Segmenting and Clustering Neighborhoods in Tokyo

March 4, 2019

0.0.1 INTRODUCTION

Is Kichijoji the Only Place to Live? (Japanese: ?; Romaji: Kichijoji dake ga Sumitai Machi Desu ka?), a Japanese TV series adapted from a manga by Hirochi Maki, tells how the Shigeta twins from an apartment rental agency in Kichijoji - the most sought-after neighborhood in Tokyo - show their customers some great under-the-radar neighborhoods other than Kichijoji in Tokyo.

The shigeta twins can always meet the needs of the clients because the twins are extremely familiar with the neighborhoods in Tokyo and they stand in the customers' shoes. What if an apartment rental agency without the Shigeta twins wants to make recommendations as good as they did? A machine learning approach is to segement and cluster the neighborhoods in Tokyo based on their features, so that it can help apartment rental agencies - our target audience - to make smart recommendations at high proficiency and at low cost.

This project applies K-means clustering on the neighborhoods in Tokyo and the results will help the apartment rental agencies in Tokyo identify the neighborhoods that match the needs of customers in an efficient and economical way.

0.0.2 DATA

Tokyo neighborhood data and Foursquare location data will be used together to explore and cluster the neighborhoods in Tokyo.

Tokyo neighborhood data can be accessed from this website: *http://japanzipcodes.blogspot.com/2013/07/the-complete-zip-codes-of-tokyo-japan.html*. This web page lists out zip codes of all neighborhoods in Tokyo. This web page will be scraped and the data will be wrangled, cleaned, and read into a pandas dataframe as my first step. As shown below, our cleaned dataset has 1456 rows, which means we have 1456 neighborhoods in Tokyo. There are three columns: neighborhood, borough, and zipcode. The table below shows the first five rows of our dataframe.

In [13]:

```
(1456, 3)
```

Out[13]:

	neighborhood	borough	zipcode
0	Adachi	Adachiku	120-0015
1	Aoi(1-3Chome)	Adachiku	120-0012
2	Aoi(4-6Chome)	Adachiku	121-0012
3	Ayase	Adachiku	120-0005
4	Chuohoncho(1-2Chome)	Adachiku	120-0011

As you can see from the above cleaned dataset, there are two columns: neighborhood and borough. Tokyo is often referred to as a city but it is officially known as "Tokyo-to" - Tokyo Metropolis or the Greater Tokyo Area. It contains 23 special wards, 26 cities, 5 towns, and 8 villages, each of which has a local government. The Tokyo Metropolitan Government administers the whole metropolis including the special wards, cities, towns, and villages. In Japan, a ward/city/town/village as an administrative unit of a metropolis is closely equivalent to a London borough or a New York borough. Therefore in this cleaned dataset, the second column which contains the names of wards/cities/towns/villages in Tokyo was named as "borough" so that it is easier to understand.

Now we have built the dataframe combining postcodes, neighborhoods, and boroughs, I will obtain the latitude and longitude coordinates using the *pgeocode* package for each neighborhood in order to utilize the Foursquare location data. Here are the first five rows of our data table which contains neighborhood, borough, zipcode, latitude, and longitude.

In [22]:

```
(1456, 5)
```

Out[22]:

	neighborhood	borough	zipcode	latitude	longitude
0	Adachi	Adachiku	120-0015	35.7632	139.8076
1	Aoi(1-3Chome)	Adachiku	120-0012	35.7651	139.8129
2	Aoi(4-6Chome)	Adachiku	121-0012	35.7874	139.8195
3	Ayase	Adachiku	120-0005	35.7691	139.8264
4	Chuohoncho(1-2Chome)	Adachiku	120-0011	35.7651	139.8129

Now we have a cleaned dataset containing the geography information of Tokyo neighborhoods that we need in this project. The next step is to access venue information of those neighborhoods by using Foursquare location data.

I already obtained Foursquare credentials by setting up an account on Foursquare Developer API. Having both the coordinates of the neighborhoods and the Foursquare credentials enables me to access Foursquare location data. Foursquare location data offers comprehensive and accurate information about venues of given locations, for examples, restaurants, entertainment, hotels, stores, and others. The massive dataset of location data built by Foursquare also powers third-party apps, including Evernote, Uber, Flickr and Jawbone.

I will explore the neighborhoods and conduct cluster analysis by leveraging the Foursquare location data in combination with Tokyo neighborhood data. Visualizations and recommendations based on results of the cluster analysis will be made to help the apartment rental agencies in Tokyo identify the neighborhoods that match the needs of customers.

How I accessed and utilized Foursquare location data and how I conducted cluster analysis on Tokyo neighborhoods will be further explained in details in the methodology part.

0.0.3 METHODOLOGY

I created a map of Tokyo with neighborhoods superimposed on top to get started. But since that is an interactive map, it can not be shown in this PDF report. The codes to create the map can be seen from the notebook of this project. The map basically tells us the geography of Tokyo and the area that we are going to segment. It is also a helpful practice since we are going to mark different

clusters on the map of Tokyo. The visualization is an important part of the results too, although it can not be directly seen from this report.

