# BattleofNeighborhoods

April 14, 2021

The Battle of the Neighborhoods

## 1 Table of Contents

## 2 Intoduction: Business Problem

**Background** Nowadays, there is a substantial increment of remote work due to the fact that the society is going through a pandemic. A lot of companies have offered the possibility to their employees to work from home (mainly). However, some employees could have had the idea of taking this opportunity to keep working for their employers but from other locations where in some way, could represent a better quality of life, or better personal or professional satisfaction.

People in Latin America that have enough level of incomes could have thought that this new way of working can be a opportunity to move to another country (where english is the official language) to learn a new language. A perfect combination has arised because those people could re locate to another country without the risk of losing their jobs.

**Problem** The question now is how a person that wants to learn english and is looking to move to a english spoken country can find the best option? Over the internet, there are a lot of courses offered in different cities, so, the main purpose of this exercise is finding the best one taking into account the variety of venues around the course location as a decisor.

**It is important to aclare that this exercise assumes that the closer the housing to the course location the better for the "new student".**

## 3 Data sources

### Where do we find the data?

The two main data sources are:

- Web scraping
- Foursquare API

The exercise is made for people who is looking to move to Canada or United States so, the website that is used to scrape the data of the courses will only be for USA and Canada.

***Only the best 15 courses for each country according to the website will be considered***

The URL for the courses in **Canada**: https://www.languageinternational.com/english-courses-canada

The URL for the courses in **United States**: https://www.languageinternational.com/english-courses-usa

If we look at the web sites, the main page does not have the addresses of the institutes, so, the first task is finding their addresses within the website.

### How it will be used?

Once all the addresses are saved in the dataframe of the courses (the structure of the dataframe will be **institutename, city, name, address, postalcode**), the next step is finding the coordinates of each institute.

After having all the neccesary data of the courses with their specific addresses the next step is use Foursquare as the provider of the venues around specific locations (through the API)

**Note:** the locations that are going to be used in the API are the institue ones.

## 4 Data extraction

### Let's import necessary Libraries

[2]:

```
# This section is used to import libraries that are going to be used along the␣
 ↪project

!pip install bs4
from bs4 import BeautifulSoup # this module helps in web scrapping.

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

!conda install -c conda-forge folium=0.5.0 --yes
```

```
import folium # map rendering library

import numpy as np # library to handle data in a vectorized manner

import pandas as pd # library for data analsysis

import re # library for regular expressions

import json # library to handle JSON files

import requests # library to handle requests
from pandas.io.json import json_normalize # tranform JSON file into a pandas␣
 ↪dataframe

# Matplotlib and associated plotting modules
import matplotlib.cm as cm
import matplotlib.colors as colors
import matplotlib.pyplot as plt

#!conda install -c conda-forge plotly
#import plotly.express as px
#import plotly.graph_objects as go

# import k-means from clustering stage
from sklearn.cluster import KMeans

from IPython.display import Image

print('Libraries imported.')
```

```
Requirement already satisfied: bs4 in
c:\users\sebastian.bedoya\anaconda3\lib\site-packages (0.0.1)
Requirement already satisfied: beautifulsoup4 in
c:\users\sebastian.bedoya\anaconda3\lib\site-packages (from bs4) (4.9.3)
Requirement already satisfied: soupsieve>1.2; python_version >= "3.0" in
c:\users\sebastian.bedoya\anaconda3\lib\site-packages (from beautifulsoup4->bs4)
(2.0.1)
Collecting package metadata (current_repodata.json): …working… done
Solving environment: …working… done

# All requested packages already installed.

Collecting package metadata (current_repodata.json): …working… done
Solving environment: …working… done

# All requested packages already installed.

Libraries imported.
```

**Now it's time to start doing the web scraping to obtain the addresses of the institutes**

The first url to scrape is for the courses of Canada

[3]:
```
urlcanada="https://www.languageinternational.com/english-courses-canada"

html_data_canada  = requests.get(urlcanada).text
soup_canada = BeautifulSoup(html_data_canada,"html5lib")
```

Let's create a dataframe where the scraped data is going to be stored.

