Opening an Italian Restaurant in Toronto

Sixuan Zhu

September 19th, 2020

# Introduction

## Background

Toronto, the capital of the province of Ontario, is a major Canadian city along Lake Ontario’s northwestern shore. It's a dynamic metropolis with a core of soaring skyscrapers, all dwarfed by the iconic, free-standing CN Tower, and a various selection of global cuisine restaurants. Sam Xiong is looking to open a new Italian restaurant in Toronto City.

## Problem

With all these selections, the competition for opening a new restaurant is tough. Sam would like to know which location would have the highest probability of success. In order to provide a data-driven and fact-based solution, we will need to take into accounts attributes such as the venue popularity of the neighborhood, the characteristics of each neighborhood, the similarity and dissimilarity of each neighborhood, and the top common venues within each neighborhood.

# Data acquisition and cleaning

## Data Source

Two main data sources for this project would be Toronto neighborhood data and Foursquare location data. For the Toronto neighborhood data, a [Wikipedia page](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) exists that has all the information we need to explore and cluster the neighborhoods in Toronto. Foursquare is the location data provider that share information about businesses and attractions around a certain a location point. We will need to retrieve Foursquare location data around Toronto City through API.

## Data cleaning

The first problem with the dataset is to get Toronto neighborhood data from the Wikipedia page. I scraped the Wikipedia page and wrangle the data, cleaned it, and then read it into a pandas dataframe so that it is in a structured format. Then I combined the neighborhood under the same postal code. Here the new data frame contains 103 neighborhoods.

The second step is to get coordinates data and join them to the neighborhood dataset. I have obtained a dataset containing the coordinates data. The step is an intermediate step to connect the neighborhood dataset and Foursquare location data.

The third step is to get specific venue information for all neighborhood in Toronto city from Foursquare through the API program. Then I have written a function to create the API request and then return only the relevant information for each nearby venue, i.e. venue name, venue location, venue category, venue lat, venue lng, and venue name. I also limited the number of venues returned to 100 to avoid the list get too exhaustive.

With all the data digested into pandas data frames, I can start to perform exploratory analysis and use clustering technique to analyze the characteristics of a neighborhood.