Cluster Postal Areas in Toronto Based on Food

Mengchao Jiang

September 2019

1 Introduction

1.1 Background

Toronto is one of the most popular city in world. It is the provincial capital of Ontario in Canada with large population and great number of travelers. It has a population of about 3 million in 2016 and a record of 43 million visitors in 2017. Meanwhile, Toronto is always an important destination for immigrants to Canada. His hospitality contributes to a population with great diversity. More than 50 percent of residents belong to a visible minority population group. The diversity of population also leads to a variety of different cuisines in Toronto.

Food location is always a major consideration while re-residence or travel. For those already live there or want to relocate to Toronto, it may be difficult for them to find the appropriate location close to their favourite cuisines as hundreds of cuisines are distributed in Toronto. For travelers, it is even harder to find the best places to get food they like with great convenience in the restricted time period. Therefore, it is meaningful to find out the distribution of different kinds of cuisines in Toronto.

1.2 Problem

This project is intended to find areas with similar cuisines thus people can easily find the place to live or go based on their personal preferences. We will divide areas in Toronto into several clusters with labels indicate most popular cuisines in each cluster. And those areas will shown on a map of Toronto with cluster labels.

2 Data Preparation

2.1 Data sources

In this project, two datasets and Foursquare API are used to get the geographical coordinates of postal areas in Toronto and top venues in each area. The first dataset is a list of postal codes with boroughs and neighbourhoods in Toronto. It can be scraped from wikipedia at following url: `https://`

en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. The second dataset contains geographical coordinates (latitudes and longitudes) of each postal codes in Toronto, which can be found at https://cocl.us/Geospatial_data. Foursquare API is used to get the most common venues of specified postal code and the category of each venue.

2.2 Data cleaning

In this project, we only research postal areas in boroughs whose name contain word 'Toronto', e.g. Central Toronto. First, the two dataset are combined into one table, which contains following features: Postcode, Borough, Neighbourhood, Latitude, Longitude. Then, top 100 venues in a radius of 500 meters of all postal codes are scraped using Foursquare API for further analysis. For each venue, it has a venue category given by Foursquare, like Pizza Place, Wings Joint, Pub etc. Those venues are very explicit and there are hundreds of venue categories. Foursquare also has a category hierarchy dividing all categories into 10 main category, which can be returned as a json file using its API. Since our focus is cuisines, we extract food venues which belong to food category.

After data wrangling, geographical information and most common food venues of each postal code are found and used to cluster areas in Toronto and visualize on a Toronto map.

3 Methodology

3.1 Exploratory data analysis

A map of Toronto with postal areas superimposed on it is created. Each red circle marker on the map represents a postal area. A total of 38 postal areas are visualized on the map and research in the project.

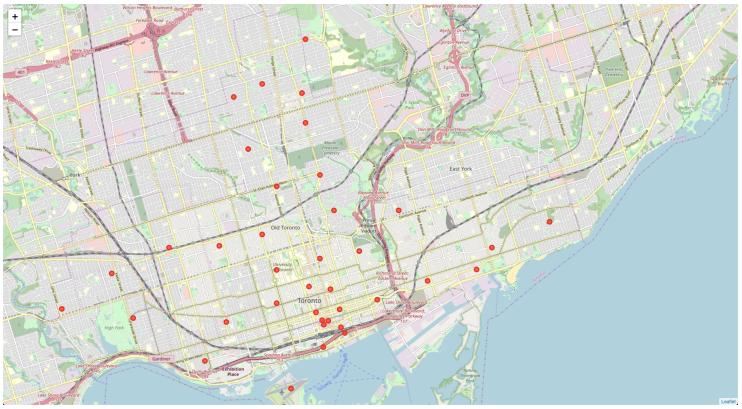

Figure 1: Postal Areas in Toronto

The number of different food venues are also counted. There are 88 different food categories in Toronto. 10 most popular food venues are shown in Figure 2.

