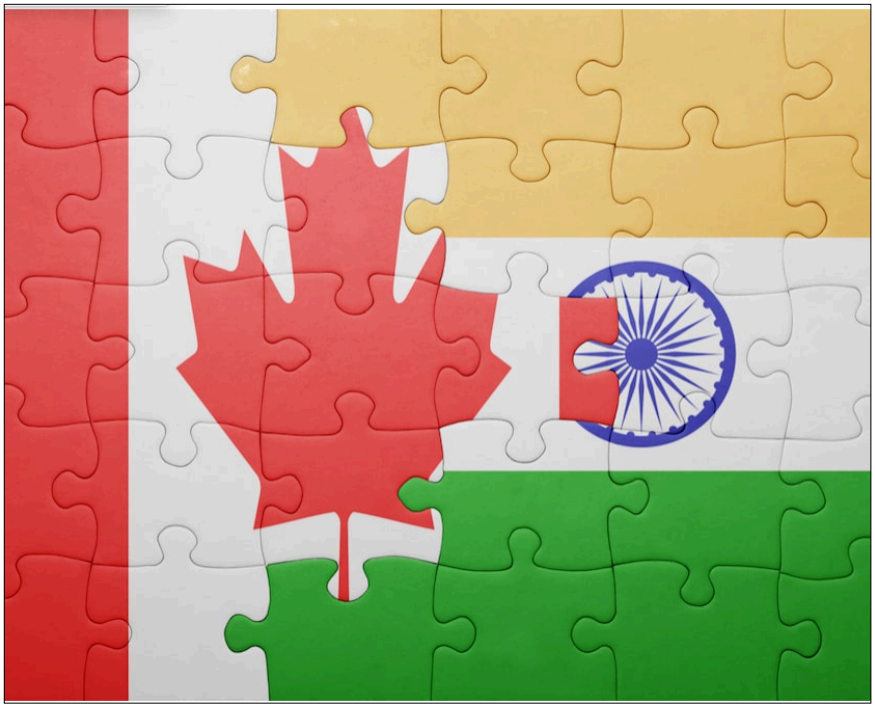


Relocation Recommendation from Delhi to Toronto Using Neighbourhood Clustering

Prepared for: Coursera Capstone
Prepared by: Naman Sharma
28 April 2020

TABLE OF CONTENTS

- Introduction to Problem
- Data Selection and Preparation
- METHODOLOGY
- RESULTS
- DISCUSSION
- CONCLUSION

INTRODUCTION

IN THIS PROJECT, WE WILL TRY TO FIND AN OPTIMAL NEIGHBOURHOOD FOR A PERSON RELOCATING FROM THE CITY OF DELHI(INDIA) TO TORONTO(CANADA)

Background

People often move until they find a place to settle down where they truly feel happy, satisfies professionally or to seek higher education. Our wants and needs change over time, prompting us to eventually leave a town we once called home for a new place that will bring us growth. We often move to a new area without knowing exactly what we're getting into, forcing us to turn tail and run at the first sign of discomfort. To minimise the chances of this happening, we should always do proper research when planning our next move in life.

Problem

Specifically, this report will be a suggestion to those students and employees to whom relocating does bring a huge change in lifestyle and they want to adapt to their new-habitat as easily as possible. Finding a suitable location for relocation for anyone in major cities like Toronto proves to be a daunting task. Various factors such as Public Facilities, Favourite Cuisine, restaurants, religious practices, etc matters to the people. Hence, migrant can bolster their decisions using the descriptive and predictive capabilities of data science.

Interest

We need to find locations(Neighbourhood) that have a similar nearby venues in Toronto compared to persons native Neighbourhood in Delhi. Also, we need locations that are close to Work/Study location in Toronto. We will use our data science powers to generate a few most promising neighbourhoods based on this criteria. Advantages of each area will then be clearly Expressed so that best possible final location can be chosen.

DATA SELECTION AND PREPARATION

Based on definition of our problem, we will gather data for both the cities, the city of Delhi and the city of Toronto. We will consider following factors before we begin are search for data:

- Borough nearby Office/University location in the city of Toronto.
  - FOR OUR PROJECT PURPOSE WE ASSUME THAT THESE BOROUGHS ARE **NORTH YORK AND SCARBOROUGH**
- Similar neighbourhood in Toronto as in Delhi.
- Picking up data pertinent to following factors:

1. Neighbourhood: Name of the neighbourhood in the Borough.
2. Name of the Borough.
3. Latitude of the Borough.
4. Longitude of the Borough.
5. Venue nearby our neighbourhood.

