Segmenting-and-Clustering-Neighbourhoods-in-Toronto 1 118

November 29, 2020

0.1 Segmenting and Clustering Neighbourhoods in Toronto

The project includes scraping the Wikipedia page for the postal codes of Canada and then process and clean the data for the clustering. The clustering is carried out by K Means and the clusters are plotted using the Folium Library. The Boroughs containing the name 'Toronto' in it are first plotted and then clustered and plotted again.

0.2 All the 3 tasks of web scraping, cleaning and clustering are implemented in the same notebook for the ease of evaluation.

0.3 Installing and Importing the required Libraries

[3]:
```
!pip install beautifulsoup4
!pip install lxml
import requests # library to handle requests
import pandas as pd # library for data analsysis
import numpy as np # library to handle data in a vectorized manner
import random # library for random number generation

#!conda install -c conda-forge geopy --yes
from geopy.geocoders import Nominatim # module to convert an address into␣
 ↪latitude and longitude values

# libraries for displaying images
from IPython.display import Image
from IPython.core.display import HTML



from IPython.display import display_html
import pandas as pd
import numpy as np

# tranforming json file into a pandas dataframe library
from pandas.io.json import json_normalize

!conda install -c conda-forge folium=0.5.0 --yes
import folium # plotting library
from bs4 import BeautifulSoup
from sklearn.cluster import KMeans
```

```
import matplotlib.cm as cm
import matplotlib.colors as colors

print('Folium installed')
print('Libraries imported.')
```

```
Collecting beautifulsoup4
  Downloading https://files.pythonhosted.org/packages/d1/41/e6495bd7d3781c
ee623ce23ea6ac73282a373088fcd0ddc809a047b18eae/beautifulsoup4-4.9.3-py3-none-
any.whl (115kB)
     |                  | 122kB 7.3MB/s eta 0:00:01
Collecting soupsieve>1.2; python_version >= "3.0" (from beautifulsoup4)
  Downloading https://files.pythonhosted.org/packages/6f/8f/457f4a5390eeae1cc3ae
ab89deb7724c965be841ffca6cfca9197482e470/soupsieve-2.0.1-py3-none-any.whl
Installing collected packages: soupsieve, beautifulsoup4
Successfully installed beautifulsoup4-4.9.3 soupsieve-2.0.1
Collecting lxml
  Downloading https://files.pythonhosted.org/packages/bd/78/56a7c88a57d0d1
4945472535d0df9fb4bbad7d34ede658ec7961635c790e/lxml-4.6.2-cp36-cp36m-manylinux1_
x86_64.whl (5.5MB)
     |                  | 5.5MB 6.5MB/s eta 0:00:01
Installing collected packages: lxml
Successfully installed lxml-4.6.2
Collecting package metadata (current_repodata.json): done
Solving environment: failed with initial frozen solve. Retrying with flexible
solve.
Collecting package metadata (repodata.json): - ^C
| Folium installed
Libraries imported.
```

0.4 Scraping the Wikipedia page for the table of postal codes of Canada

BeautifulSoup Library of Python is used for web scraping of table from the Wikipedia. The title of the webpage is printed to check if the page has been scraped successfully or not. Then the table of postal codes of Canada is printed.

[43]:
```
source = requests.get('https://en.wikipedia.org/wiki/
 ↪List_of_postal_codes_of_Canada:_M').text
soup=BeautifulSoup(source,'lxml')
print(soup.title)
from IPython.display import display_html
tab = str(soup.table)
display_html(tab,raw=True)
```

```
<title>List of postal codes of Canada: M - Wikipedia</title>
```

0.5 The html table is converted to Pandas DataFrame for cleaning and preprocessing.

[45]:
```
dfs = pd.read_html(tab)
df=dfs[0]
df = df.rename(columns={'Postal Code': 'Postcode'})
df.head()
```

[45]:

	Postcode	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront

0.6 Data preprocessing and cleaning

[47]:
```
# Dropping the rows where Borough is 'Not assigned'
df1 = df[df.Borough != 'Not assigned']

# Combining the neighbourhoods with same Postalcode
df2 = df1.groupby(['Postcode','Borough'], sort=False).agg(', '.join)
df2.reset_index(inplace=True)

# Replacing the name of the neighbourhoods which are 'Not assigned' with names␣
 ↪of Borough
df2['Neighbourhood'] = np.where(df2['Neighbourhood'] == 'Not␣
 ↪assigned',df2['Borough'], df2['Neighbourhood'])

df2
```

