Finding the Best Place to Open a New Juku Campus

Hikaru Hashimoto

1 Introduction

Juku is a kind of private fee-paying school in Japan for children to learn subjects such as maths or English after their regular school hours. Entrance exams for private junior high schools (7th grade to 9th grade), high schools (10th grade to 12th grade) and universities are tough, especially for higher rank universities in Japan. So, students study hard to pass these exams in jukus offering supplementary lessons to students.

I work for a juku where instructors teach maths and English to students from 5th grade to 12th grade. We have five campuses in Tokyo now, and we look for places to build new campuses because of the growing number of students. New campuses should be in similar areas to where we work as our business is going well now. In this context, we search for the best place to open a new campus in Tokyo by leveraging location data obtained on the web in this project. This will be informative for business owners in other fields as well because this approach is easily applicable to any field.

2 Data

2.1 Dataset

We use venues information of Tokyo to find similar places to our campuses. To obtain the information, we firstly obtain the geographical coordinate of postal codes of Tokyo fetched from the Japan Post by Geopy and put it in a data frame (figure 1). We use the coordinates to gather venues information by an API offered by Foursquare, a location platform service.

	zip_code	District	Area	lat	long
0	1020082	CHIYODA KU	ICHIBANCHO	35.729056	139.378416
1	1010032	CHIYODA KU	IWAMOTOCHO	35.695600	139.775379
2	1010047	CHIYODA KU	UCHIKANDA	35.691038	139.767290
3	1000011	CHIYODA KU	UCHISAIWAICHO	35.669426	139.755460
4	1010044	CHIYODA KU	KAJICHO	35.691689	139.771942
...	...	...	...	...	...
1028	1001102	MIYAKEJIMA MIYAKE MURA	IZU	31.999927	139.999227
1029	1001213	MIYAKEJIMA MIYAKE MURA	OYAMA	35.748624	139.702435
1030	1001622	HACHIJOJIMA HACHIJO MACHI	SUEYOSHI	35.658664	139.723526
1031	1001623	HACHIJOJIMA HACHIJO MACHI	NAKANOGO	33.065879	139.813648
1032	1001511	HACHIJOJIMA HACHIJO MACHI	MITSUNE	33.118764	139.802629

Figure 1: Screen Shot of the Geographical Coordinate Data.

2.2 Preprocessing

We have obtained 4005 postal codes of Tokyo. The geographical coordinate is obtained only from a part of them because Geopy could not find correct location data for the other postal codes. We have 1033 postal codes for this project in total, five of which are places our campuses are in.

3 Methodology

Firstly, we obtain venues information about all the areas in the data by Foursquare leveraging the geographical coordinates. Then, we measure the cosine similarity between Shibuya campus and the other locations. To reduce the size of the data, we pick the top 40 locations by the cosine similarity score. The cosine similarity may be good enough for decision making for choosing the next place to build a campus, but in this project, we use the k-means algorithm to reduce the candidates from another perspective. Technically, the cosine similarity computes the angles of two vectors, and it does not consider the length of the vectors. On the other hand, K-means considers the length of vectors. So, we suppose we can find better candidate places by using two of them than only one of them.

We use k = 10 for building our k-means model. The cluster including the location of Shibuya campus is the target. All the areas included in it are candidates for us to build a new campus because they are similar to Shibuya campus both in the cosine similarity and the k-means algorithm.

4 Result

Out of 40 candidate location we got by cosine similarity score, three areas fall in the same cluster as Shibuya campus; Sangenjaya, Higashi and Chiyoda (table 1). The cosine similarity ranges 0 to 1, and higher scores mean that the area is more similar to Shibuya. Sangenjaya has the highest cosine similarity score in the cluster except for Shibuya itself. (cos_score of Shibuya in the figure is 1 because we computed it with two identical vectors, Shibuya and Shibuya). Higashi has the second-highest cosine similarity score, but it is close to the Shibuya campus. Chiyoda has the lowest of the three, but it has an advantage in that it has a distance from our campuses (figure).

zip_code	Area	cos_score
1500002	SHIBUYA	1.000000
1540024	SANGENJAYA	0.801115
1500011	HIGASHI	0.773443
1000001	CHIYODA	0.743232

Table 1: A Candidate Area Cluster. Shibuya is the area where our Shibuya campus is placed. Cos_score is the cosine similarity score, ranging 0 to 1 (1 is the highest).

Figure 2: Candidate Areas and Shibuya Campus. Candidates are plotted in blue, and our campus in red. The left candidate is Sangenjaya, the middle one is Higashi and the right one is Chiyoda.

5 Discussion

5.1 Where to Build a New Campus

We propose to build a campus in Chiyoda, the east candidate in figure 2 because we have not covered the area. As we can see in figure 3, we already have campuses in the west area (top left in the figure) and the middle area. Especially, the left candidate area (Sangenjaya) is covered by two of our campuses (red circles at the bottom and in the middle of the map). The middle candidate area is also covered by two of our campuses (red circles in the middle and the northeast).

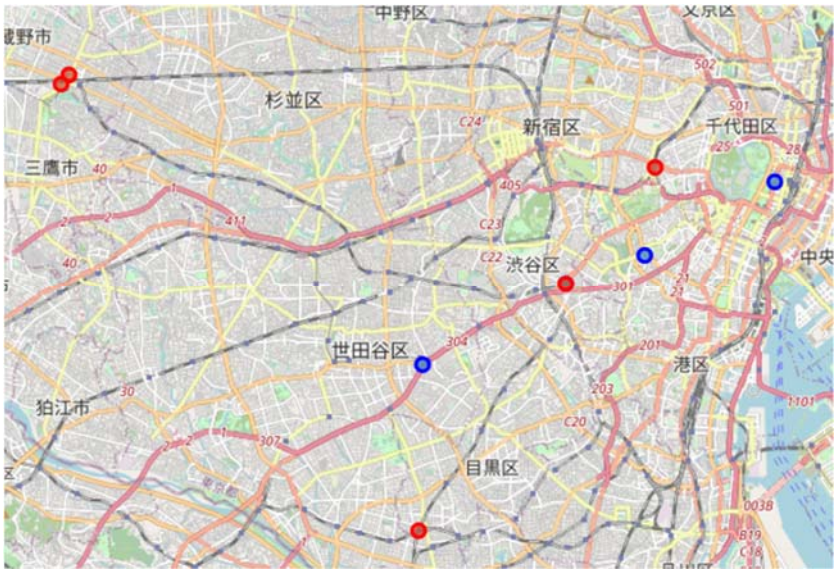

Figure 3: All of Our Campuses and Candidate locations.

5.2 Limitation

We have two limitations because of python libraries. Firstly, a python library Geopy does not find accurate geographical coordinates from zip codes. Two of them are plotted in the far west, and a zip code is plotted in the north. Secondly, venues data fetched by Foursquare may not be complete or may be old, and we might miss critical information. For example, we should highly consider schools because we teach English and maths for younger generations. It is important to build a new campus in an area where there are many teenagers. If Foursquare misses this information, it is better to think over the way we obtain our data. That said, the result offers us useful information for decision making, which is called data-informed decision making.

6 Conclusion

In this project, we observed three candidate places for a new campus by the cosine similarity and the k-means algorithm. Chiyoda (zip code 1000001) seems the best candidate because we have not covered the area. The data may not be complete enough for data-driven decision making, but it is enough for data-informed decision making.