The Battle of the Neighborhoods

Finding the best location to open an Indonesian Restaurant in Toronto, Canada

September 19

IBM Coursera Capstone Project

Fitria Kurniasari

Contents

1. Introduction

In the early post-World War II period, most migrants from Indonesia to Canada were Indo people of mixed Dutch and original ancestry. Many did not come directly from Indonesia, but rather went to the Netherlands and then re-migrated due to racial prejudice they faced there. Community members believe that perhaps 3,000 live in the Ontario area. Indonesians of Chinese descent formed the main group in the stream of migration which began in the late 1960s and early 1970s. They have come to comprise an estimated 80% of Canada's population of Indonesian background. 7,610 respondents to the 1991 census stated their place of birth as "Indonesia". Around half of those were settled in the Greater Toronto Area. Data from the 2006 Census suggested that 14,320 people of Indonesian ethnic origin reside in Canada (3,225 single responses, 11,095 in combination with other responses), primarily in Ontario (6,325, or 44%), British Columbia (4,640, or 32%), and Alberta (1,920, or 13%). (Wikipedia)

Business Problem

Due to the numbers of Indonesian immigrant in Greater Toronto, Canada, our client want to open Indonesian restaurant in which area. The objectives of this capstone project is to select and analyze the best location in Toronto to open a new restaurant. Using data science methodology and machine learning techniques like clustering in this project to provide solutions to answer the business question: where would you recommend to open a new Indonesian restaurant in Toronto, Canada?

2. Data

The following data is needed to solve the problem

a. List of the neighborhoods in Toronto, Canada
b. Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venues data.
c. Venues data, particularly related to Indonesian Restaurants. For this project we will use the Foursquare Places API. One of the features of this API is to provide a list of venues within a specific location, based on the Latitude and longitude coordinates and a radius. The data is used to clustering on the neighborhoods.

Sources of data and methods to extract them

The Wikipedia page contains a list of neighbourhoods in Toronto, Canada (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M), with a total of 4 borough and 38 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Indonesian Restaurant category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

3. Methodology

How to get neighborhood list

Fortunately, the list of neighbourhoods in the Toronto, Canada. is available in the Wikipedia page (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). We do web scraping using Python requests and beautifulsoup packages to extract the list of neighborhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the Toronto city.

Used Foursquare API to get the top 100 venues that are within a radius of 2000 meters

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the “Indonesian Restaurant” data, we will filter the “Indonesian Restaurant” as venue category for the neighbourhoods.

Clustering on the data by using k-means clustering

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, 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 is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 8 clusters based on their frequency of occurrence for “Indonesian Restaurant”. The results will allow us to identify which neighbourhoods have higher concentration of Indonesian Restaurant while which neighbourhoods have fewer number of Indonesian Restaurant. Based on the occurrence of Indonesian Restaurant in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new restaurant.

4. Results

5. Discussion

6. Conclussion