[47]:

	Postcode	Borough	\
0	M3A	North York	
1	M4A	North York	
2	M5A	Downtown Toronto	
3	M6A	North York	
4	M7A	Downtown Toronto	
..	…	…	
98	M8X	Etobicoke	
99	M4Y	Downtown Toronto	
100	M7Y	East Toronto	
101	M8Y	Etobicoke	
102	M8Z	Etobicoke	

	Neighbourhood
0	Parkwoods
1	Victoria Village

```
2                          Regent Park, Harbourfront
3                    Lawrence Manor, Lawrence Heights
4          Queen's Park, Ontario Provincial Government
..                                                 …
98       The Kingsway, Montgomery Road, Old Mill North
99                                 Church and Wellesley
100  Business reply mail Processing Centre, South C…
101  Old Mill South, King's Mill Park, Sunnylea, Hu…
102  Mimico NW, The Queensway West, South of Bloor,…

[103 rows x 3 columns]
```

[48]:
```
# Shape of data frame
df2.shape
```

[48]:
```
(103, 3)
```

0.7 Importing the csv file conatining the latitudes and longitudes for various neighbourhoods in Canada

[49]:
```
lat_lon = pd.read_csv('https://cocl.us/Geospatial_data')
lat_lon.head()
```

[49]:
```
  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
```

0.8 Merging the two tables for getting the Latitudes and Longitudes for various neighbourhoods in Canada

[51]:
```
lat_lon.rename(columns={'Postal Code':'Postcode'},inplace=True)
df3 = pd.merge(df2,lat_lon,on='Postcode')
df3.head()
```

[51]:
```
  Postcode           Borough                                Neighbourhood  \
0      M3A        North York                                   Parkwoods
1      M4A        North York                             Victoria Village
2      M5A  Downtown Toronto                Regent Park, Harbourfront
3      M6A        North York          Lawrence Manor, Lawrence Heights
4      M7A  Downtown Toronto  Queen's Park, Ontario Provincial Government

    Latitude  Longitude
0  43.753259 -79.329656
```

```
1   43.725882 -79.315572
2   43.654260 -79.360636
3   43.718518 -79.464763
4   43.662301 -79.389494
```

0.9 The notebook from here includes the Clustering and the plotting of the neighbourhoods of Canada which contain Toronto in their Borough ## Getting all the rows from the data frame which contains Toronto in their Borough.

[52]:
```
df4 = df3[df3['Borough'].str.contains('Toronto',regex=False)]
df4
```

[52]:

	Postcode	Borough
2	M5A	Downtown Toronto
4	M7A	Downtown Toronto
9	M5B	Downtown Toronto
15	M5C	Downtown Toronto
19	M4E	East Toronto
20	M5E	Downtown Toronto
24	M5G	Downtown Toronto
25	M6G	Downtown Toronto
30	M5H	Downtown Toronto
31	M6H	West Toronto
36	M5J	Downtown Toronto
37	M6J	West Toronto
41	M4K	East Toronto
42	M5K	Downtown Toronto
43	M6K	West Toronto
47	M4L	East Toronto
48	M5L	Downtown Toronto
54	M4M	East Toronto
61	M4N	Central Toronto
62	M5N	Central Toronto
67	M4P	Central Toronto
68	M5P	Central Toronto
69	M6P	West Toronto
73	M4R	Central Toronto
74	M5R	Central Toronto
75	M6R	West Toronto
79	M4S	Central Toronto
80	M5S	Downtown Toronto
81	M6S	West Toronto
83	M4T	Central Toronto
84	M5T	Downtown Toronto
86	M4V	Central Toronto

```
87       M5V  Downtown Toronto
91       M4W  Downtown Toronto
92       M5W  Downtown Toronto
96       M4X  Downtown Toronto
97       M5X  Downtown Toronto
99       M4Y  Downtown Toronto
100      M7Y      East Toronto
```

