

IBM Capstone Project Week 4

1. A description of the problem and a discussion of the background.

Some stakeholders wanted to open a new Pizza chain Restaurants in the city of Toronto. They were looking for the best place to open their first new Pizza restaurant. They know that Toronto is a very dynamic city and that Pizza food is highly appreciated by the populations. They know that there are lots of Pizza restaurants in the city, but they would like to avoid unnecessary competition against existing Pizza places. So, they wanted to open their first Pizza restaurant in locations where there as less pizza restaurants.

As a data scientist, I have been asked by these stakeholders to resolve this business problem for them by finding the right places in the city of Toronto where they should start opening their Pizza restaurant.

2. A description of the data and how it will be used to solve the problem.

To resolve this problem, I will use the following data:

- a. The Toronto city data from Wikipedia :
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
From this Wikipedia page, I will extract all the Boroughs, Neighborhoods and Postal Codes of the Toronto city required to resolve this problem
- b. The Toronto Geospace data: https://cocl.us/Geospatial_data. This is a csv file which contains all the geographical coordinates of all the Toronto Neighborhoods. That will be required to know all the latitudes and longitudes of all the Neighborhoods in Toronto.
- c. The Foursquare API: <http://developer.foursquare.com>. I will use the Foursquare's explorer API tool to collect all the venues, etc near by each location.

3. Methodology used to resolve the problem

- a. First, like what we did in week 2 and week 3 labs, I will collect and extract all the Neighborhoods and boroughs data of Toronto from the Wikipedia page. Then, I will clean and merge these data with the Geospace data in one single Panda DataFrame.
- b. Then, I will explore the data. I will first display all the Neighborhoods in a Foliummap. I will use the Foursquare API tool to extract all the venues. My Foursquare account is a free account, so I will be limited to only 100 venues by the Foursquare API in a radius of 1000 meters. Once I get all the venues, I will focus on the Pizza Places only. Continuing the exploration, I will merge my first Data Frame with the venues returned by Foursquare. I will analyze these data focusing on the Pizza Places only. I will use the matplotlib library to plot the amount of Pizza Places we have per Borough.
- c. And finally, I will use the unsupervised machine learning algorithm, K-mean clustering, to cluster the neighborhoods in different clusters of similarities. I will use first the Elbow method to determine the best K-mean value to use and will use the Folium package to get an overview of the neighborhoods by cluster on the map.