

|  |
| --- |
|  |
| IBM Applied  Data Science Capstone Project Report on NYC Restaurants Search |
| December 23  Authored by:  Rohini Mandge |

Table of Contents

[Introduction 3](#_Toc533347607)

[Data Gathering 3](#_Toc533347608)

[Methodology 4](#_Toc533347609)

[Result 5](#_Toc533347610)

[Discussion 6](#_Toc533347611)

[Conclusion 7](#_Toc533347612)

[Reference 7](#_Toc533347613)

# Introduction

As modern consumers, we greatly benefit from restaurant recommendation applications. It is so convenient to get a list of restaurants that match our preferences without much clicking, comparing, and browsing through a long list of reviews for each single business.

In this project, I want to apply the algorithms to develop predictive models learned from the past courses in this IBM Data Science certification. A Recommendation System to build a restaurant recommendation system that suggests the most suitable/popular neighborhoods for given type of cuisine/category by users.

This model can be used by the end users who are interested in finding neighborhoods for their preferred type of cuisines.

# Data Gathering

The dataset that I am going to use for this project is available free on the web. Here is the link to the dataset: <https://geo.nyu.edu/catalog/nyu_2451_34572>. I downloaded the json file from this link. This dataset contains information about the 5 boroughs of New York city and its neighborhoods. It also contains the latitude and longitude coordinates of each neighborhood.

In this json file, all the relevant information is stored in the ‘feature key’. I am going to extract this required information and convert it into python pandas data frame format for easy analysis of the data. The features that I extracted form the json file are -

* NY city Borough
* Neighborhood
* Latitude and Longitude coordinates.

I used the Foursquare location data and run a query on it for the required type of cuisine/category by the user. The information which is needed is in the ‘item’ key and will convert this as well in pandas data frame.

Some of the features which I extracted from ‘item key’ of foursquare data are -

* Venue Name
* Venue Category
* Venue location Latitude and Longitude coordinate.

# Methodology

Following stages are performed for this project. Each stage plays a vital role in the context of the overall methodology. A methodology is a general strategy that guides the processes and activities within a given domain.

As mentioned in Data gathering part, the data is downloaded from the web from NYU spatial data repository. This data is in JSON form and contains the 5 boroughs and the neighborhoods that exist in each borough. It also contains the latitude and longitude coordinates of each neighborhood. This data is ideal for my analysis because I need to segment the neighborhoods and categories of the venues and explore them.

After loading the data from the json file, the data preprocessing step is performed. All the required data for the analysis was in the ‘feature key’ of the json file. The feature key is the nested python dictionary. I transformed that into the pandas data frame for easy manipulation and analysis.

I have used the Foursquare API to get the information about different venues. The Foursquare API provides location based experiences with diverse information about venues, users, photos, and check-ins. The API supports real time access to places, Snap-to-Place that assigns users to specific locations, and Geo-tag. Additionally, Foursquare allows developers to build audience segments for analysis and measurement. JSON is the preferred response format.

All of the information which is relevant to my analysis is in the ‘item’ key of the foursquare json file. Using the get request command I extracted following fields and converted them into python data frame for easy manipulation and analysis.

|  |  |
| --- | --- |
| Field | Description |
| id | A unique string identifier for this venue. |
| name | The best known name for this venue. |
| categories | An array, possibly empty, of [categories](https://developer.foursquare.com/docs/resources/categories) that have been applied to this venue. |
| rating | Numerical rating of the venue (0 through 10). Returned as part of an explore result, excluded in search results. Not all venues will have a rating. |
| tips | Contains the total *count* of tips and *groups* with *friends* and others as *groupTypes*. Groups may change over time. |
| location | An object containing none, some, or all of address (street address), crossStreet, city, state, postalCode, country. All fields are strings.  Location Latitude  Location Longitude |

The motive of this project is to present an implemented design of the recommendation system. The system solely focuses on the user’s interests and does not take into account the reviews of other people. Since if there may exist such a case that a restaurant may not be preferred by the majority but only cater to the needs of a few, including the user, then the system may give a negative remark for the place, thus misguiding the user. Using Foursquare API and kmeans clustering I found the neighborhoods for the given type of cuisine.

