CAPSTONE PROJECT - THE BATTLE OF NEIGHBORHOODS

1. INTRODUCTION

1.1 Problem

In this project, I am interested in New York City data. First, we will find the most visited commercial shop according to the number of check-ins, then we will try to find the neighborhoods that are lacking the selected type of shop which could be potential business opportunity. So, the aim is to explore the two cities for tourist who wants to visit in them keeping in mindthe areas of food, hotels, museums and much more.

1.2 Target Audience

The target audience of this report is any one that is interested in opening a shop but have no idea what kind of and in which neighborhood After selecting the place they want to visit, people can analyse what all places can be visited and activities that can be done in the same city

2. DATA SECTION

For this project, the following data will be taken in account.

1) List of Boroughs and their corresponding Neighborhoods in Toronto, Canada which has been taken from the following Wikipedia page.
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

2) Geographical coordinates of every postal code in Toronto, Canada which have been taken from the following csv file.
https://cocl.us/Geospatial_data

3) Dataset containing the Boroughs and the Neighborhoods that exist in each borough along with the geographical coordinates of each neighborhood which has been taken from the following json file.

https://cocl.us/new_york_dataset

It contains 227,428 check-ins in New York city. The data contains two files in tsv format. Each file contains 8 columns, which are:

1. User ID (anonymized)
2. Venue ID (Foursquare)
3. Venue category ID (Foursquare)
4. Venue category name (Foursquare)
5. Latitude
6. Longitude
7. Time zone offset in minutes (The offset in minutes between when this check-in occurred and the same time in UTC)
8. UTC time

After extracting and reading the data, we will translate the above data into a Pandas data frame
for processing which would look like this. These are the data elements that are needed when we
call Foursquare web service call in order to get the venues available in that neighborhood
(Neighborhoods are not included here)

Then, the machine learning technique, “Clustering” is used to segment the neighborhoods with similar objects on the basis of each neighborhood data. These objects are given priority on the basis of foot traffic (activity) in their respective neighborhoods. This will help to locate the tourist’s areas and hubs, and then we can judge the similarity or dissimilarity between two cities on that basis.

3. METHODOLOGY SECTION

The required data to be worked upon is taken from various resources mentioned in the data section and converted into the data frames to perform the exploration, visualization and analysis using machine learning techniques.

After getting the much needed data frames of the two boroughs, the geographical coordinates of Downtown Toronto and Manhattan are determined using the geolocator package. The coordinates derived are as follows :-

Downtown Toronto - latitude 43.6563221 & longitude -79.3809161
Manhattan - latitude 40.7896239 & longitude -73.9598939

Using the geographical coordinates of both the boroughs, we then visualize the data using the folium package before and after completing the step of analysis.

Now, we start doing the analysis of the data frames using the one hot encoding in which '1' is assigned if a venue category exists or else '0' is assigned. Once the encoding has been done, the mean of the frequency of occurrence of each category is calculated and top ten venues are picked for each neighborhood on that basis using the data in Foursquare API .

Finally, to compare the neighborhoods of both the boroughs, namely Downtown Toronto and Manhattan, the neighborhoods of the two boroughs are clustered using the machine learning technique, k-means clustering using the KMeans function available in the scikit-cluster package.

In this case, the number of clusters considered are 5 for each of the boroughs that means the neighborhoods of both boroughs are clustered in 5 clusters.

Henceforth, a data frame which consists of name of the borough, name of the neighborhood, geographical coordinates of each neighborhood, cluster labels and the top 10 venues for each of the neighborhood is constructed for the two boroughs selected for the problem, i.e. Downtown Toronto and Manhattan.

After the neighborhoods have been clustered, the step of visualization is again repeated to visualize the clusters on the map of Downtown Toronto and Manhattan.

Now, since the clusters of the neighborhoods have been defined and visualized on the maps of their respective boroughs, they are finally examined individually in order to determine the discriminating venue categories that distinguish each cluster.

On examining the clusters of neighborhoods in both boroughs, Downtown Toronto and Manhattan, the results, observations, recommendations and conclusions are made which have been described in the following section of the report.

4. RESULTS SECTION

After clustering the data of the respective neighborhoods, both cities or boroughs, namely Downtown Toronto (Toronto, Canada) and Manhattan (New York, United States) have venues which can be explored and attract the tourists all over the world. The neighborhoods are much similar in features like theaters, opera houses, food places, clubs, museums, parks etc. As far as dissimilarity is concerned, it differs in terms of some unique places like historical places and monuments.

5. OBSERVATIONS AND RECOMMENDATIONS

When the tourist places in both the boroughs are compared, it can be observed that the historical place is only situated in Downtown Toronto and the Monument or landmark venue is in Manhattan neighborhoods. Similarly, Airport facility, Harbor, Sculpture garden and Boat or ferry services are also available in Downtown Toronto while venues like Nightlife, Climbing gym and Museums are present in Manhattan.

As far as recommendations are concerned, Downtown Toronto Neighborhoods will be recommended to visit first. The tourists have easy travelling access due to the Airport facility, which not only saves time but also helps to save money. This saved money can be utilized to explore more attracting venues.

6. CONCLUSION SECTION

The Downtown Toronto and Manhattan neighborhoods have more like similar venues. As we know that every place is unique in its own way, so that argument is present in both neighborhoods. The dissimilarity exists in terms of some different venues and facilities but not on a larger extent.