Untitled

April 29, 2021

This is a Data Science Capstone Project in Coursera

(for the IBM Data Science Professional Certificate)

[1]:
```
import pandas as pd
import numpy as np
import json
import requests
!pip install bs4
from bs4 import BeautifulSoup
```

```
Collecting bs4
  Downloading https://files.pythonhosted.org/packages/10/ed/7e8b97591f6f45617413
9ec089c769f89a94a1a4025fe967691de971f314/bs4-0.0.1.tar.gz
Collecting beautifulsoup4 (from bs4)
  Downloading https://files.pythonhosted.org/packages/d1/41/e6495bd7d3781c
ee623ce23ea6ac73282a373088fcd0ddc809a047b18eae/beautifulsoup4-4.9.3-py3-none-
any.whl (115kB)
    |                  | 122kB 17.7MB/s eta 0:00:01
Collecting soupsieve>1.2; python_version >= "3.0" (from
beautifulsoup4->bs4)
  Downloading https://files.pythonhosted.org/packages/36/69/d82d04022f02733bf9a7
2bc3b96332d360c0c5307096d76f6bb7489f7e57/soupsieve-2.2.1-py3-none-any.whl
Building wheels for collected packages: bs4
  Building wheel for bs4 (setup.py) … done
  Stored in directory: /home/jupyterlab/.cache/pip/wheels/a0/b0/b2/4f80b94
56b87abedbc0bf2d52235414c3467d8889be38dd472
Successfully built bs4
Installing collected packages: soupsieve, beautifulsoup4, bs4
Successfully installed beautifulsoup4-4.9.3 bs4-0.0.1 soupsieve-2.2.1
```

[2]:
```
print('Hello Capstone Project Course!')
```

```
Hello Capstone Project Course!
```

Q1 of week 3 Assignment

[3]:
```
url = 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'
results = requests.get(url).text
```

```
soup = BeautifulSoup(results, 'html.parser')
```

[4]:
```
table = soup.find_all('table')[0]
alltd = table.find_all('td')
print(alltd[0], '\n', alltd[2], '\n', alltd[20])
```

```
<td style="width:11%; vertical-align:top; color:#ccc;">
<p><b>M1A</b><br/><span style="font-size:85%;"><i>Not assigned</i></span>
</p>
</td>
 <td style="width:11%; vertical-align:top;">
<p><b>M3A</b><br/><span style="font-size:85%;"><a href="/wiki/North_York"
title="North York">North York</a><br/>(<a href="/wiki/Parkwoods"
title="Parkwoods">Parkwoods</a>)</span>
</p>
</td>
 <td style="vertical-align:top;">
<p><b>M3C</b><br/><span style="font-size:85%;"><a href="/wiki/North_York"
title="North York">North York</a><br/>(<a href="/wiki/Don_Mills" title="Don
Mills">Don Mills</a>)<br/>South<br/>(<a href="/wiki/Flemingdon_Park"
title="Flemingdon Park">Flemingdon Park</a>)</span>
</p>
</td>
```

[5]:
```
table_contents = []

for td in alltd:
    cell = {}
    if td.span.text == 'Not assigned':
        pass
    else:
        cell['PostalCode'] = td.p.b.text
        temp = td.span.text.split('(')
        cell['Borough'] = temp[0]
        cell['Neighborhood'] = (((temp[1].strip(')')).replace(' /',',')).
 ↪replace(')',' ')).strip(' ')
        table_contents.append(cell)

# print(table_contents)
df = pd.DataFrame(table_contents)
df['Borough']=df['Borough'].replace({'Downtown TorontoStn A PO Boxes25 The␣
 ↪Esplanade' : 'Downtown Toronto Stn A',
                                     'East TorontoBusiness reply mail␣
 ↪Processing Centre969 Eastern' : 'East Toronto Business',
                                     'EtobicokeNorthwest' : 'Etobicoke␣
 ↪Northwest',
```

```
                                                'East YorkEast Toronto' : 'East␣
↪York/East Toronto',
                                                'MississaugaCanada Post Gateway␣
↪Processing Centre' : 'Mississauga'})

df
```

[5]:

	PostalCode	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	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	Enclave of M4L
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]

