Report: Locating Chinese Districts in Any City

Nan He

June 2021

1 Introduction

1.1 Background

More and more Chinese works and study overseas in the United States, Canada, and other western countries in recent decades. Oversea Chinese population in the US grow from $3,347,229$ in 2010 to $4,143,982$ in 2020.[1, 2] The growing presence of these new Chinese workers has a significant impact on existing local Chinese communities in major western cities. Research suggested that the newcomers help to revitalize and transforming old Chinatown. [3] They also boost many non-traditional Chinatowns or "Chinese Districts" in various middle-sized cities where the local Chinese population is small.

For international Chinese students like me, the existence of a local Chinatown or "Chinese districts" is important for residential decisions. The combination of Authentic Chinese and Asian food, Asian supermarkets, and various services in the Chinese language provides a smooth cultural transition for many newcomers. Previous publications support my personal experiences.[4] This report focuses on finding the opportunities inside this trend.

1.2 Business Problem

One major problem of the Chinese districts in middle-sized western cities is that their distributions are generally obscure, especially for someone who is not Chinese. When I visited Charlotte and Atlanta the first time, when discussing which areas have good access to Chinese services and goods, every person almost gave different answers. Some have already lived there for over five years while still have no clear clue whether a "Chinese District" exists and where it is. Inspired by these experiences, in this report, I will investigate this problem and try to build a tool to find the location of an existing Chinese District for any given city using a data-driven approach.

1.3 Significance

This problem has values for two types of users. Firstly, investors can identify potential investments in Chinese venues, especially those who are not local to

that city or are not familiar with the Chinese communities. Secondly, for other Chinese students and oversee works who seek to find a comfortable environment in an unfamiliar city.

2 Data Acquisition and Cleaning

Two types of data are needed to achieve our goal of finding Chinese districts automatically for a given city or area,

1. Data that tells us what a Chinese district should look like.
2. A target city or area the user wants to investigate.

2.1 Chinese District Feature Data

For data 1, I retrieve the location of 5 large Chinatown in the US from this Wikipedia page. Note that ten are listed on that page, and many of them cluster on the west coast. So I choose only New York, San Francisco, Los Angeles, Chicago, and Huston, to make the data more balanced geographically. A good starting point to study what defines a Chinese district is from the venues in that area. Therefore, to extract venue data for those areas, I utilize the Foursquare API, a popular location service provider based in the US.[5]

The latitudes and longitudes of the five selected Chinatowns are:

	City	Latitude	Longitude
0	New York	40.715457	-73.996841
1	Chicago	41.850927	-87.634770
2	San Francisco	37.794137	-122.406847
3	Huston	29.703955	-95.546077
4	Los Angeles	34.062328	-118.238343

Some example resulting venues are:

	City	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	New York	40.715457	-73.996841	Eye Solutions	40.714927	-73.998774	Optical Shop
1	New York	40.715457	-73.996841	Hotel 50 Bowery	40.715936	-73.996789	Hotel
2	New York	40.715457	-73.996841	Xi'an Famous Foods	40.715232	-73.997263	Chinese Restaurant
3	New York	40.715457	-73.996841	Zu Yuan Spa	40.715469	-73.998627	Spa
4	New York	40.715457	-73.996841	The Original Chinatown Ice Cream Factory	40.715521	-73.998145	Ice Cream Shop

In total, 429 venues are selected around these five Chinatowns; those data will tell us what defines a Chinese district.

