The dataset was read directly into a table. The only cleaning needed was parsing the data and filtering the boroughs to Manhattan only.

The venue data was pulled directly using Foursquare API. It was also in a JSON format that just required parsing.

Data Usage

The neighborhood data will be useful in querying Foursquare API. Once the API data on venues are gathered, these data will be used to do a recommendation system that suggested which neighborhood in Toronto is best suitable for those coming from Manhattan based on the venues in Manhattan neighborhood.