COURSERA CAPSTONE PROJECT

May 29, 2020

Rongrong Wu
IBM Data Science Professional Certificate

Contents

0.1 Introduction/Business Problem 2
0.1.1 Business Problem 2
0.1.2 Target Audience of this project 2
0.2 Data 3
0.3 Methodology 4
0.4 Results 5
0.5 Conclusion 6

0.1 Introduction/Business Problem

I live in Chicago, and the best thing I love about Chicago is the diverse culture, and tons of great restaurants it has. The goal of the project is to find the ideal place to open a Chinese Restaurant in Chicago. The project consists of two parts. Part I is the segmentation and clustering of Chicago Restaurants, including types of Italian, French, Japanese and Chinese restaurants. Part II is about choosing the ideal location to open the new Chinese restaurant. Of course, as with any business decision, opening a new restaurant requires serious consideration and is a lot more complicated than it seems. Therefore this project aims to help alleviate the troublesome process of finding the ideal location.

0.1.1 Business Problem

The objective of this capstone project is to find the ideal place to open a Chinese Restaurant in Chicago. Using data science methodology and machine learning techniques such as k-means clustering etc. This project aims to provide in-depth research and thorough analysis.

0.1.2 Target Audience of this project

This project is particularly useful to property develoopers and investors looking to open or invest in new restaurants in Chicago. This project is timely as the city is currently flourishing with good restaurants. Research shows that people are more and more willing to spend money on fine dining restaurants.

0.2 Data

To solve the problem, we will need the following data:

1. List of Community Areas in Chicago. This defines the scope of this project which is confined to the city of Chicago, major city in the Middle West of the United States of America.

2. Latitude and longitude coordinates of those community areas. This is required in order to plot the map and also to get the venue data.

3. Venue data, particularly data related to different types of restaurants. We will use this data to perform clustering on the Community areas.

Sources of data and methods to extract them:

1. The wikipedia page (https://en.wikipedia.org/wiki/Community_areas_in_Chicago) contains a list of community areas in Chicago, with a total of 77 community areas. 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 using Python Geocoder package which will give us the latitude and longitude of the community areas.

2. We will then use Foursquare API to get the venue data for those neighborhoods. 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 restaurant category in order to help us solve the business problem put forward. The project will make use of many data science skills, from web scraping(Wikipedia), working with API(Foursquare), data cleaning, data wrangling, to machine learning and data visualization(Folium). In the Methodology section, we will discuss the analysis in detail.

3. We use k-means clustering to cluster the neighborhoods with Chinese restaurants and find 3 feasible regions for our location choice for the new restaurant.

0.3 Methodology

First, we use the packages BeautifulSoup to scrap the list of community areas of Chicago in the Wikipedia page. Then we need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. We use Geocoder package in Python to convert address into geographical coordinates in the form of latitude and longitude. We use dataframe to store the resulted Community area list.
We get the visualizations of all the community Areas:

Figure 1: Community Areas in Chicago

Then, we use Foursquare API to get the venues that are inside the region of our community area. We use the Foursquare Developer Account and Secret Key to make API calls to Foursquare passing information. Foursquare will return the venue data in Json format:

1. List of Chinese Restaurants
2. List of Japanese Restaurants
3. List of French Restaurants

We then extract the venue name, category, latitude and longitude. We filter the information and then store them in pandas dataframes and visualize the results using python folium package. By doing so, we are also preparing the data for use in clustering. Since we are analysing the 'restaurant' data. we restrict the categories to be restaurants.

We use red color to represent the community areas, blue to represent Chinese Restaurants, green to represent Japanese restaurants and yellow to represent French restaurants. See the figure below.

Figure 2: Restaurants in Chicago

0.4 Results

We then use k-means clustering to cluster the community areas in Chicago into 3 clusters. So that we could determine the best location to open a Chinese restaurant based on the density of existing restaurants in different areas of Chicago. We again use Folium to visulize the Clustering result:

1. Cluster 0: Blue
2. Cluster 1: Green
3. Cluster 2: Red

The plot is the following:

Figure 3: Clustering result

0.5 Conclusion

Most of the Chinese restaurants are concentrated in the central area of Chicago, with the highest number in cluster 0 and moderate number in cluster 2. On the other hand, cluster 1 has very low number to totally no Chinese restaurants in the neighborhoods, which seems to be an outlier.

This represents a great opportunity and high potential areas to open new restaurants in Cluster 2 areas since there will be high competition in areas like downtown Chicago in Cluster 0. Cluster 1 will result in lack of customers since there are no restaurants nearby.

Therefore, this project recommends property developers to capitalize on these findings to open new restaurants in cluster 2 which has moderate competition and potential customers.

Property developers with unique selling propositions to stand out from the competition can also consider open new restaurants in Cluster 0 areas since the crowdedness of already opened restaurants indicates there are many potential customers. Existence of places like ChinaTown will attract customers and bring cash flows to the restaurants. Lastly, property developers are advised to avoid neighborhoods in Cluster 1.

References

1. Category: Community Areas in Chicago
   https://en.wikipedia.org/wiki/Category:Community_areas_of_Chicago

2. Foursquare Developers Documentation
   https://developer.foursquare.com/docs