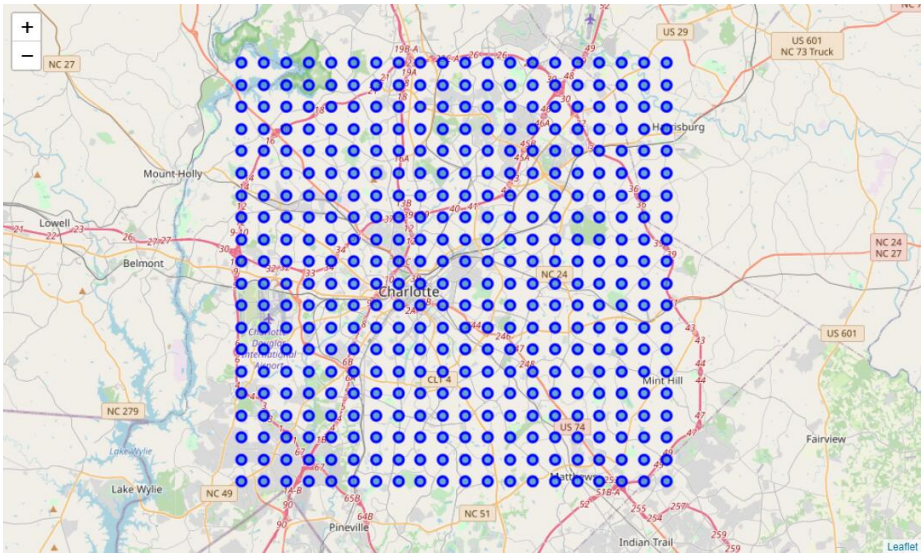
The data will be grouped by the city name and encoded in terms of the venue category. Examples look like:

	City	American Restaurant	Animal Shelter	Art Gallery	Asian Restaurant	Athletics & Sports	BBQ Joint	Bakery	Bank	Bar	Beer Garden	Beijing Restaurant	Bike Shop	Boat or Ferry	Bookstore	
0	Chicago	0.00	0.00	0.000000	0.058824	0.00	0.014706	0.029412	0.000000	0.000000	0.014706	0.000000	0.00	0.014706	0.00	(
1	Huston	0.00	0.00	0.000000	0.073529	0.00	0.000000	0.044118	0.029412	0.000000	0.000000	0.014706	0.00	0.000000	0.00	(
2	Los Angeles	0.00	0.00	0.011111	0.000000	0.00	0.011111	0.044444	0.000000	0.011111	0.000000	0.000000	0.00	0.000000	0.00	
3	New York	0.01	0.01	0.000000	0.030000	0.00	0.000000	0.070000	0.000000	0.000000	0.000000	0.000000	0.01	0.000000	0.00	(
4	San Francisco	0.01	0.00	0.000000	0.000000	0.01	0.000000	0.020000	0.000000	0.020000	0.000000	0.000000	0.00	0.000000	0.02	(

2.2 Target City Data

For data 2, the situation is more complicated. For each target city, the data sources of neighborhoods vary, making the web-scraping work city-specific. Here I choose a simple alternative approach. For each input coordinates user interested, we generate a grid around that latitude and longitude coordinate and analyze the target city based on this grid partition. Here I take Charlotte, NC, as an example to show how this approach works.

We select two points on google map to select areas include most of the metropolitan area of Charlotte. The grid is constructed by equally partitioning 20×20 sample coordinates within the chosen area.



Then we loop over those coordinates and find venues within a certain radius of each point. Note that we will make that radius slightly larger than the half-distance of the two diagonal grid points to make sure that no venues are ignored. In the Charlotte case, the radius is 1500 m. By taking this approach, there will

