Clustering neighborhoods in San Francisco and Houston

By Arnold Jiadong Yu

Introduction

Background

- With the development of technologies, travel became affordable and easy to access. Especially in the industry of information technologies, most work can be done remotely by only one computer. Therefore, more and more people start to travel around and experience various cultures of cities and countries while working remotely. However, a lot of researches need to be done before moving to next location to ensure best experience. As a result, it is advantageous for individuals to compare neighborhoods of various cities. This will not only save up a lot of time, but also give an initial idea of how the neighborhoods are formed. For example, this can give a person guide if he or she wants to move to Pairs for a month.

Problem

- A person has enough experience of the current city, which neighborhood he likes or doesn’t like. Now he wants to move to another city. Before he moves, he needs to find out how similar or dissimilar between neighborhoods in both cities.

Interest

- Individuals who like traveling will definitely like this idea. Travel companies can offer intense travel plans based on this idea such as housing, transportation, and etc.

Challenges

- A lot of data is needed to perform a rich and detailed comparison between two cities such as venues, real estate, population density, population variety, transportations, food variety, others’ tips, recommendations, and etc. Unfortunately, it is impossible for me to obtain all information. Therefore, this project preforms a basic comparison using venues of each neighborhood.

Dataset

Data Source

- Neighborhood data of two cities can be obtained from Wikipedia.
- Venues of each neighborhood can be obtained using Foursquare API.

Data Cleaning

- Neighborhood data need to be extracted from webpages using BeautifulSoup and put into a desirable format. There are total 123 neighborhoods extracted from San Francisco Neighborhood data and 88 neighborhoods extracted from Houston Neighborhood data. Afterwards, latitude and longitude are extracted of each neighborhood by using geopy library. Venues of each neighborhood are collected using its latitude and longitude. A dataframe is built using all venues of both cities respect to neighborhoods. Since not all latitude and longitude of each neighborhood can be extracted. We end up with a total number of 132 neighborhoods with its latitude and longitude.

	Neighborhood	Latitude	Longitude
0	Anza Vista	37.780836	-122.443149
1	Balboa Park	37.724949	-122.444805
2	Balboa Terrace	-38.730438	-62.233556
3	Bayview	37.728889	-122.392500
4	Belden Place	37.791744	-122.403886

	Neighborhood	Latitude	Longitude
0	Willowbrook	29.660254	-95.456096
1	Greater Greenspoint	29.944719	-95.416074
2	Carverdale	29.848687	-95.539450
3	Fairbanks	29.852726	-95.524386
4	Acres Home	32.636256	-83.692962

Dataset

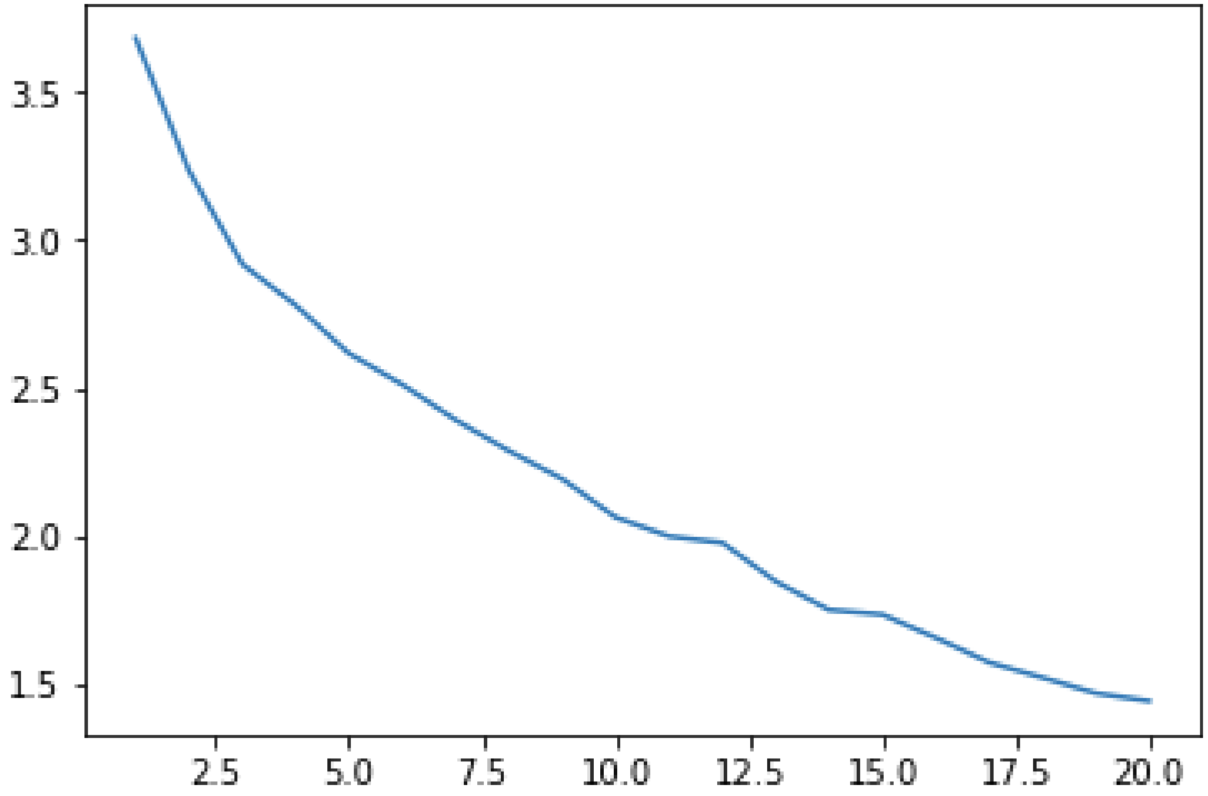
A filter is applied to each latitude and longitude to ensure that all the information which extracted is relevant to San Francisco and Houston. Then a map is built to ensure that the filter is successfully applied

