Live in a similar neighborhood-From Seoul to New York and Toronto

Lee Dong UK
SNU

1. Introducion

A lot of people plan to move. There are also many people who move abroad. There are lots of people who study abroad. Some people leave for business. What they want when they move is a familiar environment. Living in a familiar environment will help you adapt faster. Anyone who lived near the market would want the market to be close to the neighborhood they moved to. The person who often goes to the art museum in the original neighborhood would want to have an art museum in the neighborhood where he moved to. Therefore, it is important to classify and compare the characteristics of neighborhoods in particular cities through machine learning.

1.1 Background

Seoul has a different administrative division from New York. The city of Seoul is divided into 'Gu'. The 'Gu' is divided into several 'Dong'. I thought of 'Dong' as New York's 'Neighbourhood'.

1.2 Problem

1. Suppose Tom lives in Sangil-dong, Gangdong-gu, Seoul. He will move to New York. He wants to move to a neighborhood in New York that is as similar to Sangil-dong as possible. Where should he move to? Where should he go if he move to Toronto? What about from New York to Toronto? What about from Toronto to New York?

2. The neighborhood where Tom will move will belong to a particular cluster. What characteristics does the cluster have? Is there an easy way to visualize it?

[76]:

	Borough	Cluster Labels	first	second	third	forth	fifth	6th	7th	8th	9th	10th
18	Central Toronto	0	Park	Bus Line	Dim Sum Restaurant	Swim School	Distribution Center	Falafel Restaurant	Event Space	Ethiopian Restaurant	Escape Room	Electronics Store
21	Central Toronto	0	Park	Trail	Jewelry Store	Bus Line	Sushi Restaurant	Yoga Studio	Distribution Center	Event Space	Ethiopian Restaurant	Escape Room

2. Data arrangements

2.1 Data sources

Seoul Distinct(Gu, Dong)

-From wikipedia using beautifulsoup

-https://en.wikipedia.org/wiki/List_of_districts_of_Seoul

Toronto Distinct(postalcode)

-From wikipedia using beautifulsoup

-https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Toronto Postalcode and coordinate

-from CSV file from this course

NewYork Data

-Data and json file from this course

Seoul Distinct coordinate

-GeoCoder

Kinds of venue in Seoul, Toronto, NewYork

-Foursquare API

3. Method

Now let's compare New York and Toronto. Both New York and Toronto were clustered in five. However, it is difficult to see at a glance what criteria each cluster is clustered on. Therefore, I decided to score arbitrarily. Suppose there are a lot of Bakery in cluster 1st Most Common Value. Then this cluster can be thought of as a place where Bakery is gathered. One more thing needs to be considered. Suppose that there is one Neighborhood whose 1st Most Common Value is bakery. But there are 12 Neighborhoods whose 2st Most Common Value is Hotel. Then you have to conclude that the cluster is full of hotels. Thus, the nth Most Common Value was counted for certain items and multiplied by 11-n. For example, if there are three Neighbourhoods whose 7th Most Common Value is 'Cafe'. Cafe gets 3*(11-7)=12 points. Sum all these scores together. And the higher the score, the more representative the cluster. It is represented by a histogram. This allows you to view the characteristics of each cluster at a glance.

Now let's get down to business. Let's find a cluster in New York that best suits people in certain parts of Toronto. However, the number of categories in Toronto and New York was different. Therefore, I decided to adjust to the categories in New York. The categories in Toronto but not in New York have decided to throw them away. If so, data preprocessing is done. If you put Toronto's data into the Kmeans model in New York and use the predict function, you will see clusters. In contrast, the New York data was placed in the Toronto Kmeans model. All New York neighborhoods matched cluster1 in Toronto, according to the results.

