Coursera - IBM Data Science

Applied Data Science Capstone

Capstone Project - The Battle of Neighborhoods

Final Project (Week 1)

by:

Robson Sampaio

RIO DE JANEIRO, BRASIL

15/06/2020

SUMÁRIO

PROBLEM & BACKGROUND

The City of New York, is the most populous city in the United States, one of the greatest metropolises over the world, is a dream place for gourmet to seek delicious cuisine. It is diverse and is the financial capital of USA. It is multicultural. It provides lot of business oppourtunities and business friendly environment. It has attracted many different players into the market. It is a global hub of business and commerce. The city is a major center for banking and finance, retailing, world trade, transportation, tourism, real estate, new media, traditional media, advertising, legal services, accountancy, insurance, theater, fashion, and the arts in the United States. This also means that the market is highly competitive.

There are hundreds of restaurants and I chose one to carry out this analysis and build this report, but for each type of restaurant, the same line of analysis can be used.

Data Preparation

Neighborhoods in New York City (Wikipedia) - I cleaned the data and reduced it to boroughs of NYC so that I can use it to find geological locations for further venue analysis.

For this problem, we will obtain the services of the Foursquare API to explore the data of cities, in terms of their neighborhoods. The data also includes information about places around each neighborhood, such as restaurants, hotels, coffee shops, parks, theaters, art galleries, museums and more. We selected a municipality in each city to analyze its neighborhoods. New York's Manhattan and Toronto's Downtown Toronto. We will use the machine learning technique, “Clustering” to segment neighborhoods with similar objects based on data from each neighborhood. These objects will have priority based on pedestrian traffic (activity) in their respective neighborhoods. This will help to locate areas and tourist centers, and then we can judge the similarity or difference between two cities on that basis.

Os dados recuperados do cotem informações dos locais, especificado por longitude e latitude. As informações sobre os locais continham a distância de aproximadamente 100m. O dataset tinha os seguintes atributos: Neighborhood,

Neighborhood Latitude, Neighborhood Longitude, Venue, Name of the venue e.g. the name of a store or restaurant, Venue Latitude, Venue Longitude, Venue Category.

Methodology

Categorized the methodologysection into two parts:

Exploratory Data Analysis, visualise the crime repots in different Vancouver boroughs to idenity the safest borough and normalise the neighborhoods of that borough. We will Use the resulting data and find 10 most common venues in each neighborhood.

Modelling, to help stakeholders choose the right neighborhood within a borough we will be clustering similar neighborhoods using K - means clustering which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. We will use K-Means clustering to address this problem so as to group data based on existing venues which will help in the decision making process.

Exploration

To compare similarities between two cities, I explored neighborhoods, segmented and grouped to find related neighborhoods in a large city like New York and Toronto. To do this, we need to group data that is a form of unsupervised machine learning: k-means clustering algorithm.

Download data

```
url = "https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"
source = requests.get(url).text
soup = BeautifulSoup(source, 'lxml')
table = soup.find('table',{'class':'wikitable sortable'})
```

Listing data in pandas dataframe

```
df = df[~df['Borough'].isnull()]  # to filter out bad rows
df.drop(df[df.Borough == 'Not assigned'].index, inplace=True)
df.reset_index(drop=True, inplace=True)
df = df.groupby(['PostalCode','Borough'])['Neighborhood'].apply(lambda x: ','.join(x)).reset_index()
df['Neighborhood'].replace('Not assigned',df['Borough'],inplace=True)
df
```

	PostalCode	Borough	Neighborhood
0	M1B	Scarborough	Malvern, Rouge
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae
...	...	...	...

Geo Coordinates

```
df_geo_coordinate = pd.read_csv('http://cocl.us/Geospatial_data')
df_geo_coordinate.head()
```

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

```
df_geo_coordinate.shape
```

```
(103, 3)
```

```
df_geo_coordinate.head()
```

	Postal Code	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

Postal Code

```
df_toronto = pd.merge(df, df_geo_coordinate, how='left', left_on = 'PostalCode', right_on = 'Postal Code')
# remove the "Postal Code" column
df_toronto.drop("Postal Code", axis=1, inplace=True)
df_toronto.head()
```

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.806686	-79.194353
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

```
#Find null values
df_toronto.isnull().sum()
```

