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.

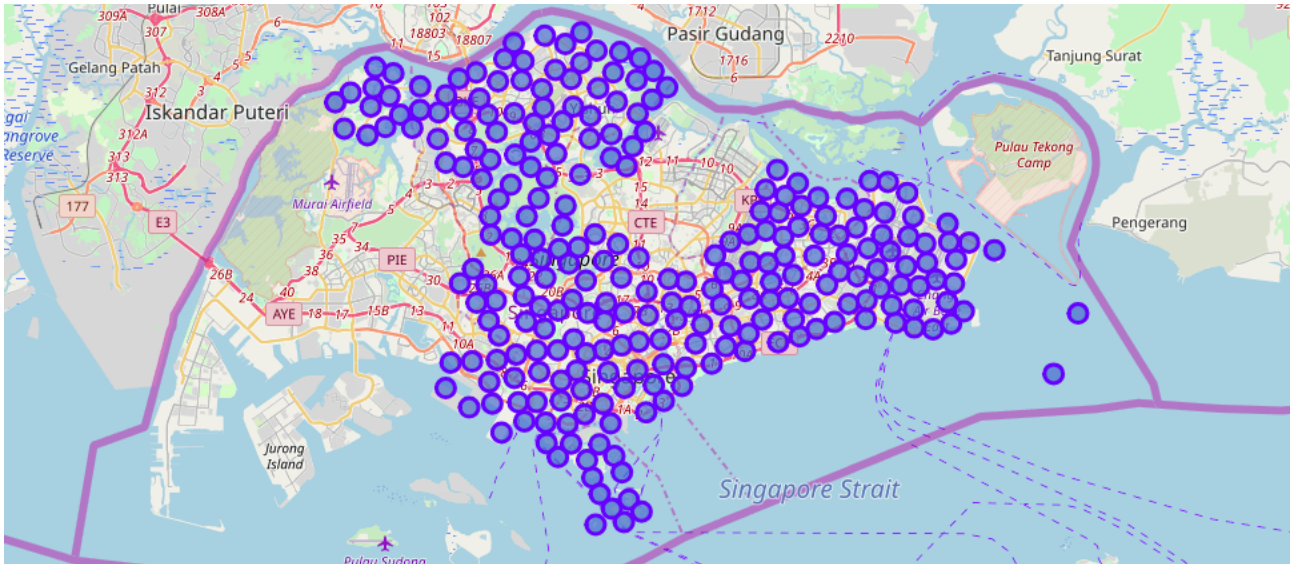

Figure 1. The areas (blue marks) were not overlapping and confined in Central region, North region, and East region of Singapore. Final areas recommendation would be drawn from them.

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.

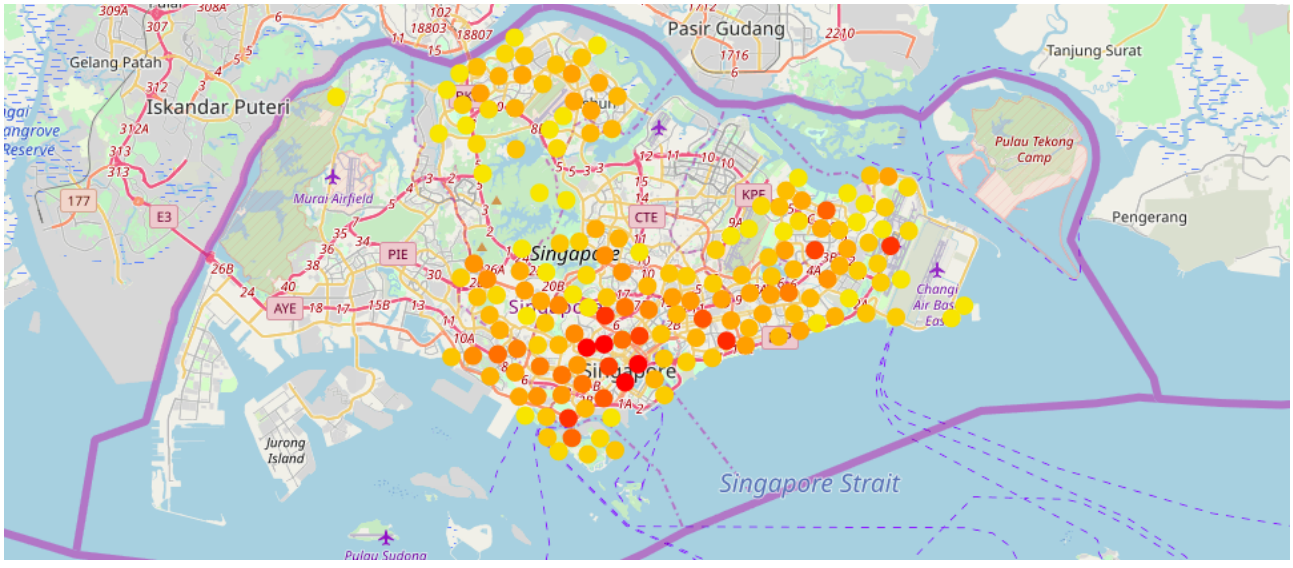

Figure 2. API's venue data. The color scale represented the number of venues in each area. Yellowish marker refered to area with less venues, and reddish more. The range of venues number was 1 to 140.

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.

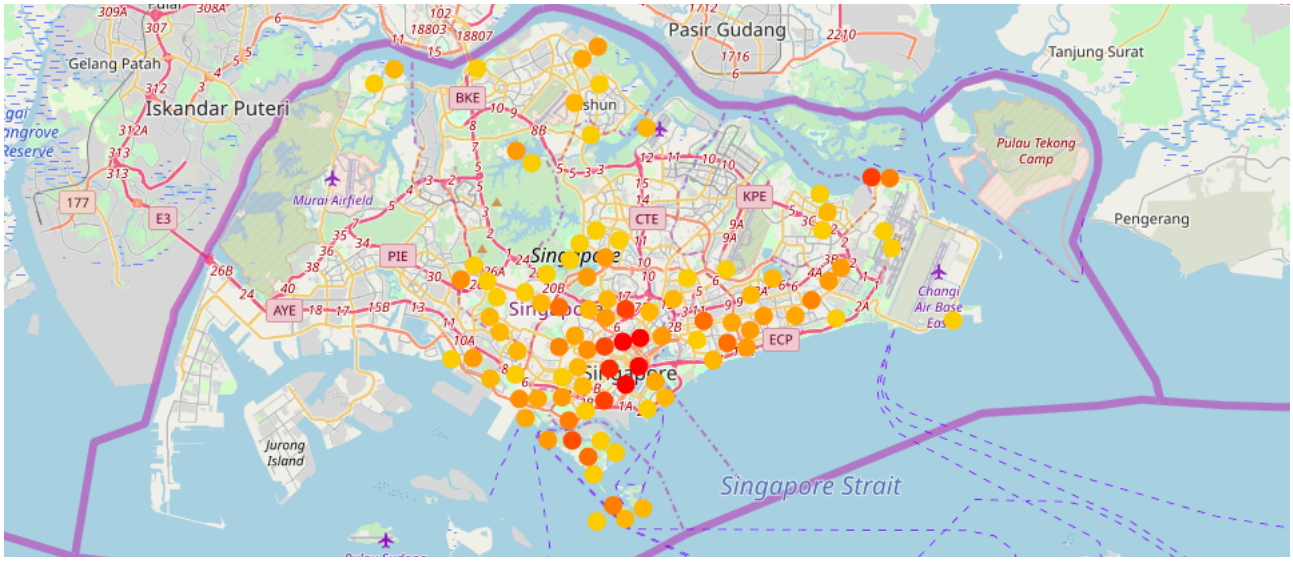

Figure 3. Downloaded venue data. The color scale represented the number of venues in each area. Yellowish marker refered to area with less venues, and reddish more. The range of venues number was 1 to 46.

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.