Untitled2

December 17, 2019

[52]:
```
pip install BeautifulSoup4
```

```
Requirement already satisfied: BeautifulSoup4 in
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (4.8.1)
Requirement already satisfied: soupsieve>=1.2 in
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (from
BeautifulSoup4) (1.9.5)
Note: you may need to restart the kernel to use updated packages.
```

[53]:
```
pip install lxml
```

```
Requirement already satisfied: lxml in
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (4.4.2)
Note: you may need to restart the kernel to use updated packages.
```

[54]:
```
pip install requests
```

```
Requirement already satisfied: requests in
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (2.22.0)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (from requests)
(1.25.7)
Requirement already satisfied: chardet<3.1.0,>=3.0.2 in
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (from requests)
(3.0.4)
Requirement already satisfied: idna<2.9,>=2.5 in
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (from requests)
(2.8)
Requirement already satisfied: certifi>=2017.4.17 in
/home/jupyterlab/conda/envs/python/lib/python3.6/site-packages (from requests)
(2019.9.11)
Note: you may need to restart the kernel to use updated packages.
```

[55]:
```
import pandas as pd
import numpy as np
import requests
from bs4 import BeautifulSoup
```

[ ]:

[56]:
```
#send request

url = 'https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M'

page = requests.get(url) #from here page.text gives the html text. We need to␣
 ↪parse the html using BeautifulSoup

soup = BeautifulSoup(page.text, 'html')
```

[57]:
```
#read table text

table = soup.find('table', {'class':'wikitable sortable'}).tbody

rows = table.find_all('tr')
columns = [v.text.replace('\n', '') for v in rows[0].find_all('th')]  # use␣
 ↪replace to remove \n
print(columns)
```

```
['Postcode', 'Borough', 'Neighbourhood']
```

[58]:
```
df = pd.DataFrame(columns=columns)

for i in range(1, len(rows)):
    tds = rows[i].find_all('td')

    if len(tds) ==4:
        values = [tds[0].text, '', ''.replace('\n', '')]  #use replace to␣
 ↪remove '\n'
    else:
        values = [td.text.replace('\n', '') for td in tds]          #use .
 ↪replace to remove '\n'

    df = df.append(pd.Series(values, index=columns), ignore_index=True)

df
```

[58]:

	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	Harbourfront
..	…	…	…
282	M8Z	Etobicoke	Mimico NW

```
283      M8Z          Etobicoke     The Queensway West
284      M8Z          Etobicoke  Royal York South West
285      M8Z          Etobicoke          South of Bloor
286      M9Z      Not assigned            Not assigned

[287 rows x 3 columns]
```

[59]:
```
# rename Postcode column
df.rename(columns={'Postcode': 'PostalCode'}, inplace=True)
```

[60]:
```
#dropping cells with Borough=Not assigned
df = df[df.Borough != 'Not assigned']
df
```

[60]:
```
    PostalCode           Borough              Neighbourhood
2          M3A        North York                  Parkwoods
3          M4A        North York           Victoria Village
4          M5A  Downtown Toronto               Harbourfront
5          M6A        North York           Lawrence Heights
6          M6A        North York             Lawrence Manor
..         …                 …                          …
281        M8Z          Etobicoke  Kingsway Park South West
282        M8Z          Etobicoke                  Mimico NW
283        M8Z          Etobicoke        The Queensway West
284        M8Z          Etobicoke     Royal York South West
285        M8Z          Etobicoke             South of Bloor

[210 rows x 3 columns]
```

[61]:
```
print('The dataframe has {} boroughs and {} neighborhoods.'.format(
        len(df['Borough'].unique()),
        df.shape[0]
    )
)
```

```
The dataframe has 11 boroughs and 210 neighborhoods.
```

[62]:
```
df['Neighbourhood'] = df['Neighbourhood'].astype(str)
neighborhoods1 = df.groupby(['PostalCode'], sort=False).agg( ','.join)

neighborhoods1
```

```
/home/jupyterlab/conda/envs/python/lib/python3.6/site-
packages/ipykernel_launcher.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-
```

```
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """Entry point for launching an IPython kernel.
```

[62]:

PostalCode	Borough	Neighbourhood
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Harbourfront
M6A	North York,North York	Lawrence Heights,Lawrence Manor
M7A	Queen's Park	Not assigned
…	…	…
M8X	Etobicoke,Etobicoke,Etobicoke	The Kingsway,Montgomery Road,Old Mill North
M4Y	Downtown Toronto	Church and Wellesley
M7Y	East Toronto	Business Reply Mail Processing Centre 969 Eastern
M8Y	Etobicoke,Etobicoke,Etobicoke,Etobicoke,Etobic…	Humber Bay,King's Mill Park,Kingsway Park Sout…
M8Z	Etobicoke,Etobicoke,Etobicoke,Etobicoke,Etobicoke	Kingsway Park South West,Mimico NW,The Queensw…