Then I explored the first neighoborhood "Adachi" as both an example and a practice for exploring all the neighborhoods later. To do that, I first defined my Foursquare credentials so that I could request Foursquare to return me with venue information of a given location. I got the top 100 venues that are in the neighborhood Adachi within a radius of 500 meters. Then I borrowed the "get_category_type" function from the Foursquare lab to extract the categories of the venues. Afterwards, I cleaned the information and structured it into a pandas dataframe. Below shows the first five shows of the dataframe, which has the venue name, venue category, latitude and longitude. There are 20 venues returned by Foursquare for the first neighborhood Adachi.

In [34]:

Out[34]:

	name	categories	lat	lng
0	Noodle House		35.765638	139.808144
1	Supermarket		35.767233	139.809015
2	7-Eleven ( )	Convenience Store	35.764449	139.807945
3		Ice Cream Shop	35.766368	139.808839
4	Gotanno Station (TS11) ()	Train Station	35.766107	139.809383

After this practice, I started to explore all the neighborhoods in Tokyo. First I created a function to repeat the same process which I explained in the previous paragraph for all the neighborhoods in Tokyo. Then I ran that function on each neighborhood and created a new dataframe which has 50004 rows, which means there are 50004 venues in Tokyo. The table has 7 columns including neighborhood, latitude, longitude, venue name, venue latitude, venue longitude, and venue category. The table below shows the first five rows of the dataframe.

In [38]:

```
(50004, 7)
```

Out[38]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adachi	35.7632	139.8076	Noodle House	35.765638	139.808144	Noodle House
1	Adachi	35.7632	139.8076	Supermarket	35.767233	139.809015	Supermarket
2	Adachi	35.7632	139.8076	7-Eleven ( )	35.764449	139.807945	
3	Adachi	35.7632	139.8076		35.766368	139.808839	
4	Adachi	35.7632	139.8076	Gotanno Station (TS11) ()	35.766107	139.809383	

```
2  Convenience Store
3      Ice Cream Shop
4       Train Station
```

After checking, I found there are 440 unique venue categories. Then I conducted one hot encoding to add dummies of venue category into the table. The 440 unique categories all became the columns and each record will fall into one of the 440 columns. Then I grouped the rows by neighborhood and took the mean of the frequency of occurrence of each category so that we could know in each neighborhood, which are the most popular venue types. I also printed out each neighborhood along with their top five most common venue types. Due to the length limit here, you could refer to my notebook of this capstone project for detailed information.

Then I wrote a function to sort the venues in descending order and created a new dataframe that displayed the top 10 venue categories for each neighborhood. The first five rows of the new dataframe is shown below.

In [49]:

Out[49]:

	Neighborhood	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	10th Most Common Venue
0	Aburadai	Convenience Store	Intersection	Park	Indian Restaurant	Concert Hall	Food & Drink Shop	Udon Restaurant	Zoo Exhibit	Farm	Fast Food Restaurant
1	Adachi	Convenience Store	Donburi Restaurant	Restaurant	Bakery	Ice Cream Shop	Noodle House	Fast Food Restaurant	Dumpling Restaurant	Discount Store	Dessert Shop
2	Agebacho	Italian Restaurant	Japanese Restaurant	Sake Bar	French Restaurant	Soba Restaurant	Bar	Yakitori Restaurant	Ramen Restaurant	Coffee Shop	Kaiseki Restaurant
3	Aiharamachi	Convenience Store	Pharmacy	Donburi Restaurant	Intersection	Video Store	Food & Drink Shop	Bus Stop	Park	Auto Garage	Steakhouse
4	Aioicho	Convenience Store	Intersection	Chinese Restaurant	Bus Stop	Bus Station	Auto Garage	Hobby Shop	Liquor Store	Grocery Store	Golf Driving Range

This is the dataframe ready for cluster analysis. I ran K-means clustering analysis to segment the neighborhoods into 10 clusters by using the KMeans function in Python. That function returned me a label ranging from 0 to 9 for each neighborhood. I merged that results with our previous dataset to get latitude and longitude for each neighborhood. The first five rows of the

dataframe is shown below. In the column named "Cluster Labels" you can find each neighborhood has been assigned with a cluster type.

