# Use Foursquare APIs and ML to find a good place to open new restaurant in Toronto

## 1. Introduction

### 1.1 Background

Let's say we are going to open a new restaurant in Toronto; and before doing that, we need to explore the current market in that area. Assumed that we will open a Chinese restaurant, so we need to find a location where Asian food is popular.

### 1.2 Business Problem

We need to find the most suitable location within a suitable neighborhood to open a new Chinese restaurant in Toronto. By using data science methods and machine learning methods such as k-Means or KNN, this project aims to provide solutions to answer the business question: In Toronto, if an entrepreneur wants to open a Chinese restaurant, where should they consider opening it?

### 1.3 Target Audience

The entrepreneur who wants to find the location to open authentic Chinese restaurant

## 2. Data

To solve this problem, I will need below data:

- List of neighborhoods in Toronto, Canada.
- Latitude and Longitude of these neighborhoods.
- Venue data related to Asian restaurants. This will help us find the neighborhoods that are the most suitable to open a Chinese restaurant.

## 3. Scrapping the data

- Scrapping data of Toronto neighborhoods on Wikipedia
- Extract lat-long of them
- Call Foursquare APIs to get venue data related to these neighborhoods

## 4. Methodology

First, I need to get the list of neighborhoods in Toronto, Canada. This is possible by

extracting the list of neighborhoods from wikipedia page

(" https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M ") I did the web

scraping by utilizing pandas html table scraping method as it is easier and more

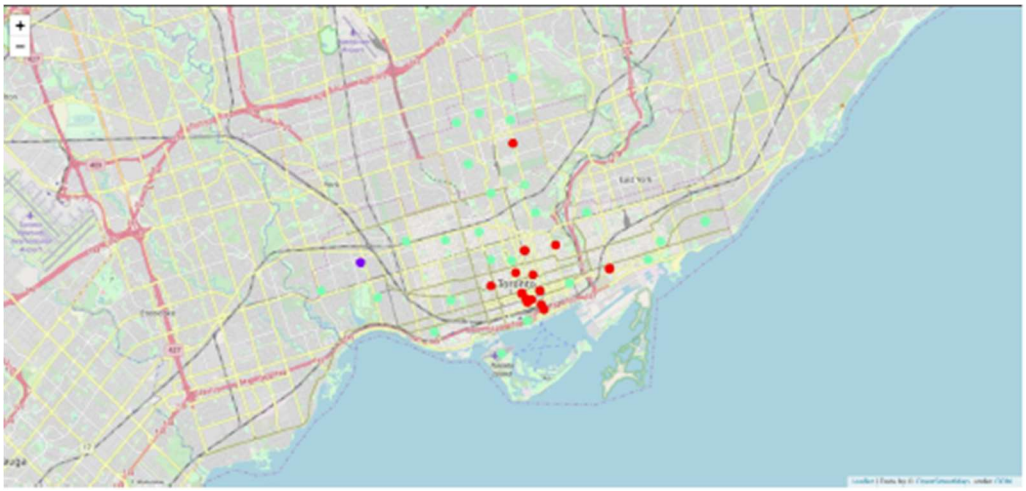
convenient to pull tabular data directly from a web page into dataframe.

Next, I use Foursquare API to pull the list of top 100 venues within 500 meters radius. I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I am able to pull the names, categories, latitude and longitude of the venues. With this data, I can also check how many unique categories that I can get from these venues. Then, I analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later.

Here, I made a justification to specifically look for “Thai restaurants”. Previously, when I ran the model, I was looking for “Asian restaurants” but there are very few results (maybe due to Foursquare categorization) so I looked for the restaurants closest to Chinese cuisine taste (side note: Chinese food and Thai food are very similar in taste, so my justification is that if there are people who enjoyed Thai food, they likely are going to enjoy Chinese food too!)

Lastly, I performed the clustering method by using k-means clustering. K-means clustering algorithm identifies k number of centeriods, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well. I have clustered the neighborhoods in Toronto into 3 clusters based on their frequency of occurrence for “Thai food”. Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.

## 5. Results



The results from k-means clustering show that we can categorize Toronto neighborhoods into 3 clusters based on how many Thai restaurants are in each neighborhood:

- Cluster 0: Neighborhoods with little or no Thai restaurants
- Cluster 1: Neighborhoods with no Thai restaurants
- Cluster 2: Neighborhoods with high number of Thai restaurants

The results are visualized in the above map with Cluster 0 in red color, Cluster 1 in purple color and Cluster 2 in light green color.

## 6. Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing the machine learning by utilizing k-means clustering and providing recommendation to the stakeholder.