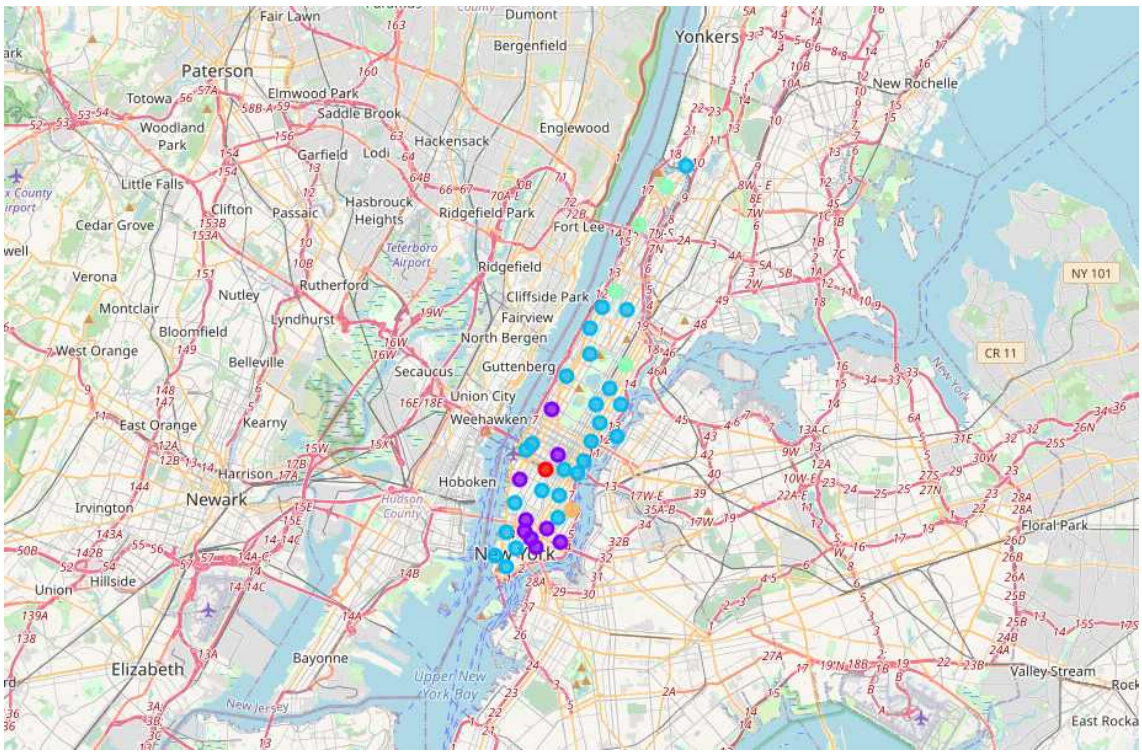
The analysis of Seoul was conducted in the same way. The data frame was created using the coordinates obtained from the Seoul Metropolitan Government and

GeoCoder received from wikpidea. Based on the coordinates, we obtained information from nearby stores at FourSquare. The uninformed areas were removed. After One Hot Encoding, it was inserted into the Toronto clustering model and the New York clustering model to obtain the clustering label. And I put them on the map.