	Neighbourhood	Latitude	Longitude
2	Regent Park, Harbourfront	43.654260	-79.360636
4	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
9	Garden District, Ryerson	43.657162	-79.378937
15	St. James Town	43.651494	-79.375418
19	The Beaches	43.676357	-79.293031
20	Berczy Park	43.644771	-79.373306
24	Central Bay Street	43.657952	-79.387383
25	Christie	43.669542	-79.422564
30	Richmond, Adelaide, King	43.650571	-79.384568
31	Dufferin, Dovercourt Village	43.669005	-79.442259
36	Harbourfront East, Union Station, Toronto Islands	43.640816	-79.381752
37	Little Portugal, Trinity	43.647927	-79.419750
41	The Danforth West, Riverdale	43.679557	-79.352188
42	Toronto Dominion Centre, Design Exchange	43.647177	-79.381576
43	Brockton, Parkdale Village, Exhibition Place	43.636847	-79.428191
47	India Bazaar, The Beaches West	43.668999	-79.315572
48	Commerce Court, Victoria Hotel	43.648198	-79.379817
54	Studio District	43.659526	-79.340923
61	Lawrence Park	43.728020	-79.388790
62	Roselawn	43.711695	-79.416936
67	Davisville North	43.712751	-79.390197
68	Forest Hill North & West, Forest Hill Road Park	43.696948	-79.411307
69	High Park, The Junction South	43.661608	-79.464763
73	North Toronto West, Lawrence Park	43.715383	-79.405678
74	The Annex, North Midtown, Yorkville	43.672710	-79.405678
75	Parkdale, Roncesvalles	43.648960	-79.456325
79	Davisville	43.704324	-79.388790
80	University of Toronto, Harbord	43.662696	-79.400049
81	Runnymede, Swansea	43.651571	-79.484450
83	Moore Park, Summerhill East	43.689574	-79.383160
84	Kensington Market, Chinatown, Grange Park	43.653206	-79.400049
86	Summerhill West, Rathnelly, South Hill, Forest…	43.686412	-79.400049
87	CN Tower, King and Spadina, Railway Lands, Har…	43.628947	-79.394420
91	Rosedale	43.679563	-79.377529
92	Stn A PO Boxes	43.646435	-79.374846
96	St. James Town, Cabbagetown	43.667967	-79.367675
97	First Canadian Place, Underground city	43.648429	-79.382280
99	Church and Wellesley	43.665860	-79.383160

```
100  Business reply mail Processing Centre, South C…  43.662744 -79.321558
```

0.10 Visualizing all the Neighbourhoods of the above data frame using Folium

[53]:
```
map_toronto = folium.Map(location=[43.651070,-79.347015],zoom_start=10)

for lat,lng,borough,neighbourhood in␣
 ↪zip(df4['Latitude'],df4['Longitude'],df4['Borough'],df4['Neighbourhood']):
    label = '{}, {}'.format(neighbourhood, 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
```

[53]:
```
<folium.folium.Map at 0x7fa6bfdef550>
```

0.11 Using KMeans clustering for the clsutering of the neighbourhoods

[57]:
```
k=5
toronto_clustering = df4.drop(['Postcode','Borough','Neighbourhood'],1)
kmeans = KMeans(n_clusters = k,random_state=0).fit(toronto_clustering)
kmeans.labels_
df4.insert(0, 'ClusterLabels', kmeans.labels_)
```

[58]:
```
df4
```

[58]:

	ClusterLabels	Cluster Labels	Postcode	Borough	\
2	0	0	M5A	Downtown Toronto	
4	0	0	M7A	Downtown Toronto	
9	0	0	M5B	Downtown Toronto	
15	0	0	M5C	Downtown Toronto	
19	2	4	M4E	East Toronto	
20	0	0	M5E	Downtown Toronto	
24	0	0	M5G	Downtown Toronto	
25	3	3	M6G	Downtown Toronto	
30	0	0	M5H	Downtown Toronto	
31	4	1	M6H	West Toronto	
36	0	0	M5J	Downtown Toronto	
37	3	3	M6J	West Toronto	

