Coursera Capstone

IBM Applied Data Science Capstone

Opening a New Chinese restaurant in San Diego, California

By: Shaun Clarke

November 2019

Data

To solve this problem, we will need the following data:

- List of neighborhoods in San Diego. This defines the scope of this project, which is confined to the city of San Diego, which is in California.

- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and to get the venue data.

- Venue data, particularly data related to Chinese restaurants. We will use this data to perform clustering on the neighborhoods.

Sources of data and methods to extract them

This wiki page (https://en.wikipedia.org/wiki/Category:Neighborhoods_in_San_Diego) contains a list of neighborhoods in San Diego, with a total of 170 neighborhoods. I will be using web scraping techniques to extract the data from the wiki page, using Python requests and beautifulsoup packages. For the San Diego neighborhoods I will get the geographical coordinates of the neighborhoods using the Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods. After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. The Foursquare API will provide many categories of the venue data, we are particularly interested in the desert bar category which will help us solve the business problem stated above. This is a project that will make use of the data science skills covered in this course. Skills such as working with APIs (Foursquare), data cleaning, data wrangling, machine learning (K-means clustering) and visualizing maps with folium. In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning techniques that were used.