IDEAL SHOP LOCATION PREDICTION USING MACHINE LEARNING ALGORITHM

By

AKHILA JAYAKUMAR

INTRODUCTION

In this project, we will use the knowledge gained through earlier courses in this certification such as Python for Data Science course, the data visualization with Python course, and finally, the data analysis with Python course along with some additional skills segmentation and clustering. Through previous modules, we explored New York City and the city of Toronto and segmented and clustered their neighborhoods. Both cities are very diverse and are the financial capitals of their respective countries. Given the above analysis, our client is looking to to open a coffee shop in any one of the countries or both, where would you recommend that they open it? This is the basic Data Science problem statement, and we need to come up with an ideal suggestion based on our analysis which gives maximum profit to our client.

DATA DESCRIPTION

We have two data sets which contain the borough and neighbourhood information of both cities in question. While the data pertaining to New York city was readily available for analysis, we scraped the necessary data for Toronto from the relevant Wikipedia page using the library function **BeautifulSoup** and built our own data set by cleansing and scrubbing the unformatted data. We have also combined the Geospacial coordinates data into the newly built Toronto data set in order to create a pure data set which is further used for analysis, segmentation and clustering.

Both datasets needed for the analysis are pushed to the Github repository along the necessary Geospacial coordinates data for Toronto city as json and csv files.

DATA SCIENCE METHODOLOGY

Data Analysis and Testing:

First we are going to analyze and cluster the neighbourhoods in the New York city and then do the same with Toronto city. For this we are using the k-means clustering algorithm and map function, Folium to plot the results. We have to import the readily available data set for New York city from https://geo.nyu.edu/catalog/nyu_2451_34572 and for Toronto, we are going to scrap the needed data by using the ultility **BeautifulSoup** from the Wikipedia web page, https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.

We are going to do the necessary scrubbing and cleansing of raw data to remove the unwanted details and build dataframes so as to build clusters. Now the data frame consists of Borough, Neighbourhoods, Latitude and Longitude coordinates. The New York dataframe has 5 boroughs and 306 neighbourhoods, while Toronto has 11 boroughs and 103 neighbourhoods. The optimum k is found as 5 for New York city and as 3 for Toronto after iterative analysis and testing.

RESULTS AND DISCUSSION

The New York City neighbourhood, Manhattan has been segmented into 5 clusters and Toronto city neighbourhood, West Toronto has been divided into 3 clusters. Based on the analysis done on both cities, we have identified two Neighbourhoods one each in New York city and Toronto city respectively which are ideal candidates for starting a coffee shop for our customer. We conclude that **Soho** in Manhattan is best suited in case of New York city and **Dufferin** in West Toronto as the ideal locations for starting up the business.

The places are chosen based on the lack of coffee shops in the locality and competitors, closeness to city as well as the coastal line in order to bring out maximum profit attracting potential customers and tourists. Both places do not have a coffee shop or a restaurant in 500 m radius span and the most frequented places also lack a coffee shop.

CONCLUSION

Machine Learning algorithm is used efficiently to solve a potential business problem. While many other techniques can also be used to reach at this conclusion, K- means clustering and visual representation of results using Folium are best suited for the data science problem in hand.

REFERENCE LINKS ON GITHUB

Data sets:

▶ For New York city:

▶ https://github.com/Akhi18/First_Github/blob/master/Segmentation%20and%20Clustering%20of%20neighborhoods%20in%20New%20York%20City.ipynb

▶ For Toronto:

▶ https://github.com/Akhi18/First_Github/blob/master/Segmenting%20and%20Clustering%20Neighbourhoods%20in%20Toronto%20city.ipynb

Project Report:

▶ https://github.com/Akhi18/First_Github/blob/master/Capstone_Project.ipynb