Predicting the Similarity between Areas in Different Cities

Research Report

Nikita Ravi

Words:

Contents

Introduction

Background

The world is becoming more globalized day by day. With an increase in the rate at which information and ideas are exchanged, people are becoming more aware of the infinite number of cultures that exist in our society, and etc. Keeping this in mind, many multinational companies take this globalization as an opportunity and aim to open their branches in different parts of the world for better profit and this accounts for increased Gross Domestic Product (GDP) and better stabilizes the host country's economy. As of 2006, there were 63,000 recorded multinational corporations with over 700,000 branches worldwide according to the United Nations Conference on Trade and Development (Shoo, 2017).

Problem

In order to start a chain in another city of another country, often times the company would like to choose a location that is similar to that of the original location of the first chain store because it is the best way to prevent an economic loss and gain profit. This project aims to predict the similarity between places in different cities, by recording the different types of venues in the area to assess the location's entertainment preferences. This way, if a company wants to start a branch in another area in a different part of the world, they can choose an area similar to that of their original location so that they can potentially target the same audience and gain more profit / minimize economic loss.

Interest

This report is targeted to multinational corporations who plan on opening a new branch in a different part of the world and help them assess which location would be the best at minimizing the company's economic loss.

Data Acquisition and Cleaning

Data Sources

The postal codes and area names for Tokyo, Japan were obtained from UPS website. For London, the data was obtained from the Doogal website, Toronto information was retrieved from Wikipedia, and the Los Angeles information was obtained from a travel website.

Data Cleaning and Feature Selection

The data on Japan downloaded from the UPS site had to be cleaned by getting rid of unwanted information such as an introduction to the data on the csv file. This was crucial to do so because otherwise unwanted information might appear in the Pandas DataFrame when calling on the *pd.read_csv()* function. However, there were no missing values in the data UPS provided. Some columns were removed while reading the data set as they were insignificant to this program. Finally, two more columns were added to include

	Area Name	Latitude	Longitude
0	Chiyoda	35.693810	139.753216
1	Chuo	35.666255	139.775565
2	Minato	35.643227	139.740055
3	Taito	35.717450	139.790859
4	Bunkyo	35.718810	139.744732

Figure 1 - Tokyo DataFrame

information on the latitude and longitude of each location.

	Post Code	Area Name	Latitude	Longitude
0	E1 6AN	Bishopsgate	51.518895	-0.078378
1	E1 7AA	Portsoken	51.515567	-0.075635
2	E1 7AD	Portsoken	51.515457	-0.076718
3	E1 7AE	Portsoken	51.515613	-0.076899
4	E1 7AF	Portsoken	51.515613	-0.076899

Figure 2 - London DataFrame

The information on London from the Doogal website needed no cleaning as all the information was there with no introductions and etc. However, some columns were removed while reading the data as they were unnecessary to the program.

The information obtained about Toronto from Wikipedia had a lot of missing information. For instance, if the borough name was not assigned, then that entire row was dropped. In addition to this, if the neighborhood column has a value of not assigned, then it was replaced by its Borough name. Also, the last extra character '\n' was removed from each neighborhood name. Furthermore, the same boroughs with different neighborhoods are combined and put into one row. Two extra columns were later included to indicate the longitude and latitude of each borough and the neighborhood columns were dropped as it was redundant relative to the borough names.

	Area Name	Latitude	Longitude
0	Scarborough	43.773077	-79.257774
1	North York	43.770817	-79.413300
2	East York	43.691339	-79.327821
3	East Toronto	43.624790	-79.393492
4	Central Toronto	43.653963	-79.387207

Figure 3 - Toronto DataFrame

	Area Name	Latitude	Longitude
0	Bell	33.974781	-118.186636
1	Bell Gardens	33.969456	-118.150395
2	Huntington Park	33.982704	-118.212034
3	Los Angeles	34.053691	-118.242767
4	Maywood	33.986681	-118.185349

Figure 4 - LA DataFrame

The Los Angeles dataset was cleaned up to only include the city name while reading the table. The latitude and longitude columns were later included.

Methodology

After collecting the data necessary for each city that is being investigated, the Foursquare API to get the venue data for each area of each city. The program collects information on the top 100 venues in a radius of 500m. By making use of API calls, Foursquare will return the venue data in a JSON file with information on Venue Name, Venue Latitude, Venue Longitude, and Venue Category. All this information is compiled into a DataFrame. An example is shown below of Tokyo's venues:

	Area Name	Area Latitude	Area Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Chiyoda	35.69381	139.753216	Nippon Budokan (日本武道館)	35.693356	139.749865	Stadium
1	Chiyoda	35.69381	139.753216	Kitanomaru Park (北の丸公園)	35.691653	139.751201	Park
2	Chiyoda	35.69381	139.753216	Kanda Tendonya (神田天丼家)	35.695765	139.754682	Tempura Restaurant
3	Chiyoda	35.69381	139.753216	Jimbocho Kurosu (神保町 黒須)	35.695539	139.754851	Ramen Restaurant
4	Chiyoda	35.69381	139.753216	Shimizumon Gate (清水門)	35.692685	139.752681	Historic Site