41	2	4	M4K	East Toronto
42	0	0	M5K	Downtown Toronto
43	3	3	M6K	West Toronto
47	2	4	M4L	East Toronto
48	0	0	M5L	Downtown Toronto
54	2	4	M4M	East Toronto
61	1	2	M4N	Central Toronto
62	1	2	M5N	Central Toronto
67	1	2	M4P	Central Toronto
68	1	2	M5P	Central Toronto
69	4	1	M6P	West Toronto
73	1	2	M4R	Central Toronto
74	3	3	M5R	Central Toronto
75	4	1	M6R	West Toronto
79	1	2	M4S	Central Toronto
80	3	3	M5S	Downtown Toronto
81	4	1	M6S	West Toronto
83	1	2	M4T	Central Toronto
84	3	3	M5T	Downtown Toronto
86	1	2	M4V	Central Toronto
87	0	0	M5V	Downtown Toronto
91	0	0	M4W	Downtown Toronto
92	0	0	M5W	Downtown Toronto
96	0	0	M4X	Downtown Toronto
97	0	0	M5X	Downtown Toronto
99	0	0	M4Y	Downtown Toronto
100	2	4	M7Y	East Toronto

	Neighbourhood	Latitude	Longitude
2	Regent Park, Harbourfront	43.654260	-79.360636
4	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
9	Garden District, Ryerson	43.657162	-79.378937
15	St. James Town	43.651494	-79.375418
19	The Beaches	43.676357	-79.293031
20	Berczy Park	43.644771	-79.373306
24	Central Bay Street	43.657952	-79.387383
25	Christie	43.669542	-79.422564
30	Richmond, Adelaide, King	43.650571	-79.384568
31	Dufferin, Dovercourt Village	43.669005	-79.442259
36	Harbourfront East, Union Station, Toronto Islands	43.640816	-79.381752
37	Little Portugal, Trinity	43.647927	-79.419750
41	The Danforth West, Riverdale	43.679557	-79.352188
42	Toronto Dominion Centre, Design Exchange	43.647177	-79.381576
43	Brockton, Parkdale Village, Exhibition Place	43.636847	-79.428191
47	India Bazaar, The Beaches West	43.668999	-79.315572
48	Commerce Court, Victoria Hotel	43.648198	-79.379817
54	Studio District	43.659526	-79.340923

```
61                                    Lawrence Park  43.728020 -79.388790
62                                         Roselawn  43.711695 -79.416936
67                                  Davisville North  43.712751 -79.390197
68     Forest Hill North & West, Forest Hill Road Park  43.696948 -79.411307
69                       High Park, The Junction South  43.661608 -79.464763
73                  North Toronto West, Lawrence Park  43.715383 -79.405678
74                The Annex, North Midtown, Yorkville  43.672710 -79.405678
75                           Parkdale, Roncesvalles  43.648960 -79.456325
79                                        Davisville  43.704324 -79.388790
80                    University of Toronto, Harbord  43.662696 -79.400049
81                               Runnymede, Swansea  43.651571 -79.484450
83                          Moore Park, Summerhill East  43.689574 -79.383160
84          Kensington Market, Chinatown, Grange Park  43.653206 -79.400049
86   Summerhill West, Rathnelly, South Hill, Forest…  43.686412 -79.400049
87   CN Tower, King and Spadina, Railway Lands, Har…  43.628947 -79.394420
91                                        Rosedale  43.679563 -79.377529
92                                  Stn A PO Boxes  43.646435 -79.374846
96                          St. James Town, Cabbagetown  43.667967 -79.367675
97              First Canadian Place, Underground city  43.648429 -79.382280
99                               Church and Wellesley  43.665860 -79.383160
100  Business reply mail Processing Centre, South C…  43.662744 -79.321558
```

[59]:
```
# create map
map_clusters = folium.Map(location=[43.651070,-79.347015],zoom_start=10)

# set color scheme for the clusters
x = np.arange(k)
ys = [i + x + (i*x)**2 for i in range(k)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lon, neighbourhood, cluster in zip(df4['Latitude'], df4['Longitude'],␣
 ↪df4['Neighbourhood'], df4['Cluster Labels']):
    label = folium.Popup(' Cluster ' + str(cluster), parse_html=True)
    folium.CircleMarker(
        [lat, lon],
        radius=5,
        popup=label,
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.7).add_to(map_clusters)

map_clusters
```

[59]: <folium.folium.Map at 0x7fa6bfeca208>