Coursera Capstone Project Week 4

PROJECT NAME: IDENITFYING LOCATION FOR A MALL IN THE SUBURBS OF NEW ZEALAND AS PART OF CAPSTONE PROJECT IBM DATA SCIENCE COURSE

BY

ANAND SUNDARESAN

03 FEBRUARY 2020

INTRODUCTION

Cities often have malls in the city center and this is a nice place for people to engage themselves in looking up various items of their interests. Malls are central to every aspect of a social life right from dining, clothing, beauty, movies and playing games and amusements for children to name a few. Hence malls are a favourite destination of choice for businessmen to open their businesses and market products where they gain with a vast array of customers who frequent the malls anytime of the day. The Malls are growing due to the popularity and based on location and frequency of customers we will determine using the data available employing data science methods to identify the pertinent location for creating a new mall in the most opportune location

BUSINESS PROBLEM

The key business question is to identify the location within the city most suited to setup a shopping mall. Based on exploratory methods employed using data science, techniques learnt within the data science and using geolocation and other APIs the aim of this project is to solve this business problem. Which is the best place to open the shopping mall?

Intended Audience

Data obtained from the data science techniques will aim to assist the following actors.

 Mall developers especially property development agencies Advertising agencies Business promoters Government councils and neighbourhood centres

TECHNOLOGY SOLUTION

DATA

Sources: https://en.wikipedia.org/wiki/Category:Suburbs_in_New_Zealand

Location Mapping Sources:

- Location identification using Latitude and Longitude coordinates using APIs
- Plotting the map of the area
- Getting Venues that will aid in neighbourhood identification employed with Clustering techniques

PROCESS

a. Use the Wikipedia to get a list of suburbs from New Zealand as per the link shown above
b. WebScrape the data using the beautifulsoup and python libraries
c. Get geougraphical coordinates using Python packages to obtain latitude and longitude
d. Use API from Foursquare to get Venue data for the neighbourhoods identified in Step c

SOLUTION

A List is obtained from the Wikipedia page that provides us with the intial data sample. The link is freely available on the internet https://en.wikipedia.org/wiki/Category:Suburbs_in_New_Zealand.

Using the techniques learned so far, the web scraping technique is employed using the scripts available in Python. The list of neighbourhoods from the sample data is analysed using the beautifulsoup package libraries and this data is used which is a list of data that can be the next step in sampling. In addition to this, the location information is easily available with just calls to the Geocoder package and with a host of features to get the correct address from geographical coordinates simply by passing the required information. The information is then plotted in a graphical manner with plot packages available easily with few lines of code to give us a pictorial view of the neighbourhood

Foursquare API is a very useful API that will get us the venues from this data in a given radius in this case upto 2000 meters. Using the secred id and password to access this API, we call Foursquare simply with the coordinates of the neighbourhoods within the python code. The data returned in a JSON format can be easily interpreted and extracted for further analysis. We now proceed to get the venue name, category and the geographical coordinates. We then identify with additional simple python code, to get the unique categories from the venues we fetched earlier.

Analysis is performed with the statistical mean of the occurrence of each category once the data has been grouped successfully. The data is now enabled for clustering by filtering the “Shopping mall” as the venue category which we are interested in

The Clustering of data is performed using the K-Means clustering which we had learnt earlier. K-Means is a form of clustering employed wherein the algorithm identified k number of centroids, allocating evert data to the nearest cluster and keeping centroids as minimum as possible. This is an unsupervised machine algorithm and very useful in the exercises we have learnt to understand and hence this is employed for the project. Now we cluster the neighbourhoods into clearly 3 clusters based on the frequency of occurrence of “Shopping mall”. Results help identify that the neighbourhoods have different sets of frequencies of available shopping mall in the respective neighbourhood and likewise different sets of frequencies where there are no or fewer shopping malls.

Summarising the results

In Summary, it has been found that Most of the shopping malls are concentrated in the central area of Auckland with cluster 0 having many malls in the neighborhoods. This represents a great opportunity and high potential areas to open new shopping malls as there is very little to no competition from existing malls.