Data

Data is the heart of any data science problem as it has a major effect on the final answers.

Data Source

Our problem requires location data that describes hotels and restaurants around New York City. Specifically, location of hotels and restaurants, number of restaurants around hotels, foot traffic in restaurants and their category. For retrieving this data Foursquare API was utilized. Foursquare API search feature would be used to get the location of hotels and restaurants. Due to Foursquare API call limitations, the number of hotels in search query is set to 50 and radius parameter to 500.

Data Cleaning

For the hotels dataset, a “search” call was made to the Foursquare API with the search query “Hotel”. The aim was to get the hotels around New York City. The results got from the call were first converted into pandas data frame. There were a lot of unwanted rows and columns in the data frame. The columns other than the name and location of hotels were dropped and the rows containing only hotel as the value for the column name is taken.

For the restaurant dataset, a “explore” call was made to the API to get a glimpse on which venues were most trending. The results got from the call were first converted to a pandas data frame. There were a lot of unwanted columns and rows in the data frame. The columns other than the venue name and venue location were dropped and the rows containing restaurant as the value for the column venue name is taken.

Feature Selection

For the intended analysis to be made, two new data frames were created one for hotels and the other for restaurants, each holding only the name , latitude and longitude coordinates of the original datasets as these are the features that Machine Learning algorithms requires for our analysis.

The project will require the use of many data science skills: Data Cleaning, Data wrangling, working with Foursquare API, Visualization and Machine Learning.