IBM Applied Data Science Capstone Project -- Final Report

Daisy Zhu

April, 2020

Clustering and segmenting neighborhood in Toronto

Table of content:

I. Introduction

This report is a part of IBM applied data science capstone project, a 9 course series for data science learning created by IBM on Coursera platform. The overall background and data source are given from the course whereas the rest of analysis and approach are left for course students to explore and develop. Through the project, students need to access data from website using scraping and to visualize data on Foursquare API. Furthermore, students can further explore data by using various Python packages and machine learning techniques to draw conclusions. Before conducting analysis and modeling, data will be collected, wrangled and cleaned. Students will find best data format and features for machine learning and modeling phase. And then, students need to find out the best fit model through trying different algorithms and tuning. Along with the whole process, students would visualize data to help understand and improve the modeling process.

The original idea for this project is that someone would like to open a new Chinese restaurant in the city Toronto. He/She is wondering where the location should be. The objective of the project is to explore neighborhood in Toronto and give a sound suggestion.

The audience for this report are:

- ✓ Potential investors who plan to run business
- ✓ Potential real estate buyers
- ✓ Potential real estate renters
- ✓ Course peers and instructors

II. Data description

The dataset for this project mainly comes from two parts:

- ✓ The Foursquare API: geographical data with related information will be access via Python scraping to get most venues for each neighborhood in the city of Toronto. By doing so, we can visualize geographical data on the map and clearly see how venues are distributed in neighbors.

- ✓ Toronto Neighborhood profiles: a csv file from data open source on Toronto city website. The data file contains many categories such as neighborhood info, population, families, household and marital status, language, income etc. We can explore these data to dig out features which affect distribution of venues and type of business in each area.

- ✓ Other relational data: a csv file containing geospatial coordinate data. We can join coordinate data with dataset above.

You can search and look at data at
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
https://open.toronto.ca/dataset/neighbourhood-profiles/

Daisy Z.