A trip planning recommender

Man-wai Kwok

April 26, 2019

Table of Contents

1. Introduction

1.1 Problem

This report presents a trip planning recommender system for travelers. The system **suggests traveling areas** concentrated with attractive and functional venues **of travelers' interest**, aiming to help them plan their trips easier and quicker.

1.2 Background

Solo travel means a lot to different groups of people. Some spend their holidays for pleasure; some make wedding trips or graduation trips for their moments of life; some take rests at places where everyday troubles can go out of their minds so that they can think for what next. However, even you have very good ideas of which country or which city to go, before stepping out of the door, you still need to figure out which area in the city to stay as the beginning of the journey. This task is no easier even nowadays there is a massive amount of information, reviews and tips on the Internet, because it takes time to read them, to put them on a map and find out if some places are within walking distance, and so on.

Therefore, a trip planning recommender is going to be useful for them as it can show on the map walkable areas in which there are good attractions, good restaurants or interesting venues, so that a traveler can easily pick one suggested area to start exploring with. What's more, many solo travelers need to visit laundry shops, cafes, supermarkets and other places of daily life functions from time to time throughout their trips. While this kind of information is relatively uneasy to find, the recommender will take care of it. All in all, the trip planning recommender will be giving you area suggestions that suit your needs.

1.3 Targets

The system targets **any travelers** who need to plan for their trips, including but not limited to solo traveler, business travelers, exchange students and foreign workers. It eases their burdens to research on places that will fulfill every day of their traveling lives. These places include sightseeing spots, food places, and any functional venues such as supermarkets.

1.4 Summary

The system makes suggestions based on **user's inputs** of traveling place preference **in five key aspects**, namely Food, Hotel, Life, Rest, and Sightseeing. To demonstrate this work, Central Region, North Region and East Region of Singapore will be covered in recommendation making. The relevance of these regions to the key aspects will be **evidenced by** place **data** available on the Internet. The report is structured to first cover data collection, including data sourcing, data pre-processing and feature engineering, followed by methodology, namely how the engineered features are used in analysis and methods involved. Then the result and discussion of the system are presented, before making a conclusion to this work.

2. Data collection

2.1 Data source

Two sources of data were exploited in this work, which were Singapore government data at data.gov.sg. Another one was the Four Square API, which maintained a large places database and offered free limited access to its API [1].

The API allowed locations exploration. User might provide to the API a searching location (in terms of latitude and longitude) , a searching radius and a list of targeting venue categories [2], then it would return a list of targeted venues within the radius including venue's names, locations, and categories. The number of venues returned in each call was limited to 100, and a free account could only make at most 950 such calls per day. Additional information of a venue such as tips, photos and so on was regarded as premium calls and a harsher limitation applied. In this work, only basic venue information was used.

The Singapore government data website contained various types of datasets about the country such as economy, geology, tourism and so on. The region boundary dataset contained geo-polygons that defined the five administrative regions of the country in year 2014, of which Central region, North region and East region were used to confine an initial list of area suggestions, where areas would be further matched with user's preference to make the final suggestions. Tourism-related data used by the recommender included names and locations of hotels, attractions, historic trees, historic sites, museums, libraries and parks. Some of the datasets also included photo URLs and photo description which would make up for the lack of it from the API data.

2.2 Data cleaning and pre-processing

2.2.1 Initial area suggestions

An initial set of areas within the three interested regions were drawn randomly, in the condition that the areas did not overlap. The areas were all circular and were represented by a location as its center. A total of 259 areas were found, the coded name, location, and address of each area was recorded. The areas were marked on a map as in figure 1.