In our project we will gather and process data with following techniques:

- Acquire the names and boroughs of the neighbourhoods by scrapping a wikipedia pages and assigning them in “csv" coded data frames.
- After we have the names of all the neighbourhoods, we will geocode them using the geopy.geocoder (Nominatim). And thus will add longitude and latitude data corresponding to borough location and neighbourhood data.
- Next, we use the foursquare API to find all types of venues within a 500 meter radius for every neighbourhood. Then these are merged with the existing data set
- Normalise the data using one hot coding and finally preparing it for cluster formation.
- Merging the One hot Coded data and the Most Frequent venue data for proper visualisation of problem.

METHODOLOGY

1. We will Scrape the Data of Delhi from a Wikipedia page into the Notebook and convert it to usable Data Frame for Processing and Cleaning.

Scrapping the Wikipedia Page

```
#accessing the web page by Http request made by requests library

req = requests.get("https://en.wikipedia.org/wiki/Neighbourhoods_of_Delhi").text
soup = BeautifulSoup(req, 'html5lib')
div = soup.find('div', class_="mw-parser-output" )
print("web Page Imported")
```

```
web Page Imported
```

```
   #Code to extract the relevent data from the request object using beautiful soup

data = pd.DataFrame(columns=['Borough','Neighborhood'])
i=-1
flag = False
no=0
prev_borough = None
for child in div.children:
    if child.name:
        span = child.find('span')
        if span!=-1 and span is not None:
            try:
                if span['class'][0] == 'mw-headline' and child.a.text!='edit':
                    prev_borough = child.a.text
                    i+=1
                    flag = True
                    continue
            except KeyError:
                continue
        if child.name=='ul' and flag==True:
            neighborhood = []
            for ch in child.children:

                try:
                    data.loc[no]=[prev_borough,ch.text]
                    no+=1
                except AttributeError:
```

2. Geocoding every neighbourhood and obtaining the latitude and longitude values accompanying the respective neighbourhood.

Geocoding every neighborhood

```
lat_lng = pd.DataFrame(columns=['latitude','longitude'])
geolocator = Nominatim(user_agent="Delhi_explorer")
for i in range(184):
    address = data['Neighborhood'].loc[i]+',Delhi'
    try:
        location = geolocator.geocode(address)
        lat_lng.loc[i]=[location.latitude,location.longitude]
    except AttributeError:
        continue
```

3. Obtain Delhi's and Toronto Latitude and Longitude for basic visualisation and folium map.

4. Merging the Latitude Longitude Values with actual Neighbourhood Values.

```
delhi_neighborhood_data.dropna(inplace=True)
delhi_neighborhood_data.head()
```

:

	Borough	Neighborhood	latitude	longitude
0	North West Delhi	Adarsh Nagar	28.714401	77.167288
1	North West Delhi	Ashok Vihar	28.699453	77.184826
2	North West Delhi	Azadpur	28.707657	77.175547
3	North West Delhi	Bawana	28.799660	77.032885
5	North West Delhi	Dhaka	28.708698	77.205749

5. Visualising the obtained data set by using the folium maps.

6. Inspect Different Neighbourhoods in the city of Delhi by gathering Venues nearby and within that neighbourhood. For this we will use FOURSQUARE API:

(The Foursquare Places API provides location based experiences with diverse information about venues, users, photos, and check-ins. The API supports real time access to places, Snap-to-Place that assigns users to specific locations, and Geo-tag. Here we are using the explore api call and filtering the search only to find venues that are identified as Venues.)

7. Exploratory Data Analysis : Normalisation of the data for clustering and obtaining Data frame of frequently visited venues nearby neighbourhoods of Delhi. This type of analysis helps us in predicting why and which model to be used for machine learning programming.

Exploratory Data Analysis : Normalization of the data for clustering

```
1]: # one hot encoding
delhi_onehot = pd.get_dummies(delhi_venues[['Venue Category']], prefix="", prefix_sep="")
delhi_onehot.drop('Neighborhood', axis=1, inplace=True)

# add neighborhood column back to dataframe
delhi_onehot=pd.concat([delhi_venues['Neighborhood'], delhi_onehot], axis=1)

delhi_onehot.head()
```

1]:

	Neighborhood	ATM	Accessories Store	Afghan Restaurant	Airport	American Restaurant	Antique Shop	Arcade	Art Gallery	Asian Restaurant	Athletics & Sports	Australian Restaurant	BBQ Joint	Baby Store
0	Adarsh Nagar	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Adarsh Nagar	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Adarsh Nagar	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Ashok Vihar	0	0	0	0	0	0	0	0	0	1	0	0	0
4	Ashok Vihar	0	0	0	0	0	0	0	0	0	0	0	0	0

We have the desired Delhi_Venues Data Frame now, our methodology involves making a similar Data Frame for the locations of Toronto which are nearby our migrant's place our school. We will for our project purpose assume that our Migrant prefers Boroughs in TORONTO by our migrant are the adjacent Borough of North York and Scarborough

Now we perform similar analysis and obtain data-set for the selected borough of Toronto