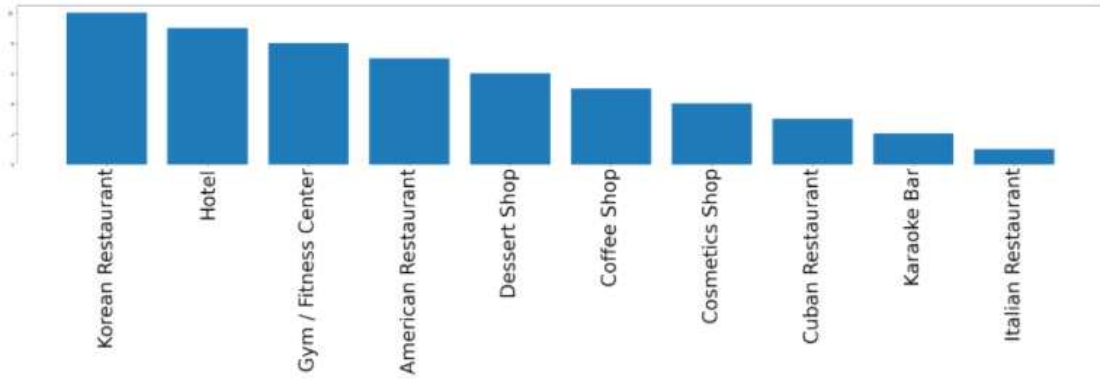
4. Results

4.1 New York Clustering

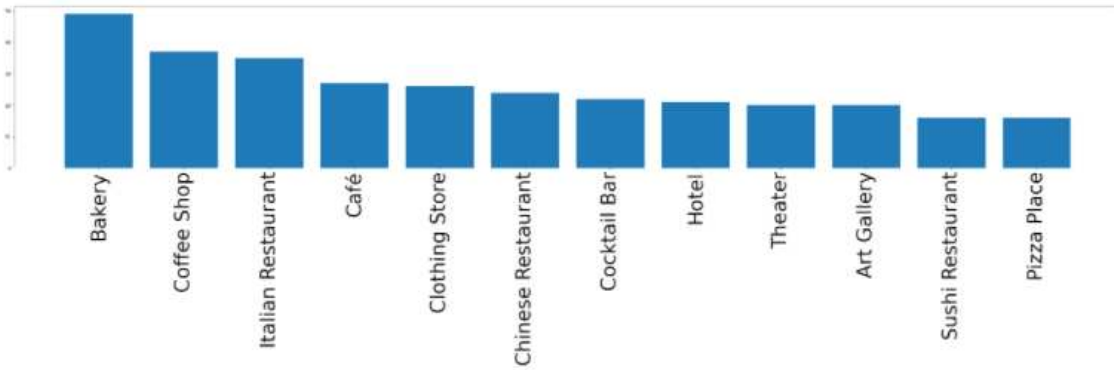
It was divided into five clusters. I used the K_Means.



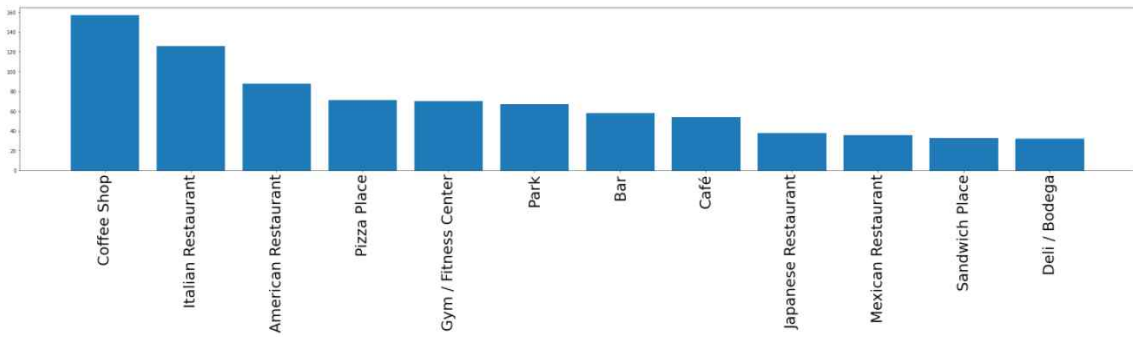
This is what cluster 0 looks like.It is a cluster with many Restaurants, Hotels and Gyms. It can be predicted that it is around a tourist destination.



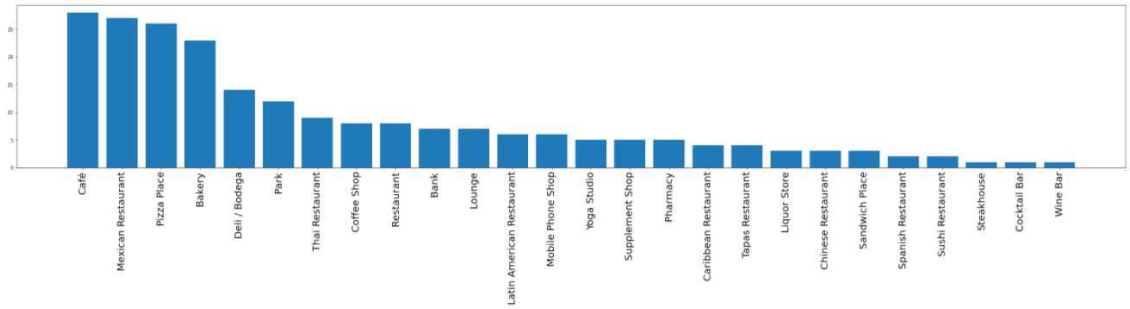
This is what cluster 1 looks like. Bakery, Cafe, Restaurant, etc. I think it's a cluster that couples often visit for dates. Food alley다.



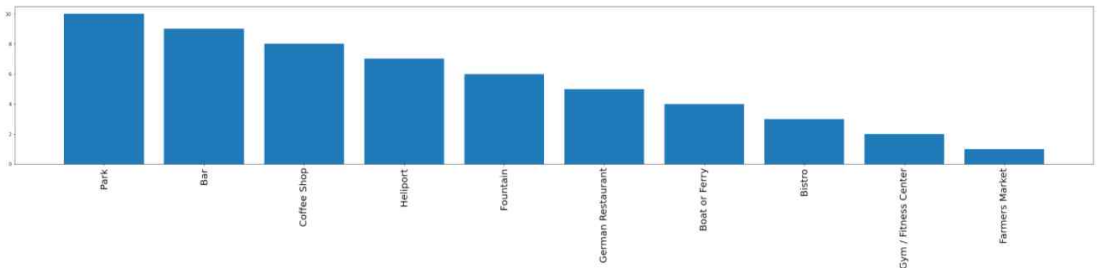
This is what cluster2 looks like. There were so many items that only 12 were selected. There are a lot of Coffee Shop, Restaurant, etc. The difference with cluster1 is that there is no Bakery. Eating in cluster2 and eating in cluster1 can also be predicted.



This is cluster 3. Unlike other regions, it is a cafe-intensive area.

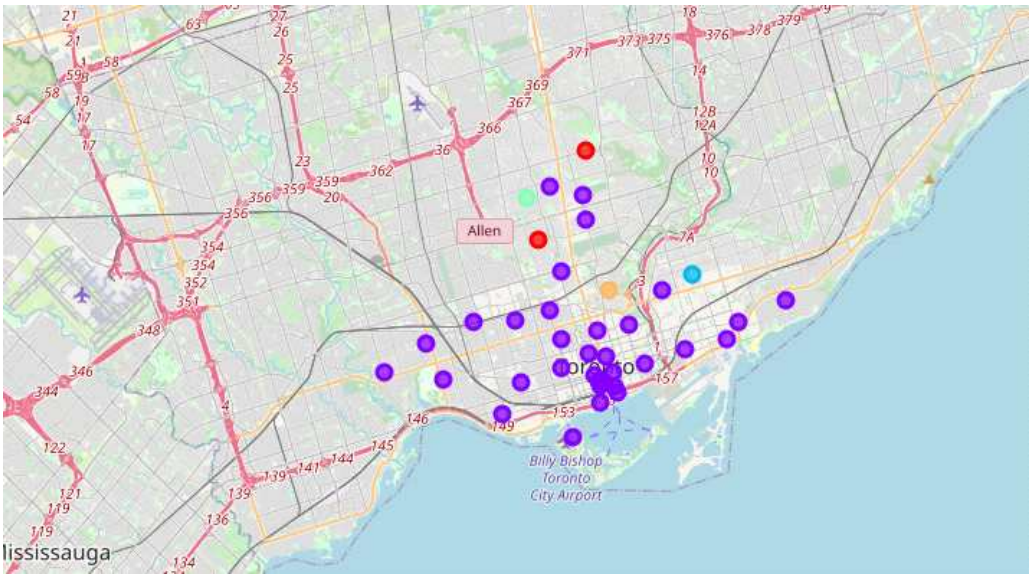


This is cluster 4. There are lots of park and bars. This area seems to be for a nightlife.