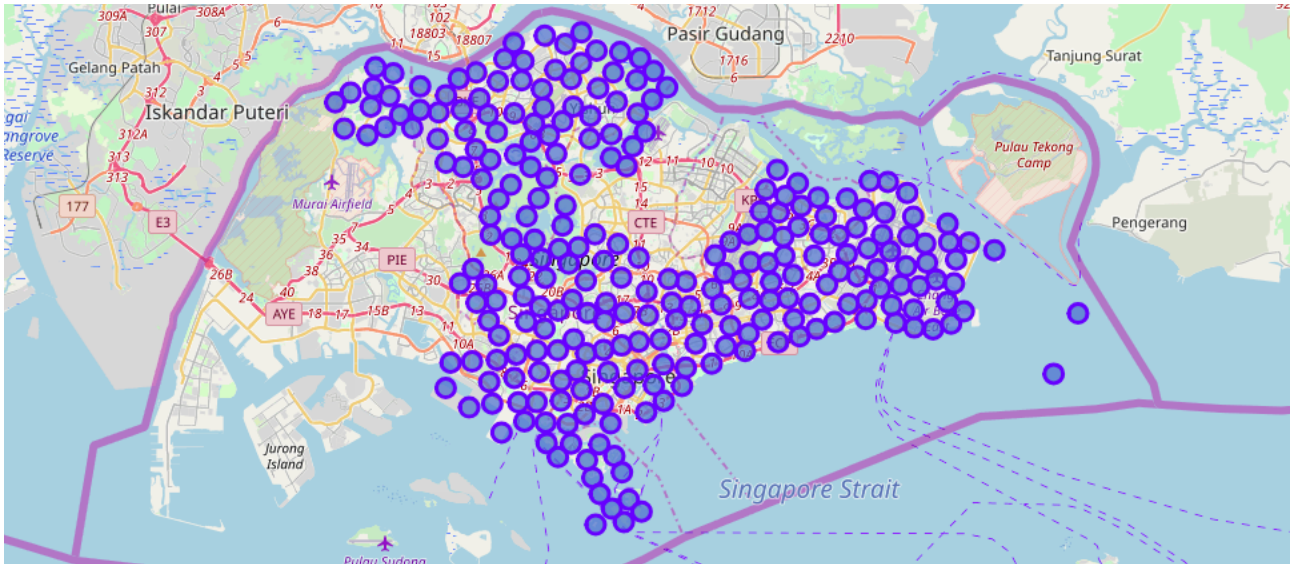


Figure 1. The areas (blue marks) were not overlapping and confined in Central region, North region, and East region of Singapore. Final areas recommendation would be drawn from them.

2.2.2 Venue data from API

To focus on potential venues interesting to a traveler, and considering limited API calls quota, the returning venue from the API was chosen to be *American Restaurant*, *Basketball Court*, *Church*, *Bank*, *Café*, *Gym*, *Asian Restaurant*, *Supermarket*, *Shopping Mall*, *Laundry Service*, *Market*, *Hospital*. More should be added to cater for different people's need in the future.

The API returned data in JSON format. The dataset was complete and did not require further processing after transforming into a table. They only contained the category of each venue, its name and location. Because there was no photo URL and description, empty columns were appended to the table. The distribution of venues was plotted as in figure 2.