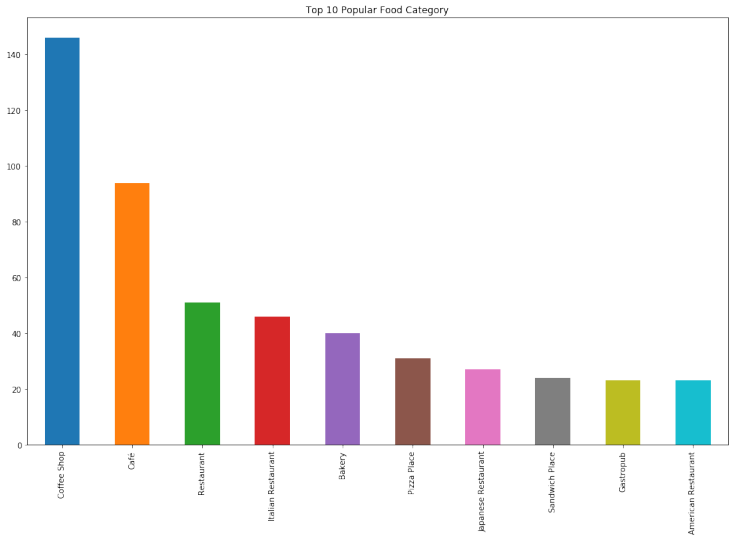

Figure 2: Top 10 Most Common Food Categories

Coffee Shop is the most popular venue category in Toronto, followed by Café. The top two categories both serve coffee products mainly and are much more popular than the rest cuisines. Cuisines like Italian, Bakery, Pizza are also very popular, followed by Japanese and Sandwich.

3.2 Model building

Since we will group all postal areas based on the cuisines, we use unsupervised learning algorithm K-means clustering. We use one hot encoding to convert venue category into features and get the mean value of each postcode. Then k-means clustering algorithm is applied to the data and distortions of different number of clusters are calculated. The optimal number of clusters is chosen from the elbow in Figure 3, then all postal areas will be divided into 4 clusters.

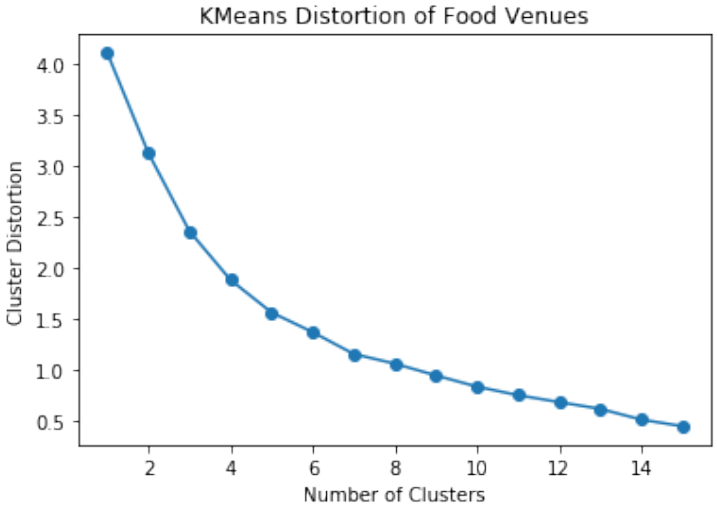

Figure 3: Distortions of K-Means Clustering

A table is also created to show the list of top venue categories in each postal area, which is combined with the table of cluster labels. Finally, we get a table of all postal areas with geographical coordinates, cluster labels, and the categories of top 5 venues of each area.

4 Results

All postal areas are grouped into 4 clusters by k-means and top venue categories are quite different in each cluster as shown below:

- Cluster 0: Coffee venues and restaurants
- Cluster 1: Cuisines with great diversity
- Cluster 2: Best for brunch and tea time
- Cluster 3: Fast-foods for quick meal or to go

Figure 4: Clustered Postal Areas in Toronto

A map of clustered postal areas is shown in Figure 4. Each circle marker represents a postal area, and different clusters are shown in different colors:

- Cluster 0: red
- Cluster 1: purple
- Cluster 2: green
- Cluster 3: yellow

Popup labels of postal code and cluster label are added to each circle marker.

5 Discussion

As we can see, most postal areas in Toronto are grouped into cluster 0, where coffee is most popular, which demonstrates that coffee is popular almost everywhere. These areas also have lots of other cuisines like Japanese and Italian. In the rest three clusters, coffee is not very popular, and only one area is contained in each cluster. Clusters are quite different from each other. Cluster 1 has cuisines with great diversity, cluster 2 consists of venues best for brunch and tea time, and cluster 3 has lots of fast-foods venues. Based on our analysis, postal areas in cluster 0 are recommended to travelers and residence like coffee. People like to try food from different country and culture can search food in cluster 0 and 1. For those have time to enjoy a brunch or tea time, cluster 2 is recommended, while cluster 3 is the place for busy ones to grab some fast food to go.

6 Conclusion

In this report, we clustered postal areas in Toronto based on popular cuisines in each area. We use different cuisines as feature and apply the k-means clustering algorithm to the data to group the areas into four clusters. Residence and travels can use the clusters to help them find the proper destination. For example, if someone is a lover of coffee, he should live in areas where can easily get a cup of coffee.