Capstone Project – The Battle of Neighborhoods Opening Chinese Restaurant in Toronto

Carol Wang

1. **Introduction :**

**Problem background:**

Toronto is Canada’s Largest city and a world leader in such areas as business, finance, technology, entertainment and culture. With a recoded population of 2,731,571 in 2016, it is the most populous city in Canada and the fourth most populous city in North America. Toronto’s large population of immigrants form all over the globe has also made Toronto one of the most multicultural cities in the world.

One of the best things about living in the world’s most multicultural city is all the tasty and diverse food options. The diversity of the cuisine available is reflective of the social and economic diversity of Toronto. Whether you are in the mood for Italian, Mexican, Chinese or Ethiopian, you won’t have a problem finding a delicious meal to satisfy your craving in Toronto.

**Business Understanding/Problem Description:**

In this project, we will find the best neighborhood in Toronto for opening a new Chinese restaurant.

There are some questions we need to address:

1. People live in which neighborhood likely eat in this kind of restaurant?
2. How many "similar" restaurants are available nearby?
3. Do the "similar" restaurants cost more? If so, what specialty do that have?

**Target Audience:**

Target audiences for this project does not limit to a person who keeps travelling but everyone. People could simply decide to look for a similar restaurant all the time because they are addicted to a specific category of food. People who rarely use restaurants would prefer to have the most rated restaurants nearby them and all this could be easily handled by our recommender system. So target for this project is basically everyone who is exploring different places or similar places.

1. **Data :**

To find a solution to the questions and build a recommender model, we need data and lots of data.

Here are some questions we need to consider but not limited to when gathering data to suggest a location for new restaurant:

1. Geographical coordinates (latitude and longitude) to find out where exactly it is located.

2. Population of the neighborhood where the restaurant is located.

3. Average income of neighborhood to know how much is the restaurant worth.

**Data Source:**

1. List of postal code for communities in Toronto from Wikipedia:

<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

1. Geospatial Coordinates CSV file for Toronto postal codes from <http://cocl/us/Geospatial_data>
2. Foursquare API venues explore methods to get venues of given neighborhoods
3. Foursquare API venues methods to get ranks and likes of restaurants by given venue id

**Methodology:**

1. I will use of demographic data by neighborhoods, which allows me to gauge people’s taste in each neighborhood. For example, I assume areas with a majority of Chinese people would be prefer to go to Chinse restaurant.
2. I will use the Foursquare API to explore neighborhoods in Toronto and apply the explore function to get the most common venue categories in each of the neighborhoods.
3. I will then use this feature to group the neighborhoods into clusters and use the k-means clustering algorithm to refine the groupings.
4. Finally, I will use the Folium library to visualize the neighborhoods in Toronto and their emerging clusters as well as push the right neighborhood for setting up the restaurant business.
5. **Exploratory Data Analysis:**

**Data Cleaning:**

I will use BeautifulSoup library to scrape Wikipedia to list of postal codes for each community. Remove the empty data in each row and group each

zip code with all neighborhoods:

**A screenshot of a cell phone

Description automatically generated**

Then geospatial data (coordinates) is read into pandas dataframe. Then joined with the previous dataframe to get a new dataframe with zip code, neighborhood, latitude, and longitude coordinates.

A screenshot of a cell phone

Description automatically generated

Then we can plot a map of the neighborhoods using Folium library:

A close up of a map

Description automatically generated

**Foursquare API**

I will use Foursquare API to explore neighborhoods and merge it to dataframe including neighborhoods and top-rated venues.

A screenshot of a cell phone

Description automatically generated

**Features**

When extracting features from a dataset, it is often useful to transform categorical features into vectors. Then group them by neighborhood by taking the mean of the frequency of occurrence.

A screenshot of a cell phone

Description automatically generated

Then a dataframe is created with top types of venues around each neighborhoods in Toronto:

A screenshot of a cell phone

Description automatically generated

1. **Clustering Modeling and Recommendations:**

In clustering, I try to find homogeneous subgroup within our data such that data points in each cluster are as similar as possible according to a similarity measure such as Euclidean-based distance or correlation-based distance.

Clustering is considered an unsupervised learning method since we don’t have the ground truth to compare the output of the clustering algorithm to the true labels to evaluate its performance. In this project, I use K-means clustering, which is a type of unsupervised learning, which is used when have unlabeled data. The K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups where each data point belongs to only one group. In this case, the clustering is based on types of nearby venues.

Then data is grouped into 5 clusters. The clusters and the top 10 venues for each neighborhood are shown below:

A screenshot of a cell phone

Description automatically generated

A close up of a map

Description automatically generated

Number of restaurant in each of the 5 clusters are listed below:

A close up of a logo

Description automatically generated

Then dataset is grouped by clusters to get the means of the frequency occurrence of each cluster, and examined by a heatmap.

A close up of a sign

Description automatically generated

A screenshot of a social media post

Description automatically generated

1. **Conclusions:**

Conclusions can be taken from the previous maps, tables, and exploratory data analysis. For example, cluster 1 and 4 has high frequency of Chinese restaurants, so try not to open a new Chinese restaurant in cluster 1 or 4 neighborhoods.