# Result

Following figure 1 and 2 shows the result table and the map showing the restaurants. For some locations, the neighborhood information is not present in the data which is extracted from the foursquare API. Hence, for easy access of the location, I also extracted the cross street information for user. This will help the users to locate the restaurant.

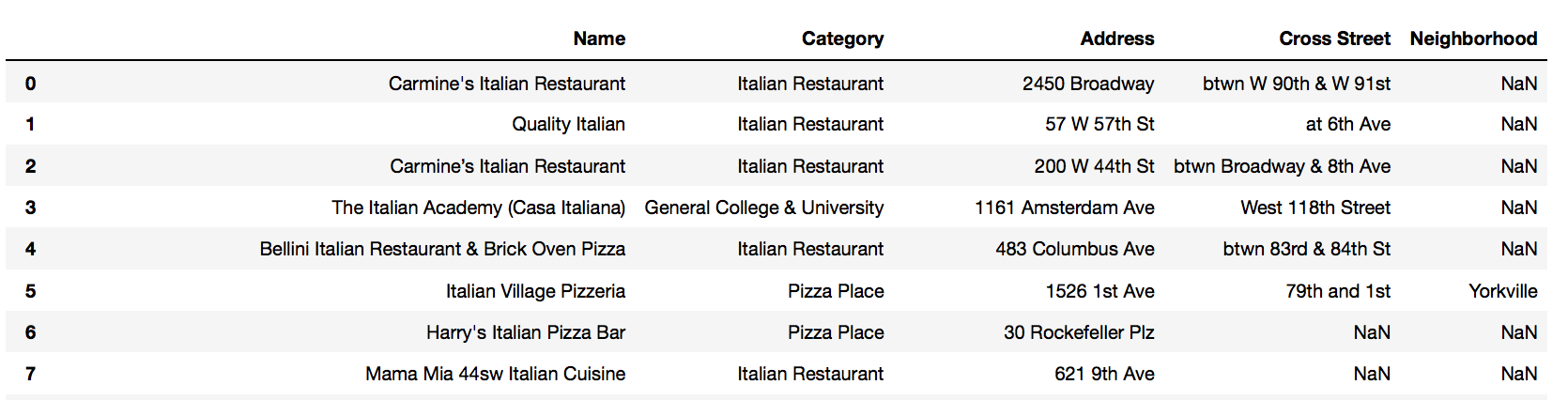


Figure 1: List of first 10 Italian Restaurants.



Figure 2: Map showing location of Italian restaurants.

# Discussion

Unsupervised learning for subcategorization is relatively more difficult. In this project, I applied clustering on the venue category in order to find the restaurants for a specific cuisine. The intuition is, let’s say, if I put ‘Italian’ word in search query, I will be presented with the restaurants with Italian cuisine. However, the difficulty is, it’s not obvious what each cluster corresponds to. In some cases the Italian word match with some other types of venues such as ‘Italian Cultural Institute’.

One way to figure this out is to look at the percentage of existing labels. For example, if in a cluster, 50% of restaurants are “bar”, 25% are “night life”, then we could reason this cluster corresponds to the bar type of restaurants. Though we do observe this in some of the clusters, there are also clusters with mixed labels that are not easy to interpret. A more fundamental question to ask is, is the clustering based on restaurant types? Or, is it perhaps more related to something else like star­ rating, cost, or other latent factors? A key lesson for me here is that unsupervised learning doesn’t always give us the result we expect.

# Conclusion

Designing this system was a very useful exercise and I believe that the literature survey I have done was very useful for me in order to understand the logic behind the recommendation systems. During my survey I came across the concept of *durable\_rules* which can be used for real time analytics. I researched on that; but I realized that constructing rules for such a food recommendation may be a difficult work. So I did not used it.

Implementing such a project demonstrated how a knowledge based system can actually encapsulate knowledge. It also showed the limitations of knowledge based systems and the assumptions that must be made when designing them. As an instance for these assumptions, I have assumed that the user likes the recommendation according to the type of cuisine.

# Reference

<https://github.com/jruizgit/rules>

<https://developer.foursquare.com/docs/api/venues/details>