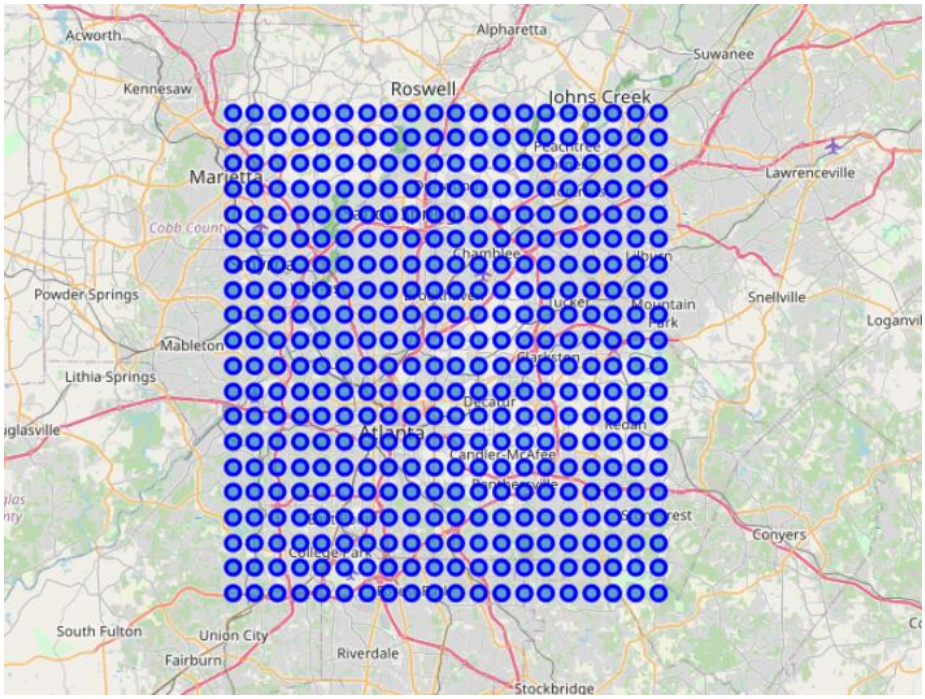
also have overlapping venues, where the same venues appear in different grid points. Since the grid points can locate in sparsely populated areas, where the number of venues is tiny, I will ignore them to avoid potential statistical problems. Only grid points with over 15 venues will be appended into the data.

Example venues look like:

	Location Index	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	0	35.11311	-80.965507	Cajun Canvas	35.116538	-80.962047	General Entertainment
1	0	35.11311	-80.965507	Wing King	35.116816	-80.956874	Wings Joint
2	0	35.11311	-80.965507	A Piece of Havana	35.116673	-80.962459	Cuban Restaurant
3	0	35.11311	-80.965507	Jersey Mike's Subs	35.116783	-80.962137	Sandwich Place
4	0	35.11311	-80.965507	Papa John's Pizza	35.116785	-80.962353	Pizza Place

In total, 10041 venues are selected. Those data are where we want to get insight from.

Here I show another example to prove the workflow. Find Atlanta on google map, and select the Input point 1 is $33.620260, -84.543165$, point 2 is $34.024354, -84.112095$, and the sample size is 20×20:



The data will be grouped by grid point indices and encoded in terms of the venue category. Examples look like:

	Location Index	ATM	Accessories Store	Adult Boutique	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Arcade	Art Gallery	Art Museum	Arts & Crafts Store	En
0	0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	
1	1	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.017857	0.0	0.0	0.0	0.0	0.0	
2	2	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.023256	0.0	0.0	0.0	0.0	0.0	
3	3	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.047619	0.0	0.0	0.0	0.0	0.0	
4	4	0.022222	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.044444	0.0	0.0	0.0	0.0	0.0	

Then the data are ready to be analyzed.

2.3 Usage of the Data

The most straightforward idea is to use the unsupervised learning techniques only based on the target-city data. In this case, we do not need data 1. However, Chinese venues are not typical features for many US cities; it is highly possible that naive unsupervised learning techniques cannot distinguish the Chinese district from other districts. Therefore, data 1 will play a vital role in screening the venue types and possibly introducing weighted distances during unsupervised learning.

3 Methodology

3.1 Naive K-means Clustering

3.2 Feature Analysis

3.3 Weighted Distance K-means

4 Result and Discussion

4.1 Results of Charlotte and Atlanta

4.2 The Pipeline

4.3 Potential Problems

5 Conclusion

References

[1] Chinese diaspora across the world: A general overview, 2010.

[2] Overseas chinese, 2020.

[3] Jackie Jia Lou. Chinatown transformed: Ideology, power, and resources in narrative place-making. *Discourse Studies*, 12(5):625–647, 2010.

[4] Min Zhou. *Chinatown: The socioeconomic potential of an urban enclave.* Temple University Press, 2010.

[5] Foursquare main site, 2021.