The Battle of Neighbourhoods

Coursera IBM ML final project

Introduction

The problem we will solve is:

which of the neighbourhoods of Toronto are more suitable to open there a restaurant?

Introduction

To solve this problem we will

1. find out by FourSquare API how many restaurants are within 1000 meters from each of Toronto Neighbourhoods
2. divide all Toronto neighbourhoods into three groups by K-means clustering:

- Neighbourhoods with a few restaurants. These neighbourhoods will be the best choice to open there a new restaurant.
- Neighbourhoods with a medium number of restaurants. These neighbourhoods will be also not so bad choice to open there a new restaurant, but to decide we need some additional consideration.
- Neighbourhoods with a lot of restaurants. They are not recommended to open there a new restaurant.

Data

1. We scrape the following Wikipedia page, https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M, in order to obtain the list of all Toronto neighbourhoods, as well as their postal codes and boroughs
2. For each neighbourhood we add its latitude and longitude by the link: http://cocl.us/Geospatial_data
3. The final dataframe contains 103 rows and 5 columns and looks like

	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	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

Methodology

1. First we import BeautifulSoup and use it to scrape the table from Wikipedia
2. Then we will refine and modify the dataframe df in the following way.

- Rename the columns as 'Postcode', 'Borough', and 'Neighbourhood'
- Drop the rows, where Borough is "Not assigned"
- If Neighbourhood is "Not assigned", then assign its value the same as for Borough
- Neighbourhoods with the same Postcode are joined and separated by comma
- Download the coordinates for each Neighbourhood group from by the link: http://cocl.us/Geospatial_data and add them to two new columns of df “latitude” and “longitude”

Methodology

3. Now we use FourSquare API. For each group of neighbourhoods in df we will be interested in number of restaurants within 1000 meters from the geographical position of that group . So, first put radius = 1000 and search_query='restaurant'

4. Now add to df a column 'sum', where there will be the numbers of restaurants within 1000 meters from the neighbourhoods. The python code looks like:

```
for k, row in df.iterrows():
    (lat,lng)= (row["Latitude"], row["Longitude"])
    url = 'https://api.foursquare.com/v2/venues/search?client_id={}&client_secret={}&ll={},\
    {}&v={}&query={}&radius={}&limit={}'\
    .format(CLIENT_ID, CLIENT_SECRET, lat, lng, VERSION, search_query, radius, LIMIT)
    results = requests.get(url).json()
    df.set_value(index=k,col='sum',value=len(results['response']['venues']))
```

Methodology

5. Now we chose the K-means clustering method with the number of clusters is equal to 3. And we will clustering the list of 103 groups of Toronto neighbourhoods by three parameters: latitude, longitude and sum (which equals to the number of restaurants within 1000 meters from the neighbourhood).

Our hypothesis is that 3 clusters may be interpreted as:

- Neighbourhoods with a few restaurants. These neighbourhoods will be the best choice to open there a new restaurant.
- Neighbourhoods with a medium number of restaurants. These neighbourhoods will be also not so bad choice to open there a new restaurant.
- Neighbourhoods with a lot of restaurants. They are not recommended to open there a new restaurant.

Results

1. The clusters are:

```
array([0, 0, 0, 0, 2, 2, 0, 0, 0, 0, 0, 2, 1, 2, 2, 0, 0, 0, 2, 0, 0, 2,
       1, 0, 0, 0, 0, 0, 0, 2, 2, 0, 0, 0, 0, 0, 0, 2, 0, 2, 1, 1, 2, 2,
       0, 1, 2, 1, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 0, 2, 1,
       1, 1, 0, 1, 1, 0, 0, 2, 2, 1, 2, 1, 2, 0, 0, 0, 0, 2, 2, 1, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0], dtype=int32)
```

Cluster "0": 0-6 restaurants
Cluster "2": 7-16 restaurants
Cluster "1": 17-30 restaurants

2. So we recommend to open a new restaurant in the Neighbourhoods of the cluster "0", we recommend to consider the neighbourhoods of the cluster "2" as the possible variant to open there a new restaurant, and we do not recommend to open a new restaurant in the neighbourhoods of the cluster "1"

Discussion

Sure the number of restaurants within 1000 meters from a neighbourhood is not the unique factor to decide is it a good idea to open there a new restaurant or not.

And our clustering may be considered just as the first step for more detailed analysis based on a lot of other factors.

But this more detailed analysis is not the subject of our consideration now.

Conclusion

In this research we have clustered all Toronto neighbourhoods by three groups. Which are interpreted as:

- Strongly recommended for launching there a new restaurant (cluster "0")
- Recommended for launching there a new restaurant (cluster "2")
- Not recommended for launching there a new restaurant (cluster "1")