[6]:
```
df.shape
```

[6]:
```
(103, 3)
```

Q2 of Week 3 Assignment

[7]:
```
!pip install geocoder
import geocoder
```

```
Collecting geocoder
  Downloading https://files.pythonhosted.org/packages/4f/6b/13166c909ad2f2
d76b929a4227c952630ebaf0d729f6317eb09cbceccbab/geocoder-1.38.1-py2.py3-none-
any.whl (98kB)
```

```
     |                | 102kB 995kB/s ta 0:00:01
Requirement already satisfied: click in
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (from geocoder)
(7.1.2)
Requirement already satisfied: requests in
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (from geocoder)
(2.25.1)
Collecting ratelim (from geocoder)
  Downloading https://files.pythonhosted.org/packages/f2/98/7e6d147fd16a10a5f821
db6e25f192265d6ecca3d82957a4fdd592cad49c/ratelim-0.1.6-py2.py3-none-any.whl
Requirement already satisfied: future in
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (from geocoder)
(0.18.2)
Requirement already satisfied: six in
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (from geocoder)
(1.15.0)
Requirement already satisfied: idna<3,>=2.5 in
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (from
requests->geocoder) (2.10)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (from
requests->geocoder) (1.26.4)
Requirement already satisfied: certifi>=2017.4.17 in
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (from
requests->geocoder) (2020.12.5)
Requirement already satisfied: chardet<5,>=3.0.2 in
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (from
requests->geocoder) (4.0.0)
Requirement already satisfied: decorator in
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (from
ratelim->geocoder) (4.4.2)
Installing collected packages: ratelim, geocoder
Successfully installed geocoder-1.38.1 ratelim-0.1.6
```

[8]:
```
#since geocoder does not work, the geospatial csv file is used
df_geo = pd.read_csv('Geospatial_Coordinates.csv')
df_geo.rename({'Postal Code' : 'PostalCode'}, axis=1, inplace=True)
df2 = pd.merge(df, df_geo, how='inner', on='PostalCode')
df2
```

[8]:

	PostalCode	Borough	\
0	M3A	North York	
1	M4A	North York	
2	M5A	Downtown Toronto	
3	M6A	North York	
4	M7A	Queen's Park	
..	…	…	

```
98         M8X              Etobicoke
99         M4Y       Downtown Toronto
100        M7Y  East Toronto Business
101        M8Y              Etobicoke
102        M8Z              Etobicoke

                                        Neighborhood   Latitude  Longitude
0                                          Parkwoods  43.753259 -79.329656
1                                    Victoria Village  43.725882 -79.315572
2                              Regent Park, Harbourfront  43.654260 -79.360636
3                       Lawrence Manor, Lawrence Heights  43.718518 -79.464763
4                          Ontario Provincial Government  43.662301 -79.389494
..                                                   …        …        …
98       The Kingsway, Montgomery Road, Old Mill North  43.653654 -79.506944
99                                 Church and Wellesley  43.665860 -79.383160
100                                      Enclave of M4L  43.662744 -79.321558
101  Old Mill South, King's Mill Park, Sunnylea, Hu…  43.636258 -79.498509
102  Mimico NW, The Queensway West, South of Bloor,…  43.628841 -79.520999

[103 rows x 5 columns]
```

Q3 of Week 3 Assignment

To plot clusters of boroughs of Toronto containing the word 'Toronto'