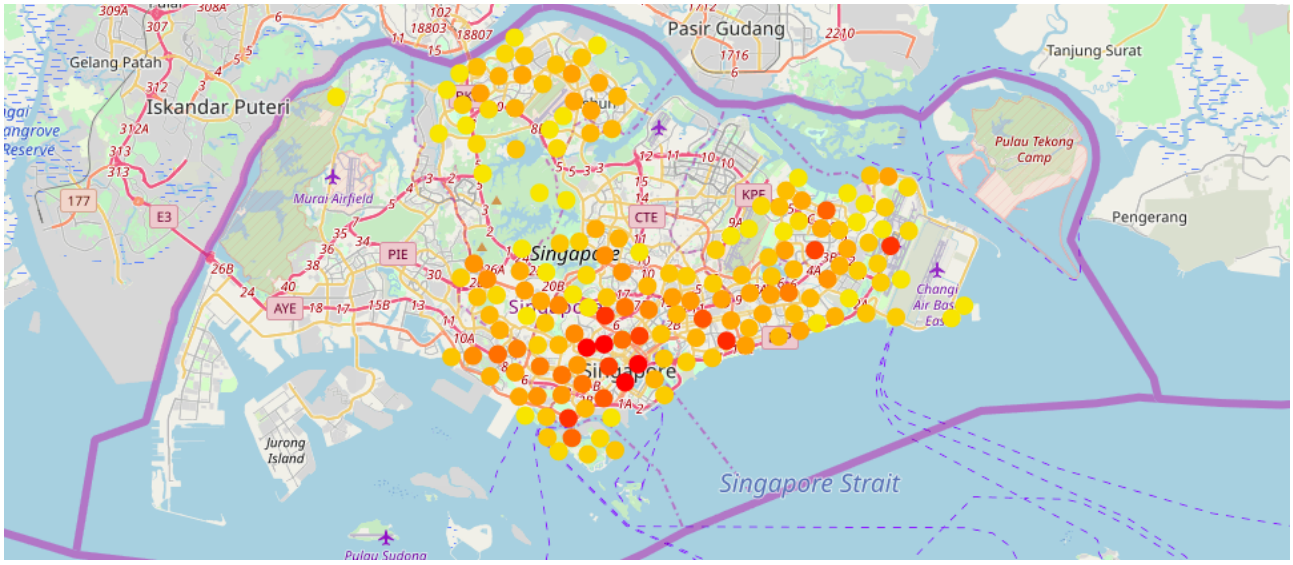


Figure 2. API's venue data. The color scale represented the number of venues in each area. Yellowish marker refered to area with less venues, and reddish more. The range of venues number was 1 to 140.

2.2.3 Venue data from downloaded dataset

Seven datasets of venue types were downloaded. While some come with popular SHP file format and some were not, however, all of them had a KML format alternative. Therefore their KML format was used to save coding effort. Different datasets contained different attributes for their venues, to standardize it, only venue's name, location, photo URL and description were used and got transformed into a table. For consistency, a venue category column was manually added. The category was the type of dataset itself, namely, *Hotel*, *Attraction*, *Historic tree*, *Historic site*, *Museum*, *Library and Park*. Their distribution was plotted as in figure 3.

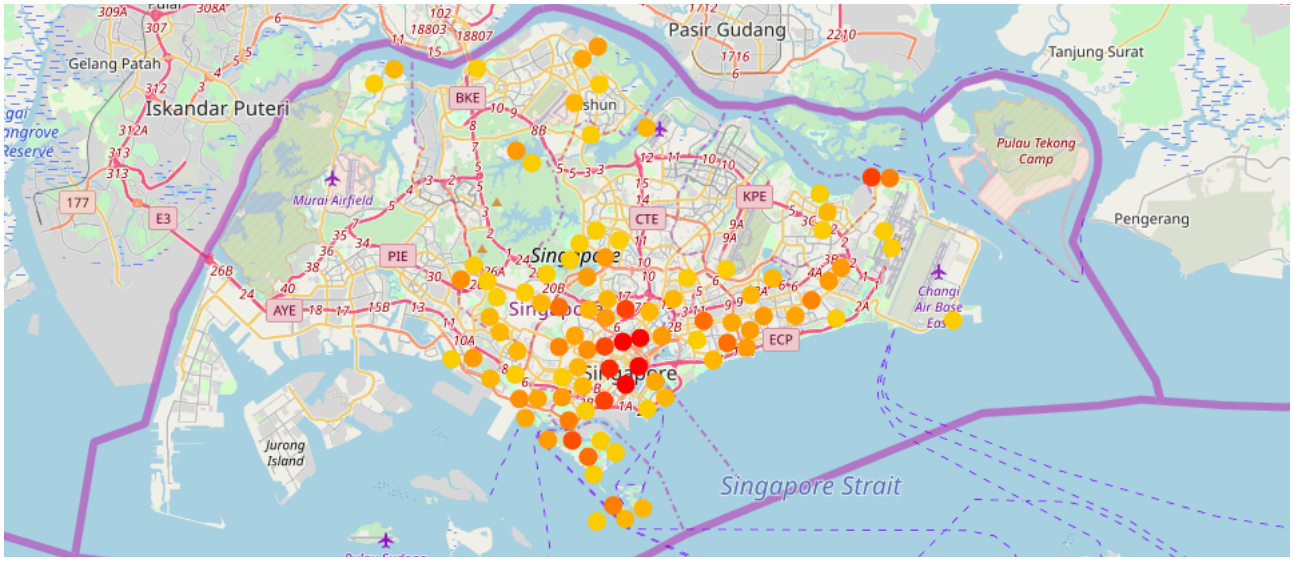


Figure 3. Downloaded venue data. The color scale represented the number of venues in each area. Yellowish marker refered to area with less venues, and reddish more. The range of venues number was 1 to 46.