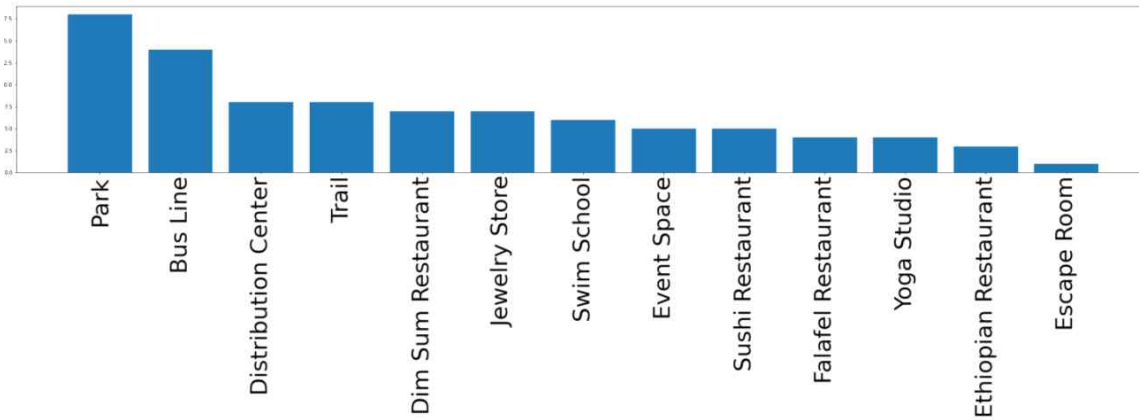


4.2 Toronto Clustering

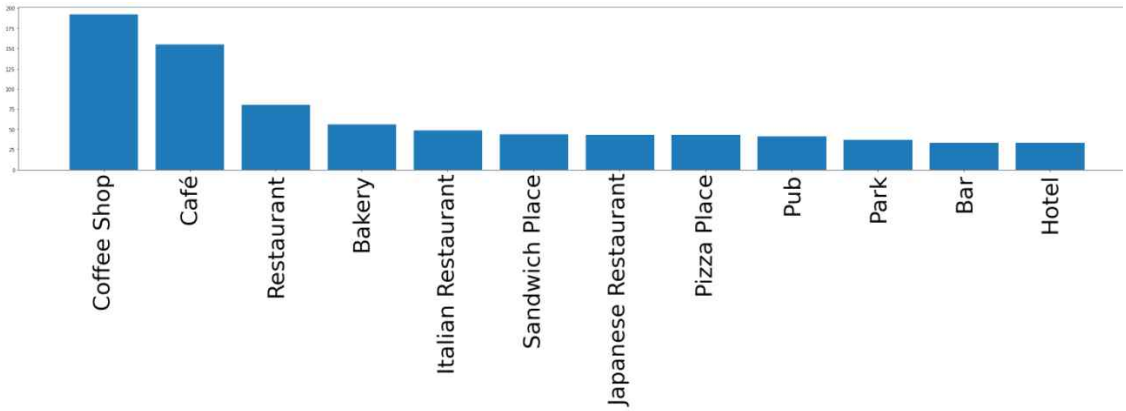
Datas were too focused on cluster 1 when we proceeded with K-Means. Even with DBSCAN, there were 36 cases per cluster. Toronto's clustering seems to need more detailed data work.



This is cluster 0. This consists of 2 boroughs. There are parks, bus lines and Trails. This cluster is for transportation and relax



This is cluster 1. This consists of most of boroughs. There are lots of restaurants and bakery. This cluster is food alley. There are many restaurants in Toronto!



The rest of the cluster contains only one bough, so it is omitted. It is not a meaningful cluster.

4.3 From Toronto to NewYork

I put Toronto's data into New York's model. Therefore, people living in each cluster are suitable for the corresponding cluster in New York City. For example, East York/East Toronto was NY model cluster2. Moving to the area will be similar to the original town.

