Untitled7

May 23, 2020

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 Brackests containing the name 'Toronto' in it are first plotted and then clustered and plotted again.

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

Installing and Importing the required Libraries

[4]:
```
!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.')
```

```
Requirement already satisfied: beautifulsoup4 in
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (4.9.1)
Requirement already satisfied: soupsieve>1.2 in
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (from
beautifulsoup4) (2.0.1)
Requirement already satisfied: lxml in
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (4.5.1)
Collecting package metadata (current_repodata.json): done
Solving environment: done

## Package Plan ##

  environment location: /home/jupyterlab/conda/envs/python

  added / updated specs:
    - geopy


The following packages will be downloaded:

    package                    |            build
    ---------------------------|-----------------
    geographiclib-1.50         |             py_0          34 KB  conda-forge
    geopy-1.22.0               |     pyh9f0ad1d_0          63 KB  conda-forge
    ------------------------------------------------------------
                                           Total:          97 KB

The following NEW packages will be INSTALLED:

  geographiclib      conda-forge/noarch::geographiclib-1.50-py_0
  geopy              conda-forge/noarch::geopy-1.22.0-pyh9f0ad1d_0



Downloading and Extracting Packages
geopy-1.22.0          | 63 KB     | ##################################### | 100%
geographiclib-1.50    | 34 KB     | ##################################### | 100%
Preparing transaction: done
Verifying transaction: done
Executing transaction: done
Collecting package metadata (current_repodata.json): done
```

```
Solving environment: failed with initial frozen solve. Retrying with flexible
solve.
Collecting package metadata (repodata.json): done
Solving environment: done

## Package Plan ##

  environment location: /home/jupyterlab/conda/envs/python

  added / updated specs:
    - folium=0.5.0


The following packages will be downloaded:

    package                    |            build
    ---------------------------|-----------------
    altair-4.1.0               |             py_1         614 KB  conda-forge
    branca-0.4.1               |             py_0          26 KB  conda-forge
    brotlipy-0.7.0             |py36h8c4c3a4_1000         346 KB  conda-forge
    chardet-3.0.4              |py36h9f0ad1d_1006         188 KB  conda-forge
    cryptography-2.9.2         |   py36h45558ae_0         613 KB  conda-forge
    folium-0.5.0               |             py_0          45 KB  conda-forge
    pandas-1.0.3               |   py36h830a2c2_1        11.1 MB  conda-forge
    pysocks-1.7.1              |   py36h9f0ad1d_1          27 KB  conda-forge
    toolz-0.10.0               |             py_0          46 KB  conda-forge
    vincent-0.4.4              |             py_1          28 KB  conda-forge
    ------------------------------------------------------------
                                           Total:        13.0 MB

The following NEW packages will be INSTALLED:

  altair             conda-forge/noarch::altair-4.1.0-py_1
  attrs              conda-forge/noarch::attrs-19.3.0-py_0
  branca             conda-forge/noarch::branca-0.4.1-py_0
  brotlipy           conda-forge/linux-64::brotlipy-0.7.0-py36h8c4c3a4_1000
  chardet            conda-forge/linux-64::chardet-3.0.4-py36h9f0ad1d_1006
  cryptography       conda-forge/linux-64::cryptography-2.9.2-py36h45558ae_0
  entrypoints        conda-forge/linux-64::entrypoints-0.3-py36h9f0ad1d_1001
  folium             conda-forge/noarch::folium-0.5.0-py_0
  idna               conda-forge/noarch::idna-2.9-py_1
  importlib_metadata conda-forge/noarch::importlib_metadata-1.6.0-0
  jinja2             conda-forge/noarch::jinja2-2.11.2-pyh9f0ad1d_0
  jsonschema         conda-forge/linux-64::jsonschema-3.2.0-py36h9f0ad1d_1
  markupsafe         conda-forge/linux-64::markupsafe-1.1.1-py36h8c4c3a4_1
  pandas             conda-forge/linux-64::pandas-1.0.3-py36h830a2c2_1
  pyopenssl          conda-forge/noarch::pyopenssl-19.1.0-py_1
  pyrsistent         conda-forge/linux-64::pyrsistent-0.16.0-py36h8c4c3a4_0
```

```
  pysocks            conda-forge/linux-64::pysocks-1.7.1-py36h9f0ad1d_1
  pytz               conda-forge/noarch::pytz-2020.1-pyh9f0ad1d_0
  requests           conda-forge/noarch::requests-2.23.0-pyh8c360ce_2
  toolz              conda-forge/noarch::toolz-0.10.0-py_0
  urllib3            conda-forge/noarch::urllib3-1.25.9-py_0
  vincent            conda-forge/noarch::vincent-0.4.4-py_1


Downloading and Extracting Packages
pysocks-1.7.1        | 27 KB     | ##################################### | 100%
toolz-0.10.0         | 46 KB     | ##################################### | 100%
chardet-3.0.4        | 188 KB    | ##################################### | 100%
folium-0.5.0         | 45 KB     | ##################################### | 100%
branca-0.4.1         | 26 KB     | ##################################### | 100%
cryptography-2.9.2   | 613 KB    | ##################################### | 100%
brotlipy-0.7.0       | 346 KB    | ##################################### | 100%
pandas-1.0.3         | 11.1 MB   | ##################################### | 100%
altair-4.1.0         | 614 KB    | ##################################### | 100%
vincent-0.4.4        | 28 KB     | ##################################### | 100%
Preparing transaction: done
Verifying transaction: done
Executing transaction: done
Folium installed
Libraries imported.
```

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

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.