2.2.4 Final venue dataset

Since data from both sources were designed to be tables having the same set of columns, they could be simply concatenated to form the final dataset. Together there were 19 different venues, of which the 7 downloaded venues might have a photo URL and description if provided. All venues had names, categories and locations in the form of latitude and longitude.

2.3 Feature Engineering

The suggestion was made on area based. First, each venue was paired up with an area if the venue was within the radius of 500 meter of the area. Unpaired venues were dropped. Then, the number of venues of each area was counted on category-based, and became the area's features. Therefore, the raw venue data would be transformed into a dataset of 259 areas (rows) by 19 categories (columns / features). This concluded the first stage of feature engineering.

The second stage happened after knowing user's preference of these venue categories. The preference of each venue was quantified to a number in the range of 1 to 9 inclusive. To simplify

user's input, they only needed to provide preference for 5 aspects, where each aspect represented a set of categories. The mapping between the aspects and the categories is shown in table 1.

Table 1. Mapping between user inputting aspects and categories.

Aspect	Category
Food	American Restaurant, Asian Restaurant
Hotel	Hotel
Life	Bank, Basketball Court, Church, Gym, Hospital, Laundry Service, Library, Market, Supermarket
Rest	Café, Park, Shopping Mall
Sightseeing	Attraction, Museum, Heritage Tree, Historic Site

If an user chose 7 for Food's preference, the same value applied to all corresponding categories. However, such mapping was arguable and **improvement** could be made by learning user behaviors from a large user dataset, which was not available during this work.

Then each feature was normalized individually. Before normalization, the feature represented the number of venues in a category and had an arbitrary range. A cutoff value was obtained based on the preference value of a feature. In that feature, any value over the cutoff will be set to the cutoff value, before dividing it by the cutoff to become in the range of 0 to 1. Then the value was scaled by the preference value, such that an important feature had a wider range.

3. Methods

The user's preference was added to the engineered features set before the set was learnt by K-Means. A quick overview of this method is discussed in below.

3.1 K-Means

K-means is an unsupervised machine learning algorithm that clusters close data instance together. Two parameters control the result of the method. The first one is the number of cluster, denoted by *k*, and the second is the initial locations of the *k* cluster's centroids, which is normally set randomly.

The algorithm proceeds by grouping data instances to the clusters based on their distance to the centroids, and an instance belongs to the closest centroid. After the grouping, the centroid locations are re-calculated by taking the means of locations of instances of each group. The process of is then repeated by regrouping the instances under the new sets of centroids, followed by deciding another new sets of centroid locations until it converges.

3.2 Strategy

K-Means was applied to the final features set. Areas in the same group as the user's preference were considered suggestions to the users. Since the grouping result depended on the value of *k* and initial cluster seeds, K-Means with multiple configurations of these settings were performed, and a voting mechanism was applied to all results to get the final suggestions.

For each result, an area earned a vote if it was a suggestion as defined in above. The areas were ranked based on the number of votes, and areas of higher ranks became the final suggestions.

4. Results

A suggestion result based on user's preferences on the five aspects is shown in figure 4. All suggesting areas were marked in green and the goodness of each was indicated by the marker's opacity. High ranking areas were represented as solid green color, and bad ones became less visible.

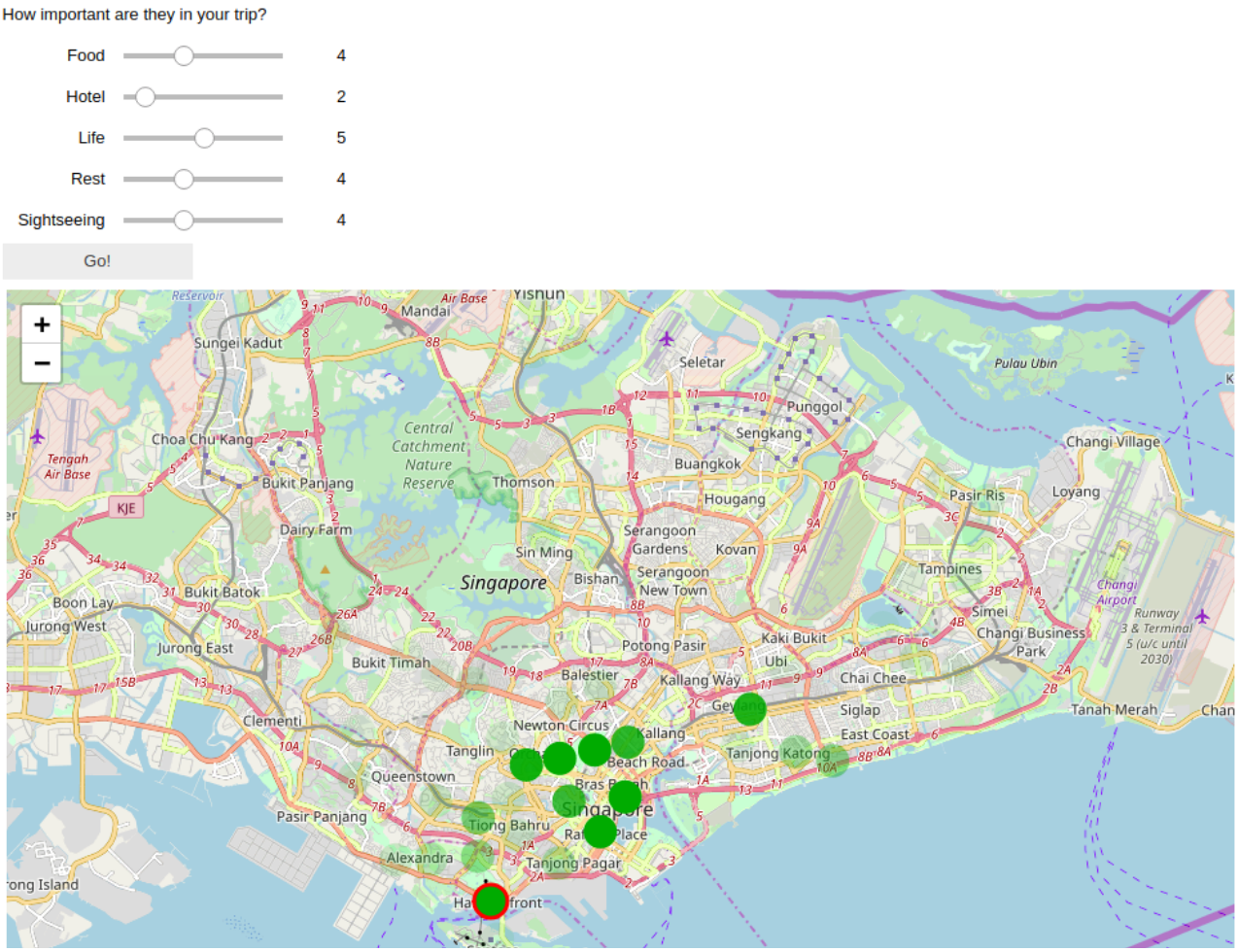


Figure 4. Final suggestion result based on user's preference on five aspects (upper). The suggesting areas were plotted in a map (lower), with opacity representing the likelihood of a good suggestion. A solid green marker represented a good suggestion. The best choice had a red-outer green marker.

For better result visualization, area information of each suggesting area such as venue name, address and venue counts were shown as in figure 5, photos in figure 6. Figure 7 plotted all venues in the area.