[103 rows x 2 columns]

[63]:
```
df['Neighbourhood'] = df['Neighbourhood'].astype(str)
neighborhoods1 = df.groupby(['PostalCode', 'Borough'], sort=False).agg( ','.
 ↪join)

neighborhoods1
```

```
/home/jupyterlab/conda/envs/python/lib/python3.6/site-
packages/ipykernel_launcher.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  """Entry point for launching an IPython kernel.
```

[63]:

```
                                                                Neighbourhood
PostalCode Borough
M3A        North York                                                Parkwoods
M4A        North York                                           Victoria Village
M5A        Downtown Toronto                                         Harbourfront
M6A        North York                              Lawrence Heights,Lawrence Manor
M7A        Queen's Park                                          Not assigned
…                                                                            …
M8X        Etobicoke                    The Kingsway,Montgomery Road,Old Mill North
M4Y        Downtown Toronto                                 Church and Wellesley
M7Y        East Toronto           Business Reply Mail Processing Centre 969 Eastern
M8Y        Etobicoke              Humber Bay,King's Mill Park,Kingsway Park Sout…
M8Z        Etobicoke              Kingsway Park South West,Mimico NW,The Queensw…

[103 rows x 1 columns]
```

[67]:

```
result = neighborhoods1.reset_index(level=['PostalCode', 'Borough'])
result
```

[67]:

```
    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
101        M8Y         Etobicoke
102        M8Z         Etobicoke

                                        Neighbourhood
0                                           Parkwoods
1                                      Victoria Village
2                                          Harbourfront
3                          Lawrence Heights,Lawrence Manor
4                                          Not assigned
..                                                  …
98         The Kingsway,Montgomery Road,Old Mill North
99                                   Church and Wellesley
100  Business Reply Mail Processing Centre 969 Eastern
101  Humber Bay,King's Mill Park,Kingsway Park Sout…
102  Kingsway Park South West,Mimico NW,The Queensw…

[103 rows x 3 columns]
```

[80]:
```
result.Neighbourhood = result.Borough + np.where(result.Neighbourhood=='Not␣
↪assigned')
```

```
      ␣
↪---------------------------------------------------------------------------

       TypeError                                 Traceback (most recent call␣
↪last)

       ~/conda/envs/python/lib/python3.6/site-packages/pandas/core/ops/__init__.
↪py in na_op(x, y)
       967         try:
   --> 968             result = expressions.evaluate(op, str_rep, x, y,␣
↪**eval_kwargs)
       969         except TypeError:


       ~/conda/envs/python/lib/python3.6/site-packages/pandas/core/computation/
↪expressions.py in evaluate(op, op_str, a, b, use_numexpr, **eval_kwargs)
       220     if use_numexpr:
   --> 221         return _evaluate(op, op_str, a, b, **eval_kwargs)
       222     return _evaluate_standard(op, op_str, a, b)


       ~/conda/envs/python/lib/python3.6/site-packages/pandas/core/computation/
↪expressions.py in _evaluate_standard(op, op_str, a, b, **eval_kwargs)
        69     with np.errstate(all="ignore"):
   ---> 70         return op(a, b)
        71


       TypeError: must be str, not int


   During handling of the above exception, another exception occurred:


       AssertionError                            Traceback (most recent call␣
↪last)

       <ipython-input-80-5aca5f2d4e76> in <module>
   ----> 1 result.Neighbourhood = result.Borough + np.where(result.
↪Neighbourhood=='Not assigned')
```

```
      ~/conda/envs/python/lib/python3.6/site-packages/pandas/core/ops/__init__.
↪py in wrapper(left, right)
     1046
     1047         with np.errstate(all="ignore"):
  -> 1048             result = na_op(lvalues, rvalues)
     1049         return construct_result(
     1050             left, result, index=left.index, name=res_name, dtype=None


      ~/conda/envs/python/lib/python3.6/site-packages/pandas/core/ops/__init__.
↪py in na_op(x, y)
      968             result = expressions.evaluate(op, str_rep, x, y,␣
↪**eval_kwargs)
      969         except TypeError:
  --> 970             result = masked_arith_op(x, y, op)
      971
      972         return missing.dispatch_fill_zeros(op, x, y, result)


      ~/conda/envs/python/lib/python3.6/site-packages/pandas/core/ops/__init__.
↪py in masked_arith_op(x, y, op)
      448
      449     else:
  --> 450         assert is_scalar(y), type(y)
      451         assert isinstance(x, np.ndarray), type(x)
      452         # mask is only meaningful for x


      AssertionError: <class 'tuple'>
```

[ ]: