**Recommend Hotels in Manhattan, NY for Visitors**

Author: Meixuan Xu

Date: July 6, 2020

**1. Introduction**

#### 1.1 Background

New York City is one of the most popular tourist cities. The tourist population in New York has continuously increased in past nine years and welcomed approximately 65.2 million visitors in 2018, including overall 37.9 million visitors who stayed overnight in 2018. When people plan a trip, the first thing they need to do is to book fly and hotel. Many visitors want to book a hotel surrounded by preferred categories of restaurants within walking distance, so they can have nice dining out, and taste different styles and delicious food during their trip.

However, there is no booking website helps visitors to find hotel based on based on the restaurant flavors. For example, Booking.com is a famous website for booking accommodations. This website put together several menus for their customers to search different accommodations separately. If you searched hotels, you got a result of hotel list. If you want to know the surrounding accommodations, such as restaurants and public transport stations, you need to do a lot of extra searches on other search engines like maps.google.com.

**1.2 Problem description**

This project will leverage foursquare location data to build a recommendation system to help tourist to find hotel in NYC and explore the surroundings of the hotel based on their preferred restaurant flavors.

Suppose a family want to take their vacation in Manhattan, NYC. They want an affordable, comfort hotel which has a lot of preferred style restaurants in a walking distance, so they can easily have nice diners in nearby restaurants during their stay.

From a data scientist's view, this problem is described as a family want a hotel (a GPS point) in a neighborhood that meets below requirements:

1) In a hotel with at least several preferred categories of restaurants.

2) These restaurants are in walking distance.

## 2. Data Sources and Data Preprocessing

**2.1 Data sources**

To complete this project, we need two geography data sources, NYC neighborhood geography data and NYC venue geography data.

NYC neighborhood geography data can be downloaded from [geo.nyu.edu](https://geo.nyu.edu/catalog/nyu_2451_34572) And 5 boroughs and 306 neighborhoods geography data of New York City have been received from the downloaded file.

NYC venue geography data can be retrieved from a location technology platform - Foursquare.com which provides powerful venue and location information. We can leverage the free Foursquare API to get all of restaurants and hotel information in every neighborhood in NYC.

**2.2 Data Preprocessing**

When raw data was downloaded from a website, usually they were just exact what you want. We need do data wrangling to extract the fields that we needed, or maybe we need to create new fields for analysis.

Step 1, since the borough Manhattan is most popular borough in NYC, in this project, I practice a scenario that the visitor wants to find a hotel in borough Manhattan. the NYC neighborhoods geography data has 16 properties. I extracted 4 properties in each record, and then select the Borough Manhattan geography data for next step analysis.

Step 2, Based on the Manhattan neighborhoods geographic data, I used Foursquare API to get all venues for every neighborhood in Manhattan, and then extracted restaurant venues and hotel venues.

Step 3, This step is data standardization. For example, the value of the venue category in the data frame is not standardized. so, I copied the current category and add as a new regional style column and reorganize the venue categories before further analysis. For example, some restaurants resisted their category as "Dim Sum Restaurant", actually this belongs to Chinese style restaurant, so "Dim Sum Restaurant" is replaced with "Chinese restaurant" in the regional style column.

Step 4, In this project, supposed the visitor’s favorite style restaurant are Chinese restaurant, Seafood restaurant, and Japanese restaurant. Thus, we select there three categories of restaurants from the standardized restaurant venues as preferred restaurants.

After data preprocessing, we got preferred restaurant list (see Figure\_2\_1) and hotel list (see Figure\_2\_2) in borough Manhattan.

|  |
| --- |
|  |
| Figure\_2\_1 Preferred restaurant list in Manhattan |
|  |
| Figure\_2\_2 Preferred restaurant list in Manhattan |

## 3. Data Analysis

**3.1 Feature variables decided by visitors**

The goal of this project is to recommend hotels that meets the user's two requirements: number of preferred categories of restaurants near the hotels and they are in a walking distance.

A hotel with how many preferred restaurants nearby should be recommended? Different visitors might have various numbers. This number is apparently should be decided by visitor. Here I use a variable "least\_num" to let user to decide at least number of the preferred restaurants near a hotel to be recommended.

Also, how many meters away is in a walking distance? Different users might have their own feelings about walking distance. So, this project also allows users to decide their own distance as walking distance. A variable "distance\_m" was created for the walking distance in meter.

**3.2 Calculation**

Since I alreadyhave the GPS information of both restaurants and hotels. Here I need to download python module geopy.distance to calculate the distance between hotel and restaurant.

When I calculate the distance between hotel and all restaurants, all of preferred restaurants within the "distance\_m" have been saved and counted, and sorted.

**3.3 Result**

Based on the variable "least\_num" and variable "distance\_m", two scenarios are demonstrated as below.

Scenario 1, when least\_num > 5 and distance\_m <= 400 meters, we got 14 hotels recommended. See Figure\_3\_1

|  |
| --- |
|  |
| Figure\_3\_1 |

Scenario 2, when least\_num > 3 and distance\_m <= 400 meters, we got 21 hotels to recommend. Figure\_3\_2 shows the first 10 records

|  |
| --- |
|  |
| Figure\_3\_2 |

## 4. Results Visualization

Now we’ve found all of the recommended hotel list, let's cast the results on the map. The recommended hotels are green with hotel icon, and the not recommended hotels are gray with hotel icon. Preferred restaurants are orange with bold circle marker. Data visualization helps user better understand data than just data. Figure\_4\_1 shows the whole picture of the recommended hotels and preferred restaurants. Figure\_4\_2 shows the enlarged part.

|  |
| --- |
|  |
| Figure\_4\_1 shows the whole view of the recommended hotels |

|  |
| --- |
|  |
| Figure\_4\_2 shows the enlarged part of result |

## 5. Discussion

Although visitor usually only need one hotel during his/her visit, however it has a lot of recommendations in this project. There are 15 recommended hotels in scenario 1 and 24 recommended hotels in scenario 2. This is because besides dining, user might have other considering, such as price, transportation station nearby. So, this project just provides a recommended hotel list for user to pick the one meet use's other requirements.

In this project, I chose free Foursquare API to retrieve restaurant venues and hotel venues information. Comparing with the results from map.google.com, I got less hotels from Foursquare API then from map.google.com. Since this project is only for study use not for business purpose, it's acceptable that we did not get all of the hotel records in Manhattan.