Course: Coursera Data Science Professional by IBM
Title: Data Science Capstone Project
Author: Lee Aik Tuan

Introduction

In this project, we are working for a relocating firm that helps people relocate from Toronto, Ontario, Canada to Manhattan, New York, USA. As part of the effort to help customers ease into their new environment, we are starting a pilot project to help them find neighborhoods that match well with their existing neighborhood in Toronto. For example, people who stay near parks now would be interested in finding places with parks in Manhattan, or those who like sushi may prefer to have a few Japanese restaurants near their homes. In this case, the business question is "Which neighbourhoods in Manhattan most closely matches with a given neighbourhood in Toronto, in terms of venues and facilities?"

Data

To answer this question, we would need to develop a kind of recommendation system that recommends the neighborhoods of Manhattan based on the company's clients' existing neighborhoods in Toronto. We would need to neighbourhood information about both Toronto and New York. Information about the New York neighborhoods is available on the internet (https://geo.nyu.edu/catalog/nyu_2451_34572) but for the project, a cached copy at the IBM server (https://cocl.us/new_york_dataset) was used. Information about Toronto neighborhoods can be scraped off following the Wikipedia page, https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M and the coordinates can be obtained either from Google Maps Geocoding API or the Geocoder Python package. In this project, the latitudes and longitudes are sourced from a cached copy on the IBM server (https://cocl.us/Geospatial_data).

We would also need information about the venues or places of interest, their categories and the number of such venues in these two cities. For such information, we would make use of the Four Square APIs.