Best choice: **CENTRAL REGION-00-26**

Address: Tan Ean Kiam Building, 15, Phillip Street, Chinatown, Raffles Place, Singapore, Central, 049514, Singapore

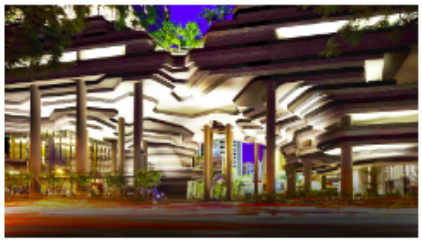
Summary:

venue_cat	Asian Restaurant	Attraction	Church	Gym	Heritage Tree	Historic Site	Hospital	Hotel	Market	Museum	Park	Supermarket
count	69	8	5	5	1	13	3	19	32	2	2	2

Figure 5. Best choice among all suggestions. A table listing the counts of venue categories was also shown to evaluate the goodness of the choice.

The following photos and descriptions are from data.gov.sg

Beautifully restored, Thian Hock Keng Temple is the oldest Chinese temple in Singapore and dedicated to Mazu, the Goddess of the Sea.



PARKROYAL on Pickering is a luxurious garden oasis in downtown Singapore.

The grande dame of markets in Singapore, Lau Pa Sat blends history, striking architecture and scrumptious local food into one heady experience.

Figure 6. Photos and descriptions of venues in a suggested area.

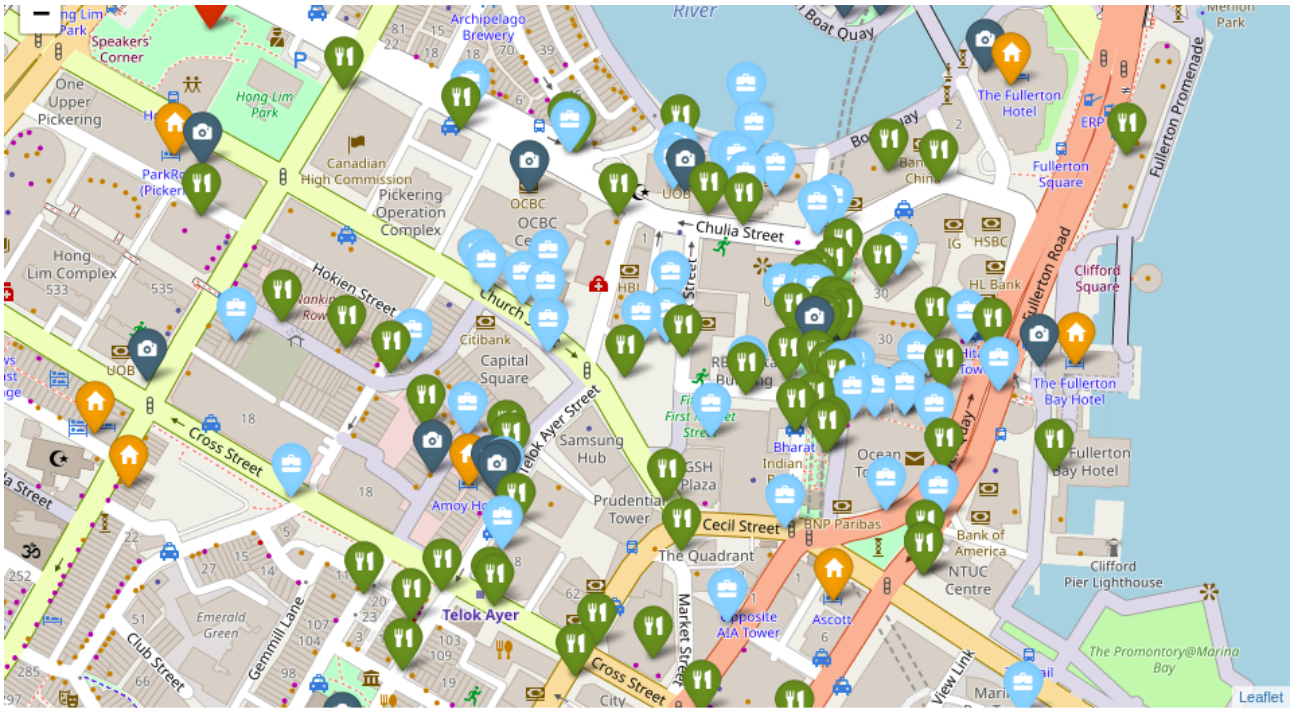


Figure 7. Venues in a suggested area were shown with markers in different signs and colors.