[9]:
```
!conda install -c conda-forge folium=0.5.0 --yes
import folium
import matplotlib.cm as cm
import matplotlib.colors as colors
```

```
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
    attrs-20.3.0               |    pyhd3deb0d_0          41 KB  conda-forge
    branca-0.4.2               |    pyhd8ed1ab_0          26 KB  conda-forge
    ca-certificates-2020.12.5  |      ha878542_0         137 KB  conda-forge
    entrypoints-0.3            | pyhd8ed1ab_1003           8 KB  conda-forge
    folium-0.5.0               |            py_0          45 KB  conda-forge
    jsonschema-3.2.0           |    pyhd8ed1ab_3          45 KB  conda-forge
    pandas-1.1.5               |  py36h284efc9_0        11.3 MB  conda-forge
    pyrsistent-0.17.3          |  py36h8f6f2f9_2          89 KB  conda-forge
    pytz-2021.1                |    pyhd8ed1ab_0         239 KB  conda-forge
    vincent-0.4.4              |            py_1          28 KB  conda-forge
    ------------------------------------------------------------
                                           Total:        12.6 MB

The following NEW packages will be INSTALLED:

  altair             conda-forge/noarch::altair-4.1.0-py_1
  attrs              conda-forge/noarch::attrs-20.3.0-pyhd3deb0d_0
  branca             conda-forge/noarch::branca-0.4.2-pyhd8ed1ab_0
  entrypoints        conda-forge/noarch::entrypoints-0.3-pyhd8ed1ab_1003
  folium             conda-forge/noarch::folium-0.5.0-py_0
  jsonschema         conda-forge/noarch::jsonschema-3.2.0-pyhd8ed1ab_3
  pandas             conda-forge/linux-64::pandas-1.1.5-py36h284efc9_0
  pyrsistent         conda-forge/linux-64::pyrsistent-0.17.3-py36h8f6f2f9_2
  pytz               conda-forge/noarch::pytz-2021.1-pyhd8ed1ab_0
  vincent            conda-forge/noarch::vincent-0.4.4-py_1

The following packages will be SUPERSEDED by a higher-priority channel:

  ca-certificates    pkgs/main::ca-certificates-2021.4.13-~ --> conda-forge::ca-
certificates-2020.12.5-ha878542_0



Downloading and Extracting Packages
pyrsistent-0.17.3    | 89 KB     | ##################################### | 100%
folium-0.5.0         | 45 KB     | ##################################### | 100%
branca-0.4.2         | 26 KB     | ##################################### | 100%
altair-4.1.0         | 614 KB    | ##################################### | 100%
ca-certificates-2020 | 137 KB    | ##################################### | 100%
pandas-1.1.5         | 11.3 MB   | ##################################### | 100%
entrypoints-0.3      | 8 KB      | ##################################### | 100%
jsonschema-3.2.0     | 45 KB     | ##################################### | 100%
pytz-2021.1          | 239 KB    | ##################################### | 100%
attrs-20.3.0         | 41 KB     | ##################################### | 100%
vincent-0.4.4        | 28 KB     | ##################################### | 100%
Preparing transaction: done
Verifying transaction: done
```

```
Executing transaction: done
```

[10]:
```
df2['Interest'] = df2['Borough'].str.contains('Toronto')

#there are 7 boroughs (clusters)
df3 = df2[df2['Interest']==True]
df3.head()
```

[10]:

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Interest
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	True
9	M5B	Downtown Toronto	Garden District, Ryerson	43.657162	-79.378937	True
15	M5C	Downtown Toronto	St. James Town	43.651494	-79.375418	True
19	M4E	East Toronto	The Beaches	43.676357	-79.293031	True
20	M5E	Downtown Toronto	Berczy Park	43.644771	-79.373306	True

[11]:
```
print('Shape of df3: ', df3.shape)
print('Number of unique boroughs:', len(df3['Borough'].unique()))
```

```
Shape of df3:  (39, 6)
Number of unique boroughs: 7
```

[12]:
```
# Coordinates of Toronto from Google
tor_lat, tor_lng = 43.7181552, -79.5184859
```

[13]:
```
# create map
map_clusters = folium.Map(location=[tor_lat, tor_lng], zoom_start=11)

# set color scheme for the clusters and map a borough to a color
colors_array = cm.rainbow(np.linspace(0, 1, 7))
rainbow = [colors.rgb2hex(i) for i in colors_array]
uqb = df3['Borough'].unique()
coldict = dict(zip(uqb, rainbow))

# add markers to the map
markers_colors = []

for lat, lng, bor, neigh in zip(df3['Latitude'], df3['Longitude'],␣
 ↪df3['Borough'], df3['Neighborhood']):
    label = folium.Popup(str(neigh) + ' of ' + str(bor), parse_html=True)
    folium.CircleMarker(
```

```
        [lat, lng],
        radius=5,
        popup=label,
        color=coldict[bor],
        fill=True,
        fill_color=coldict[bor],
        fill_opacity=0.7).add_to(map_clusters)

map_clusters
```

[13]: <folium.folium.Map at 0x7f7d0ed7f390>