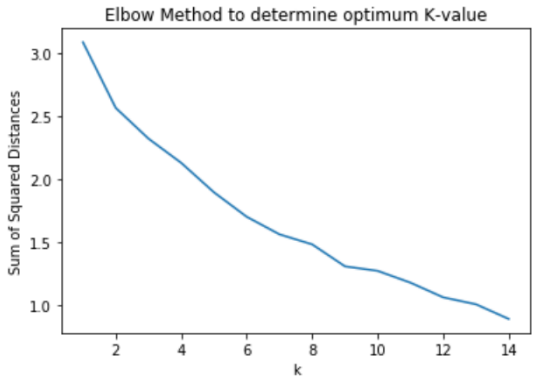
Figure 5 - Tokyo's Venues in each Area

After getting tables like the one in figure 5 for each city, merge all of them to form one huge giant table which is grouped by Area name and illustrates the mean frequency of each venue category at each location in each city, as shown below:

	Area Name	ATM	Accessories Store	African Restaurant	American Restaurant	Arcade	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	...	Video Store	Vietnamese Restaurant	Wagashi Place	Waterfall	Wine Bar	Win Sho
0	Adachi	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.0	0.0	...	0.0	0.00	0.0	0.0	0.00	0.0
1	Aldersgate	0.0	0.0	0.0	0.0	0.0	0.01	0.01	0.0	0.0	...	0.0	0.02	0.0	0.0	0.03	0.0
2	Aldgate	0.0	0.0	0.0	0.0	0.0	0.01	0.02	0.0	0.0	...	0.0	0.02	0.0	0.0	0.03	0.0
3	Arakawa	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.0	0.0	...	0.0	0.00	0.0	0.0	0.00	0.0
4	Bell	0.0	0.0	0.0	0.0	0.0	0.00	0.00	0.0	0.0	...	0.0	0.00	0.0	0.0	0.00	0.0

Figure 6 - Final Combined DataFrame

Using the data from figure 6, we now cluster the data by using the K-means clustering method. The K-means clustering algorithm finds all the k-number of centroids, and then assigns every data point to the nearest cluster. The feature variables are all the venue categories and by using these columns, each Area Name is clustered with one of k-number of clusters.



To choose the optimum number of clusters, the elbow method is used. The elbow method involves graphing out the Sum of Squared Distances (SSD) of each cluster point to its respective centroids, against each k-value. The optimum k-value is later determined to be where the SSD doesn't change as much on a drastic level.

Figure 7 - Elbow Method

Results

The results from this type of unsupervised machine learning algorithm shows there are nine cluster groups. Each different cluster is given a different color with cluster 0 being red, cluster 1 being purple, cluster 2 being blue, cluster 3 being light blue, cluster 4 being turquoise, cluster 5 being teal, cluster 6 being green, cluster 7 being light orange, and cluster 8 being dark orange. The different clusters is a visualization of how each area in each city are similar or dissimilar to each other as shown below:

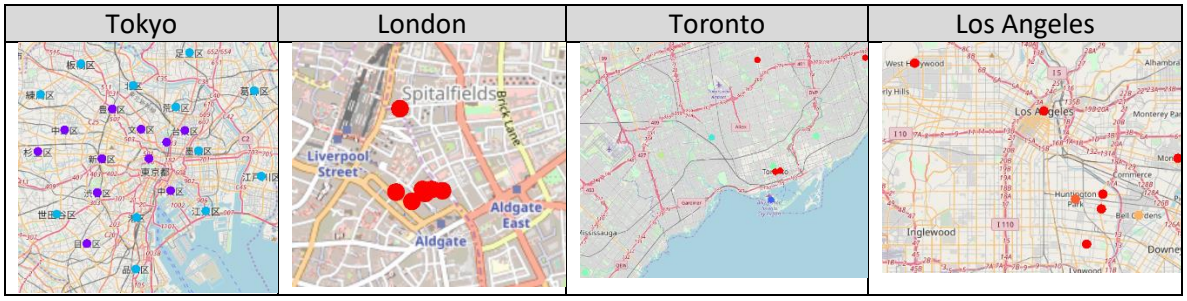


Figure 8 - Visual Clustering of Data

Conclusions

From the results shown above, majority of the areas in London and Los Angeles are in cluster 0, and some of the areas in Toronto are in cluster 0 and this indicates that these locations are similar in terms of the types of venues they share. Only one location in Los Angeles is part of the same cluster, cluster 1, as majority of the locations in Tokyo. Hence, if a company, with its main branch being in London, wants to open a different branch in a different part of the world, it is more statistically appropriate for the branch to be opened in Toronto and Los Angeles as potentially the company may be targeting the same type of audience, thus maximizing profit.

Limitations and Future Directions

In this project, only the venue categories were taken into account to assess the similarities between areas of different cities. However, this may not be the only factor that needs to be taken

into account when considering the similarity between locations. Future research for this project may have to include a methodology to investigate similarity or dissimilarity between the cultural background, history, venue categories of each particular location and see if there is an improvement in predicting the optimum location of opening a new branch for a multinational corporation.

References

Shoo, D. (2017, September 26). *Economic Effects of Multinational Corporations*. Retrieved from bizfluent: https://bizfluent.com/info-8444236-economic-effects-multinational-corporations.html