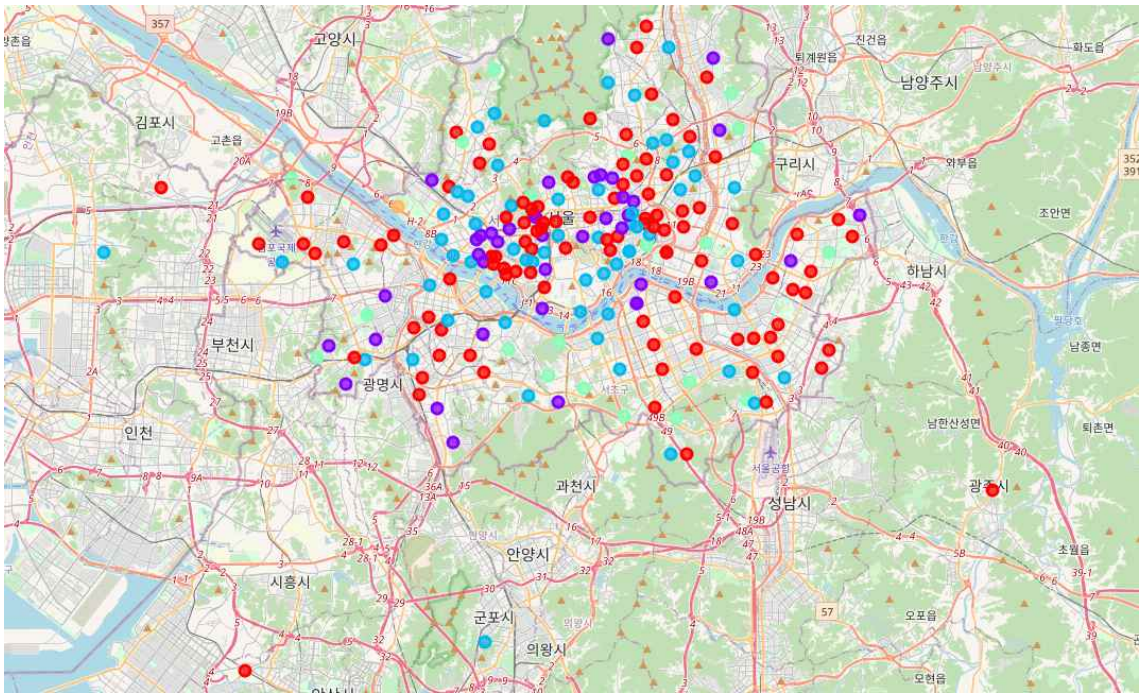
4.4 From NewYork to Toronto

All New York boroughs were in Toronto model cluster 1. It seems that a deeper analysis is needed.

4.5 From Seoul to NewYork

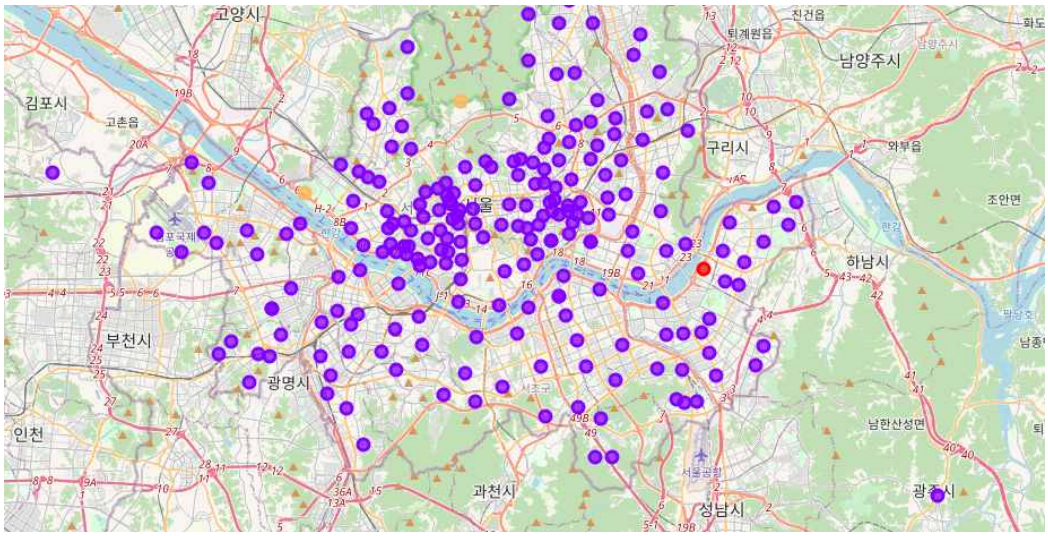
I put Seoul's data into New York's model. Therefore, people living in each cluster are suitable for the corresponding cluster in New York City. For example, Naengcheon-dong is NY model cluster1. Moving to the area will be similar to the original

town.



4.6 From Seoul to Toronto

I put Seoul's data to Toronto's model. Therefore, people living in each cluster are suitable for the corresponding cluster in Toronto. Almost all were in cluster1. But there were also exceptions. For example, Pyeongchang-dong is Toronto model cluster4. Moving to the area will be similar to the original town.



5. Conclusion

My findings have enabled people to move more appropriately. People will be able to adapt faster and live happily in similar circumstances. These models and methodologies could be applied to any city. My code has been released. If you want to move to other cities or travel to , analyze yourself the same way. It would be a good way to live in the most familiar environment.

6. Future Discussion

Toronto modeling is not perfect. Because too many areas belonged to one cluster. This needs to be supplemented. For now, there can be many causes of the problem. 1. The data received from FourSquare is too formal. 2. Indeed, their areas are all alike. 3. No suitable modeling methods have been found. It seems that follow-up research is needed to solve these things.

7. Refferences

G. References:
[1] Wikipedia
[2] Forsquare API