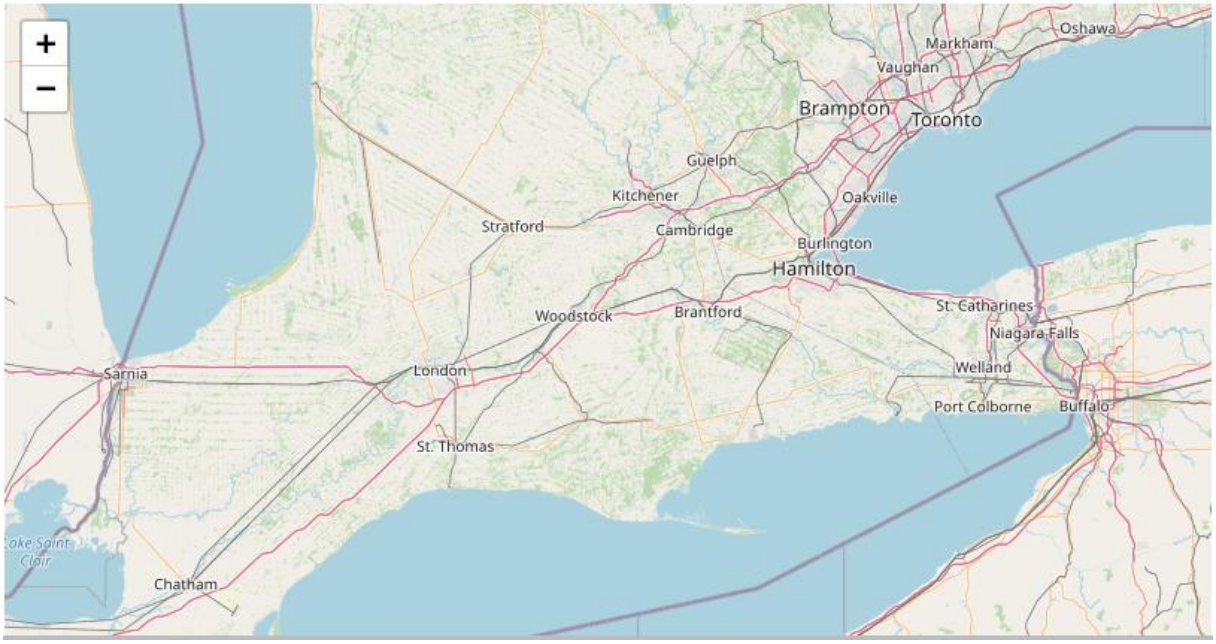
```
PostalCode      0
Borough         0
Neighborhood    0
Latitude        0
Longitude       0
dtype: int64
```

```
address = "Toronto, ON"

geolocator = Nominatim(user_agent="toronto_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of Toronto city are {}, {}.'.format(latitude, longitude))
```

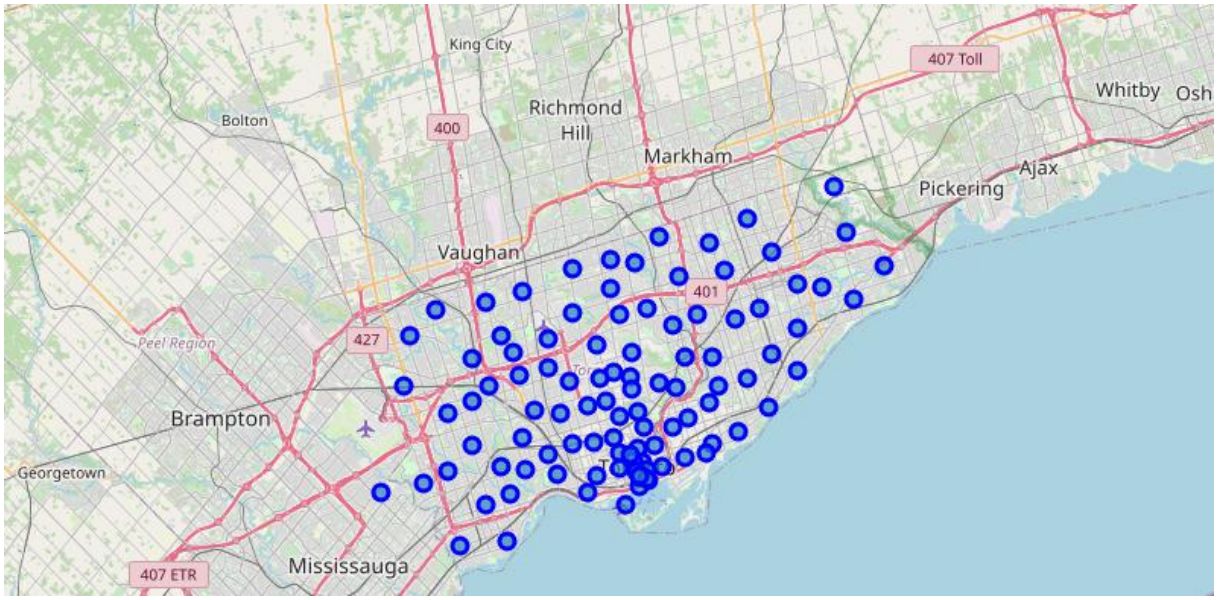
```
The geograpical coordinate of Toronto city are 43.6534817, -79.3839347.
```

```
map_toronto = folium.Map(location=[latitude, longitude], zoom_start=10)
map_toronto
```




```
for lat, lng, borough, neighborhood in zip(
        df_toronto['Latitude'],
        df_toronto['Longitude'],
        df_toronto['Borough'],
        df_toronto['Neighborhood']):
    label = '{}, {}'.format(neighborhood, borough)
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='blue',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_toronto)

map_toronto
```



Results

The objective of the business problem was to help stakeholders identify one of the safest borough in Vancouver, and an appropriate neighborhood within the borough to set up a commercial establishment especially a Grocery store. This has been achieved by first making use of Vancouver crime data to identify a safe borugh with considerable number of neighborhood for any business to be viable. After selecting the borough it was imperative to choose the right neighborhood where grocery shops were not among venues in a close proximity to each other. We achieved this by grouping the neighborhoods into clusters to assist the stakeholders by providing them with relavent data about venues and safety of a given neighborhood.

Conclusion

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's argument is present in both neighborhoods. The dissimilarity exists in terms of some different venues and facilities but not on a larger extent.

REFERÊNCIAS

Wikepedia. List of postal codes of Canada. Recuperado de https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.