Segmenting and Clustering Neighborhoods in Toronto

Import requests and panda

In [2]:
```
import requests
import pandas as pd
```

Get the HTML of the Wiki page, convert into a table with help of read_html (read HTML tables into a list of DataFrame objects), remove cells with a borough that is Not assigned.

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

df_raw = pd.read_html(wikipedia_page.content, header=0)[0]
df_new = df_raw[df_raw.Borough != 'Not assigned']

df_new.head()
```

Out[3]:

	Postal Code	Borough	Neighbourhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Find whether there is a "Not assigned" in Neighbourhood

In [4]:
```
df_new.loc[df_new.Neighbourhood == 'Not assigned']
```

Out[4]:

	Postal Code	Borough	Neighbourhood

If we have Neighbourhood Not assigned, we change it with the value of Borough

In [5]:
```
df_new.Neighbourhood.replace('Not assigned',df_new.Borough,inplace=True)
df_new.head(8)
```

```
/opt/conda/envs/Python36/lib/python3.6/site-packages/pandas/core/generic.py:6586: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy
  self._update_inplace(new_data)
```

Out[5]:

	Postal Code	Borough	Neighbourhood
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront
5	M6A	North York	Lawrence Manor, Lawrence Heights
6	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government
8	M9A	Etobicoke	Islington Avenue, Humber Valley Village
9	M1B	Scarborough	Malvern, Rouge
11	M3B	North York	Don Mills

Group Neighbourhoods with the same Postcode

In [6]:
```
df_toronto = df_new.groupby(['Postal Code', 'Borough'])['Neighbourhood'].apply(lambda x: ', '.join(x))
df_toronto = df_toronto.reset_index()
df_toronto.rename(columns = {'Postal Code':'PostalCode'}, inplace = True)
df_toronto.rename(columns = {'Neighbourhood':'Neighborhood'}, inplace = True)
df_toronto.head()
```

Out[6]:

	PostalCode	Borough	Neighborhood
0	M1B	Scarborough	Malvern, Rouge
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

In [7]:
```
df_toronto.shape
```

Out[7]:
```
(103, 3)
```