IBM APPLIED DATA SCIENCE CAPSTONE COURSE

THE BATTLE OF NEIGHBORHOODS PROJECT – WEEK 5 REPORT

1. INTRODUCTION

This is a capstone project for IBM Applied Data Science and Data Science Professional Specializations. In this project, I will discover the opportunities to found "yoga studio" in Toronto, since coworking spaces gain huge popularity between white collars, students which are trying to be healthy. Therefore, the people who are planning to find suitable places for founding the yoga studios are the ***main target audience*** for my project. At the end of this project, this target group can find the best alternative neighborhoods to open yoga studio in Toronto, Canada.

2. DATA

To extract the best alternative neighborhoods in Toronto, it is needed to have list of neighborhoods in Toronto, coordinates of these and finally venue data related to existing yoga studios in these neighborhoods.

To utilize these data, firstly, I scrap the Toronto postal codes, boroughs and neighborhoods from related Wikipedia page via Python beautifulsoup package. Then, getting the coordinates of each of these postal codes from this link (http://cocl.us/Geospatial_data) in the csv format. And finally, by utilizing Foursquare API, I get the yoga studios' venues data.

By merging the first 2 data sources, I have the coordinates of each neighborhoods in Toronto. Then, with the help of Foursquare venue data, I will list all yoga studios

in these neighborhoods. Lastly, I cluster the neighborhoods based on all venues and observe the clusters which offers the opportunity founding the new yoga studio.

3. METHODOLOGY

First, I need to get the list of neighborhoods in Toronto, Canada. As mentioned above, I used this Wikipedia page to access the required information: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.

Via beautifulsoup package of Python, I parsed HTML and I did find the table items in the HTML like <td> and <tr>. With this way, I get 3 different lists, postal codes, boroughs and neighborhoods. Then, I created a pandas data frame which has these 3 columns. By applying preprocessing steps like removing “not assigned” cells and grouping data frame over boroughs I created new data frame. When it comes to coordinates of these, I imported the required csv file provided by IBM team, located in http://cocl.us/Geospatial_data. After gathering these coordinates, I visualize the map of Toronto using Folium package to verify whether these are correct coordinates. Next, I applied Foursquare API to pull the list of top 100 venues (LIMIT) within 500 meters radius. I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. With this data, I can also check how many unique categories that I can get from these venues. Then, I analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category to apply clustering onto these neighborhoods.

Finally, I performed the clustering method by using unsupervised clustering method k-means. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster.

I have clustered the neighborhoods in Toronto into 5 clusters based on their frequency of occurrence for each venue with the aim of observing popular neighborhoods. Based on the results (the concentration of clusters), I listed the clusters which have at least one yoga studio. Finally, I have a chance to recommend the ideal neighborhoods to run yoga studio.

4. RESULTS

The results of 5-means clustering of Toronto neighborhoods based on all venues show me that cluster “0” has 11 yoga studio and cluster “4” has 3 yoga studios. The rest of the clusters does not have any yoga studio. Based on these findings, cluster “1”,”2” and “3” can offer best neighborhood alternatives to open yoga studio.