To understand the behavior of the system, an experiment was done as follows. A total of 30 preference points was distributed to the five aspects' preference values. The distribution rule was that, firstly, one aspect was chosen to be the experimenting object. The object's preference value changed from 1 to 9 points, while the other aspects shared equally the remaining points. For example, when Food was the experimenting object, and its preference value was 2, then all four others' values were 7. In each set of preference values, the best 6 suggestions were collected, and the total number of venues belonging to the experimenting aspect was plotted against the aspect's preference value. The result for all five aspects' plots are shown in figure 8. They all showed increasing trends as the aspect's preference went up, which is expected.

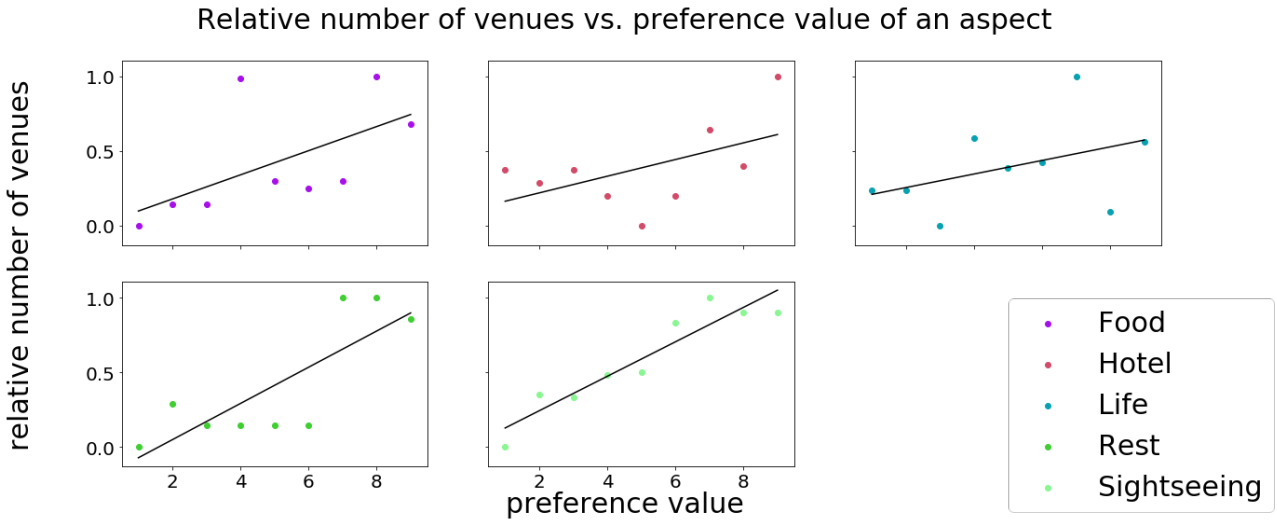


Figure 8. Relative number of venues vs. preference value of an aspect

5. Recommendations

The system can be improved in several aspects. Firstly, more data can be obtained to reflect a more complete reality, including the number of categories, and number of venues. Secondly, public reviews of venues can be incorporated to get not only the number of venues under a category in an area, but also the goodness of the category in the area.

Thirdly, the current system accepts user's preferences in five aspects and such arrangement is highly arguable. A better choice of inputs can be grounded on expert's opinions, and 'behavioral' input can be invented. Behavioral input in this context refers to user's behavior data such as browsing history that will be useful for predicting the user's preference. Such prediction model would require a large set of data and a content-based or a collaborative model could be built for that.