[5]:
```
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>
```

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

[6]:
```
dfs = pd.read_html(tab)
df=dfs[0]
df.head()
```

[6]:

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

Data preprocessing and cleaning

[36]:
```
# Dropping the rows where Borough is 'Not assigned'
df1=df[df.Borough != 'Not assigned']

# Combining the neighbourhoods with same Postalcode
df2=df1.groupby(['Postal Code','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['Neighborhood']=np.where(df2['Neighborhood']=='Not␣
 ↪assigned',df2['Borough'], df2['Neighborhood'])
df2
```

[36]:

	Postal Code	Borough	Neighborhood
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
..	…	…	…
98	M8X	Etobicoke	The Kingsway, Montgomery Road, Old Mill North
99	M4Y	Downtown Toronto	Church and Wellesley
100	M7Y	East Toronto	Business reply mail Processing Centre
101	M8Y	Etobicoke	Old Mill South, King's Mill Park, Sunnylea, Hu…
102	M8Z	Etobicoke	Mimico NW, The Queensway West, South of Bloor,…

```
[103 rows x 3 columns]
```

[13]:
```
# Shape of data frame
df2.shape
```

[13]:
```
(103, 3)
```

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

[38]:
```
lat_lon = pd.read_csv('https://cocl.us/Geospatial_data')
lat_lon.head()
```

[38]:

	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

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

[40]:
```
lat_lon.rename(columns={'Postal Code':'Postal Code'},inplace=True)
df3 = pd.merge(df2,lat_lon,on='Postal Code')
df3.head()
```

[40]:

	Postal Code	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

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.

[41]:
```
df4 = df3[df3['Borough'].str.contains('Toronto',regex=False)]
df4
```

[41]:

	Postal Code	Borough	Neighborhood	Latitude	Longitude
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494
9	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937
15	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418
19	M4E	East Toronto	The Beaches	43.676357	-79.293031
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			

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	43.696948 -79.411307
69		High Park, The Junction South	43.661608 -79.464763
73		North Toronto West	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	43.662744 -79.321558

Visualizing all the Neighbourhoods of the above data frame using Folium

[43]:
```
map_toronto = folium.Map(location=[43.651070,-79.347015],zoom_start=10)

for lat,lng,borough,neighborhood in␣
 ↪zip(df4['Latitude'],df4['Longitude'],df4['Borough'],df4['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
```

[43]: <folium.folium.Map at 0x7fa9a22cfac8>

Using KMeans clustering for the clsutering of the neighbourhoods

[48]:
```
k=5
toronto_clustering = df4.drop(['Postal Code','Borough','Neighborhood'],1)
kmeans = KMeans(n_clusters = k,random_state=0).fit(toronto_clustering)
kmeans.labels_
df4
```

[48]:

	Cluster Labels	Postal Code	Borough	\
2	0	M5A	Downtown Toronto	
4	0	M7A	Downtown Toronto	
9	0	M5B	Downtown Toronto	
15	0	M5C	Downtown Toronto	
19	4	M4E	East Toronto	
20	0	M5E	Downtown Toronto	
24	0	M5G	Downtown Toronto	
25	3	M6G	Downtown Toronto	
30	0	M5H	Downtown Toronto	
31	1	M6H	West Toronto	
36	0	M5J	Downtown Toronto	
37	3	M6J	West Toronto	
41	4	M4K	East Toronto	
42	0	M5K	Downtown Toronto	
43	3	M6K	West Toronto	
47	4	M4L	East Toronto	
48	0	M5L	Downtown Toronto	
54	4	M4M	East Toronto	
61	2	M4N	Central Toronto	
62	2	M5N	Central Toronto	
67	2	M4P	Central Toronto	
68	2	M5P	Central Toronto	
69	1	M6P	West Toronto	
73	2	M4R	Central Toronto	
74	3	M5R	Central Toronto	
75	1	M6R	West Toronto	
79	2	M4S	Central Toronto	
80	3	M5S	Downtown Toronto	
81	1	M6S	West Toronto	

```
83                2         M4T   Central Toronto
84                3         M5T  Downtown Toronto
86                2         M4V   Central Toronto
87                0         M5V  Downtown Toronto
91                0         M4W  Downtown Toronto
92                0         M5W  Downtown Toronto
96                0         M4X  Downtown Toronto
97                0         M5X  Downtown Toronto
99                0         M4Y  Downtown Toronto
100               4         M7Y      East Toronto
```

```
                                          Neighborhood   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  43.696948 -79.411307
69                      High Park, The Junction South  43.661608 -79.464763
73                                North Toronto West  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  43.662744 -79.321558
```

[50]:
```
# 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, neighborhood, cluster in zip(df4['Latitude'], df4['Longitude'],␣
 ↪df4['Neighborhood'], 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
```

[50]:
```
<folium.folium.Map at 0x7fa9a2238e48>
```

[ ]:
```
```