8. Analysing the Venues for Clustering

OUR Migrant MIGHT BE LIVING IN ANY OF THESE NEIGHBOURHOODS IN DELHI, WE WILL MAKE CLUSTERS OF SIMILAR NEIGHBOURHOODS AND COMPARE THESE CLUSTERS WITH THE NEIGHBOURHOODS CLUSTER IN CITY OF TORONTO

- Our goal here is to find the neighbourhoods with similar venues an combine them into clusters
- The most intuitive idea would be to find neighbourhoods that have similar patterns of most frequently visited venues.
- This is achieved by clustering the neighbourhoods of the basis of the venues data we have acquired. Clustering is a predominant algorithm of unsupervised Machine Learning. It is used to segregate data entries in cluster depending of the similarity of their attributes, calculated by using the simple formula of euclidian distance.
- First we organised the data of frequently visited venues in both the cities by adding the credentials like Borough name, latitude and longitude to our data set.
- This will be of help in visualisation after cluster formation, so we did that for both Delhi and Toronto !!

8. Combining both Cities Venue data for further visualisation and Analysis comparison on the Basis of K-Means Clusters

9. Merging one hot Coded / Normalised Data

10. We then Merged both the Delhi Neighbourhood Data and Toronto Neighbourhood data so that we could make clusters of similar Neighbourhoods by most common frequent Venues in both of them.

11. Applying the clustering algorithm : K-Means Algorithm¶

    Reason: K-means algorithm is an iterative algorithm that tries to partition the dataset into K predefined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum.

[40]:
```
# set number of clusters
kclusters = 7

Cluster_Data = Cluster_Data.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(Cluster_Data)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:15]
```

[40]:
```
array([4, 2, 4, 2, 5, 0, 2, 6, 2, 2, 2, 4, 2, 3, 2], dtype=int32)
```

[41]:
```
# add clustering labels
Frequent_venues.insert(0, 'Cluster Labels', kmeans.labels_)
Frequent_venues.head(10)
```

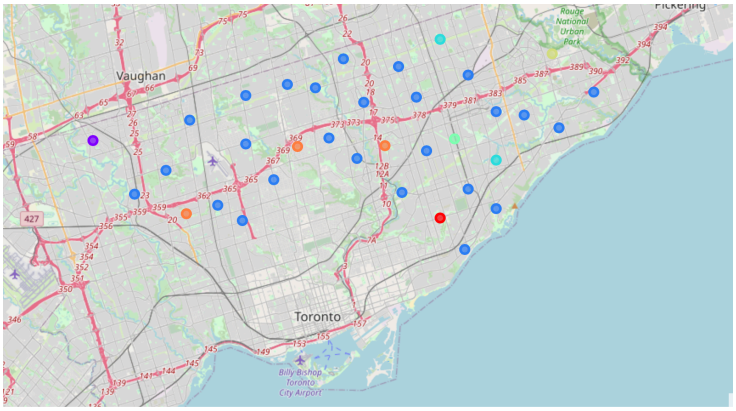
[41]:

	Cluster Labels	Borough	Neighborhood	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8 C
0	4	North West Delhi	Adarsh Nagar	28.714401	77.167288	Indian Restaurant	Train Station	Light Rail Station	Farmers Market	Fried Chicken Joint	French Restaurant	Food Truck	Foo
1	2	Scarborough	Agincourt	43.794200	-79.262029	Breakfast Spot	Lounge	Chinese Restaurant	Latin American Restaurant	Women's Store	Diner	Dog Run	Dis
2	4	South Delhi	Alaknanda	28.529336	77.251632	Indian Restaurant	BBQ Joint	Middle Eastern Restaurant	Pizza Place	Coffee Shop	New American Restaurant	Food & Drink Shop	Ste

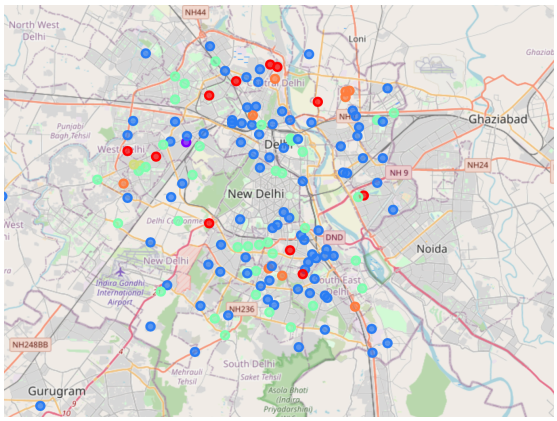
RESULTS:

1. Visualising Clusters : Folium maps were used to see the positional location of the labelled clusters in the city of Delhi and Toronto.

COLOR-CODED CLUSTERS IN TORONTO



COLOR-CODED CLUSTERS IN DELHI



2. On the Basis of Recommendation Tables:

Different clusters were divided in different tables and for each cluster tables were made for both Toronto and Delhi containing the neighbourhood information, then these tables were displayed side by side.

A few exemplar recommendations are provided in this report and the rest can be seen in the Jupiter notebook attached at the GITHUB:

Recommendation 1

	Borough	Neighborhood
0	North West Delhi	Ashok Vihar
1	North West Delhi	Keshav Puram
2	North Delhi	Nehru Vihar
3	North Delhi	Sant Nagar
4	North East Delhi	Mukherjee Nagar
5	North East Delhi	New Usmanpur
6	East Delhi	Mayur Vihar
7	South Delhi	Defence Colony
8	South Delhi	Kailash Colony
9	West Delhi	Dhaula Kuan
10	West Delhi	Dhaula Kuan
11	West Delhi	Rajouri Garden

	Borough	Neighborhood
0	Scarborough	Golden Mile , Clairlea , Oakridge

Recommendation 2

	Borough	Neighborhood
0	West Delhi	Kirti Nagar

	Borough	Neighborhood
0	North York	Humber Summit

11

Recommendation 3

	Borough	Neighborhood		Borough	Neighborhood
0	North West Delhi	Azadpur	0	North York	Hillcrest Village
1	North West Delhi	Dhaka	1	North York	Fairview , Henry Farm , Oriole
2	North West Delhi	Jahangirpuri	2	North York	Bayview Village
3	North West Delhi	Kingsway Camp	3	North York	York Mills , Silver Hills
4	North West Delhi	Model Town	4	North York	Willowdale , Newtonbrook
5	North West Delhi	Narela	5	North York	Willowdale
6	North West Delhi	Rithala	6	North York	Willowdale
7	North Delhi	Shakti Nagar	7	North York	Don Mills
8	North Delhi	Bara Hindu Rao	8	North York	Don Mills
9	North Delhi	Chawri Bazaar	9	North York	Bathurst Manor , Wilson Heights , Downsview North
10	North Delhi	Civil Lines	10	North York	Northwood Park , York University
11	North Delhi	Kamla Nagar	11	North York	Downsview
12	North Delhi	Kashmiri Gate	12	North York	Downsview
13	North Delhi	Lahori Gate	13	North York	Downsview
14	North Delhi	Paharganj	14	North York	Downsview
15	North Delhi	Pratap Nagar	15	North York	Victoria Village
16	North Delhi	Sadar Bazaar	16	North York	Bedford Park , Lawrence Manor East
17	South Delhi	Sangam Vihar	17	North York	Lawrence Manor , Lawrence Heights
18	North Delhi	Sarai Kale Khan	18	North York	Glencairn
19	North Delhi	Sarai Rohilla	19	North York	Humberlea , Emery
20	North Delhi	Shakti Nagar	20	Scarborough	Rouge Hill , Port Union , Highland Creek
21	North Delhi	Shastri Nagar	21	Scarborough	Guildwood , Morningside , West Hill

Recommendation 4

	Borough	Neighborhood		Borough	Neighborhood
0	North West Delhi	Bawana	0	Scarborough	Scarborough Village
			1	Scarborough	Milliken , Agincourt North , Steeles East , L'Amoreaux East

DISCUSSION:

Our Analysis was done on over 188 neighbourhoods Combined in Delhi and Toronto, containing 100 venues each within a radius of 500 meters. We segregated these neighbourhoods on the basis of Frequency of venues nearby. 7 clusters were obtained, each having a unique collection of Venues. Since, we were focused on finding heuristic relocation location for one neighbourhood to other we found many probable neighbourhood in Toronto which appeals the same as native neighbourhood in Delhi.

The neighbourhoods recommendation obtained here are not completely accurate. This is due to the limitations in the dataset used in the project. Due to lack of cross referencing sources, we may have missed a few neighbourhoods from our consideration. The foursquare API does not contain, or does not rely, a comprehensive dataset about the restaurants present in Delhi. Surely, in a city like Delhi with a population of over 19 million, there are much more criteria which should have been taken into account. Also in this study no cultural, and social background of particular parties were taken into account thus making this study over-generalised.

CONCLUSION:

This project helps a person get a better understanding of the neighbourhoods with respect to the most common venues in that neighbourhood. It is always helpful to make use of technology to stay one step ahead i.e. finding out more about places before moving into a neighbourhood. The future of this project includes taking other factors such as cost of living in the areas into consideration to shortlist the borough, such as filtering areas based on a predefined budget. Thus more personalisation according to each migrant can be easily approached.

Also, the resulted neighbourhood recommendation in Toronto can act as a good starting point for figuring out which location might suit best to a migrant. Each person then, by his personal preferences can further refine the study and take benefits of initial iterations provided in this Project.