In [52]:

Out[52]:

	Neighborhood	Borough	Zipcode	Latitude	Longitude	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	10th Most Common Venue
0	Adachi	Adachiku	120-0015	35.7632	139.8076	5	Convenience Store	Donburi Restaurant	Restaurant	Bakery	Ice Cream Shop	Noodle House	Fast Food Restaurant	Dumpling Restaurant	Discount Store	Dessert Shop
1	Aoi(1-3Chome)	Adachiku	120-0012	35.7651	139.8129	5	Donburi Restaurant	Convenience Store	Noodle House	Discount Store	Park	Pharmacy	Café	Chinese Restaurant	Train Station	Bakery
2	Aoi(4-6Chome)	Adachiku	121-0012	35.7874	139.8195	0	Convenience Store	Bus Stop	Japanese Restaurant	Intersection	Furniture / Home Store	Café	Fast Food Restaurant	Motorcycle Shop	Bakery	Supermarket
3	Ayase	Adachiku	120-0005	35.7691	139.8264	0	Convenience Store	Dessert Shop	Sushi Restaurant	Okonomiyaki Restaurant	Motel	Park	Gym	Baseball Field	BBQ Joint	Video Store
4	Chuohoncho(1-2Chome)	Adachiku	120-0011	35.7651	139.8129	5	Donburi Restaurant	Convenience Store	Noodle House	Discount Store	Park	Pharmacy	Café	Chinese Restaurant	Train Station	Bakery

So far we have conducted clustering analysis on Tokyo neighborhoods and segemented them into 10 clusters based on their features. Neighborhoods with close features are grouped together. The results can help rental agencies identify and recommend suitable neighborhoods based on clients' needs. The results will be displayed in Results section.

0.0.4 RESULTS AND DISCUSSION

I visualized the resulting clusters by marking the ten clusters on Tokyo map in different colors. However, it can not be shown in this PDF report. So please refer to my notebook for the codes to

creating the interactive map on which the markers of 10 clusters were added.

Then I examined the 10 clusters of neighborhoods one by one and make recommendations based on the features of each cluster. And here is the summary of my findings/conclusions. For detailed information, please refer to the tables of each cluster in my notebook which tells you the exact neighborhood names and borough names in each cluster and their 10 most common venue types.

The neighborhoods in the first cluster are a good choice for people who especially like bakery and Donburi restaurants. People who like to live near convenience stores will also like this cluster. People who like music and go to concert halls a lot will also find those neighborhoods attractive. This cluster is also perfect for park and zoo lovers. Examples of the neighborhoods in this cluster are Ayase in Adachiku, Iko in Adachiku, Ohara in Setagayaku. Overall, the neighborhoods in the first cluster is very good for people who like to eat outside and enjoy city life.

The neighborhoods in the second cluster will be loved by people who enjoy Japanese food, including Ramen, Sushi, Sake, Okonomiyaki, and others. People who like to live near convenience stores will also like this cluster. Also this cluster is perfect for young people since there are a lot of bars and cafes around. The abundant choice of restaurants and bars in this cluster will be sought after by people who like to spend money on food and wine.

The neighborhoods in the third cluster are suitable for people who like cooking by themselves since there are many supermarkets and grocery stores nearby. The great number of convenience stores will loved by both single people and families. The museums, parks, and zoos in this neighborhood will be loved by families with kids.

The neighborhoods in the fourth cluster are perfect for people who love beach, zoo, farmers market and fish market. People who enjoy city life may not like the neighborhoods in this cluster since there are not many choices of restaurants, bars, cafes, museums, theaters and others. However, if someone enjoys beach and seafood, we should definitely recommend this cluster.

The fifth cluster is a good fit for these groups of people: 1) people who like and use electronics frequently, 2) coffee lovers since there are many cafes and coffee stores, 3) people who enjoy different kinds of cuisines such as Italian and French, 4) people who love parks and hot springs.

The following groups of people can be recommended with the neighborhoods in the sixth cluster: 1) People who take bus a lot, 2) people who like Japanese, Korean and Chinese food, 3) people who like parks, museums, or art gallaries. Examples of neighborhoods in this cluster are Adachi in Adachiku, Kaga in Adachiku, and Sekido in Tama, and many others.

Cluster 7 is for people who want to live near Pharmacy and Chinese restaurants. Families with kids will like this cluster since it has kids stores and zoos. The neighborhoods in Hachioji are good representations of this cluster.

Sports lovers will like the neighborhoods in the eighth cluster since there are many parks, playgrounds gold courts, campgrounds, and gyms in those neighborhoods. The abundant chocie of convenience stores, public transportation stations, supermarkets are perfect for people who don't have cars and want to live close to everything.

People who have kids can be recommended with neighborhoods in the ninth cluster. There are many kids stores, playgrounds, supermarkets in those neighborhoods. Ina, Uenodai, Yamada, and Yokosawa in Akiruno are good represents of cluster 9.

Neighborhoods in the last cluster are perfect for golf players, and people who enjoy nature and suburban life. People like Japanese food like Ramen and Sake will also like this cluster. Examples of the neighborhoods in this cluster include Motoki in Adachiku, neighborhoods in Hino, Egota in Nakanoku and many others.

0.0.5 CONCLUSION

This projects segments and clusters the neighborhoods in Tokyo into 10 clusters based on their features utilizing Tokyo neighborhood data and Foursquare location data which provides us with the venue information of the neighborhoods in Tokyo. The results can help apartment rental agencies make recommendations of neighborhoods that match the needs of customers.