Dataset

- Then venues of each neighborhood are extracted using Foursquare API using latitude and longitude of each neighborhood with a radius 500 meter and limit 100 venues constraints. There are a total number of 4498 venues of all neighborhoods.
- A detailed count is performed to check number of venues in each neighborhood.
- Since each neighborhood have a different number of venues, it is strongly biased. As a result, we need to change the parameter limit and radius to have a close number of venues in each neighborhood. Limit 100 and Radius 2000 is used. There are 10901 venues extracted and 391 unique venues.

Methodology

- The Elbow method is used to choose the suitable k. A graph of k from 1 to 20 is plotted. K = 14 is chosen.



Result

- All results are printed based on their cluster. Below are clusters of 0 to 1. All clusters can be referenced to notebook on GitHub.

	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	...	11 Co Ve
60	0	Coffee Shop	Hotel	Park	Pizza Place	Mexican Restaurant	Gym	Cocktail Bar	Sandwich Place	Burger Joint	...	Gy Fit Ce
93	0	Hotel	Park	Coffee Shop	Sandwich Place	Mexican Restaurant	Southern / Soul Food Restaurant	Italian Restaurant	Pizza Place	Burger Joint	...	Ba
95	0	Hotel	Park	Coffee Shop	Sandwich Place	Mexican Restaurant	Southern / Soul Food Restaurant	Italian Restaurant	Pizza Place	Burger Joint	...	Ba
99	0	Hotel	Mexican Restaurant	Coffee Shop	Park	Southern / Soul Food Restaurant	Italian Restaurant	Sandwich Place	Pizza Place	Burger Joint	...	Ste
103	0	Coffee Shop	Park	Cocktail Bar	Gym	Pizza Place	Bar	Hotel	Theater	Southern / Soul Food Restaurant	...	Art

	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	...
3	1	Coffee Shop	Boutique	Garden	Bookstore	Bubble Tea Shop	Hotel	Gym	Sushi Restaurant	New American Restaurant	...
5	1	Coffee Shop	Art Gallery	Yoga Studio	Gym / Fitness Center	Art Museum	Vietnamese Restaurant	Park	Wine Shop	Baseball Stadium	...
6	1	Coffee Shop	Gym / Fitness Center	Sushi Restaurant	Cocktail Bar	Theater	Dance Studio	Gym	Art Gallery	Marijuana Dispensary	...
13	1	Coffee Shop	Food Truck	Wine Bar	Gym	Seafood Restaurant	New American Restaurant	Museum	French Restaurant	Liquor Store	...
16	1	Coffee Shop	Wine Bar	New American Restaurant	Bookstore	Food Truck	Gym	Boutique	Sushi Restaurant	Art Museum	...

Conclusion

- As a result, the map showed clusters in both cities and can offer an idea of how similar the neighborhoods are.