Lastly, the scalability of the current system is not good. Not only does it require to perform multiple K-Means with all available areas, as the number of areas and number of features scale up, it requires more time for one K-Means operation. While the best strategy for making a scalable system depends on the design of the whole system from data collection, user's input to expected outcome, one quick way to tackle the data size issue is to proxy similar areas, so as to reduce the original large dataset of many areas into a smaller dataset of fewer proxies. The number of features

can also be reduced in a similar way. Therefore keeping a small dataset is the first possible solution. Another possible solution is to embed features of an area into lower dimension numbers by use of Machine Learning models, which is another way of effectively reducing the learning dataset size.

6. Conclusions

A working trip recommender system was implemented and presented in this report. An experiment was carried out and that it could bias its recommendation to preferred venues. While enlarging the data size would reflect the reality better, a dedicated and easy-to-use user input method should be designed to handle the enlarged variability, and also, a better feature engineering approach would be required to make the system scalable.

7. Reference

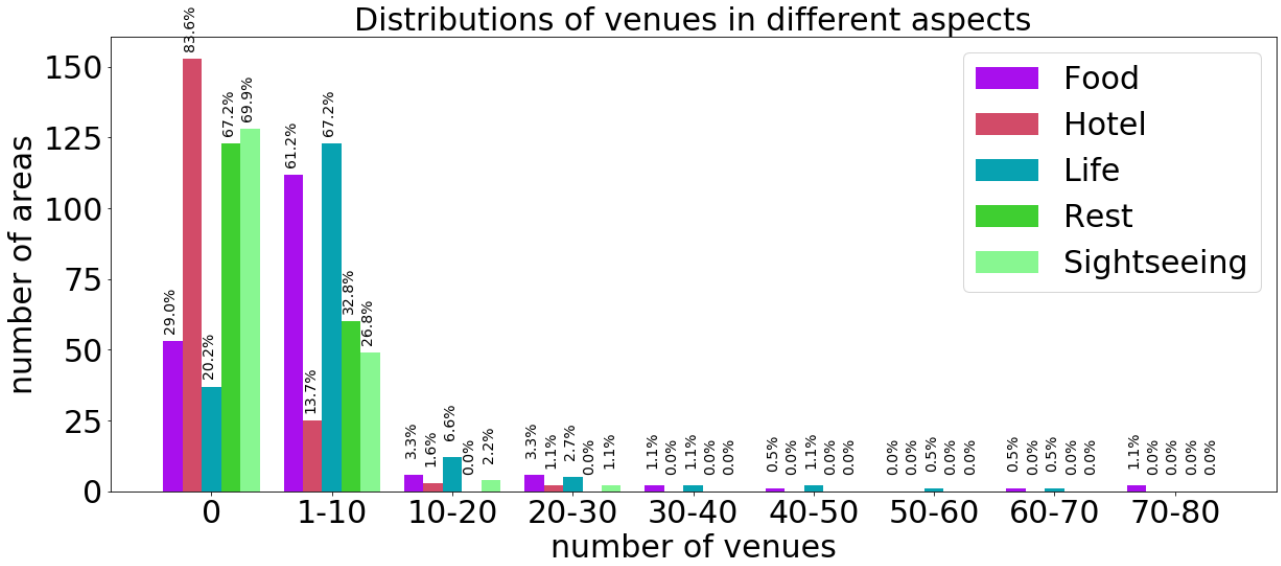
[1] Foursquare API: https://developer.foursquare.com/docs/api

[2] Foursquare API category definitions: https://developer.foursquare.com/docs/resources/categories

[3] K-Means clustering: https://en.wikipedia.org/wiki/K-means_clustering

8. Appendix

8.1 Distributions of venues in different aspects



8.2 Co-relation matrix of aspects

	Food	Hotel	Life	Rest	Sightseeing
Food	1.000000	0.667757	0.871838	0.089283	0.630810
Hotel	0.667757	1.000000	0.657781	0.226432	0.607198
Life	0.871838	0.657781	1.000000	0.146762	0.500955
Rest	0.089283	0.226432	0.146762	1.000000	0.128936
Sightseeing	0.630810	0.607198	0.500955	0.128936	1.000000