**The dataframe will have 6 columns**

[4]:
```
df_courses = pd.DataFrame(columns=["Location", "CourseName", "InstituteName",␣
 ↪"PriceFrom", "PriceTo", "Address", "urlcourse", "urlinstitute", "latitude",␣
 ↪"longitude", "Country"])
```

All the data that we need is inside the div section with the class: **row gutterh-xs**

[5]:
```
for div in soup_canada.find_all('div',class_='row gutterh-xs'):

    location = div.find_all(attrs={'class': 'searchresult_subh'})[0].text
    coursename = div.find_all(itemprop='name')[0].text
    institutename = div.find_all(itemprop='name')[1].text
    pricefrom = div.find_all(attrs={'class':␣
 ↪'courselist_course_prices_box_amount'})[0].text
    priceto = div.find_all(attrs={'class':␣
 ↪'courselist_course_prices_box_amount'})[1].text

    url = div.find_all('a', href=True)[0]

    df_courses = df_courses.append({"Location":location, "CourseName":
 ↪coursename, "InstituteName":institutename, "PriceFrom":pricefrom, "PriceTo":
 ↪priceto, "urlcourse":url['href'], "Country": "Canada"}, ignore_index=True)
```

The columns that contains prices need to be cleaned.

- Regex patters are used to replace some strings present in both columns

[6]:
```
df_courses[['PriceFrom','PriceTo']] = df_courses[['PriceFrom','PriceTo']].
 ↪replace(to_replace=["\t|\n|\r", "\W"], value=["", ""], regex=True)
```

Here the result of the web scraping from the url for the courses in Canada

[7]:
```
df_courses.head()
```

[7]:
```
            Location                 CourseName  \
0    Calgary, Canada  General English 20 (GE20)
1   Hamilton, Canada        ESL Private Lessons
2    Calgary, Canada     General English (GE25)
3   Hamilton, Canada          Intensive English
```

```
4  Vancouver, Canada                 General 20

                                     InstituteName PriceFrom PriceTo Address  \
0                            ANNE'S Language House       354     687     NaN
1                             Metropolitan College       384     619     NaN
2                            ANNE'S Language House       384     717     NaN
3                             Metropolitan College       290     525     NaN
4  Language Studies International (LSI): Vancouver       431     759     NaN

                                           urlcourse urlinstitute latitude  \
0  /course/general-english-20-ge20-anne-s-languag…          NaN      NaN
1  /course/esl-private-lessons-metropolitan-colle…          NaN      NaN
2  /course/general-english-ge25-anne-s-language-h…          NaN      NaN
3  /course/intensive-english-metropolitan-college…          NaN      NaN
4  /course/general-20-language-studies-internatio…          NaN      NaN

  longitude Country
0       NaN  Canada
1       NaN  Canada
2       NaN  Canada
3       NaN  Canada
4       NaN  Canada
```

Now, let's do exactly the same but for the courses that are offered in the United States

[8]:
```
urlusa="https://www.languageinternational.com/english-courses-usa"

html_data_usa  = requests.get(urlusa).text
soup_usa = BeautifulSoup(html_data_usa,"html5lib")
```

[9]:
```
for div in soup_usa.find_all('div',class_='row gutterh-xs'):

    location = div.find_all(attrs={'class': 'searchresult_subh'})[0].text
    coursename = div.find_all(itemprop='name')[0].text
    institutename = div.find_all(itemprop='name')[1].text
    pricefrom = div.find_all(attrs={'class':␣
 ↪'courselist_course_prices_box_amount'})[0].text
    priceto = div.find_all(attrs={'class':␣
 ↪'courselist_course_prices_box_amount'})[1].text

    url = div.find_all('a', href=True)[0]


    df_courses = df_courses.append({"Location":location, "CourseName":
 ↪coursename, "InstituteName":institutename, "PriceFrom":pricefrom, "PriceTo":
 ↪priceto, "urlcourse":url['href'], "Country":"USA"}, ignore_index=True)
```

```
[10]: df_courses[['PriceFrom','PriceTo']] = df_courses[['PriceFrom','PriceTo']].
       ↪replace(to_replace=["\t|\n|\r", "\W"], value=["", ""], regex=True)
```

**Now we have the dataframe with the data of Canada and United States courses**

```
[11]: df_courses
```

```
[11]:
              Location                                         CourseName  \
0      Calgary, Canada                          General English 20 (GE20)
1     Hamilton, Canada                                ESL Private Lessons
2      Calgary, Canada                             General English (GE25)
3     Hamilton, Canada                                  Intensive English
4    Vancouver, Canada                                         General 20
5      Toronto, Canada                                       Intensive 30
6    Vancouver, Canada                                       Intensive 30
7      Toronto, Canada                                         General 20
8    Vancouver, Canada                                       Intensive 25
9      Toronto, Canada  Club 40+ (20 lessons per week plus afternoon a…
10     Toronto, Canada                                       Afternoon 10
11   Vancouver, Canada  Club 40+ (20 lessons per week plus afternoon a…
12   Vancouver, Canada                                       Afternoon 10
13   Vancouver, Canada                        General English - Intensive
14    Montreal, Canada                            Super Intensive English
15      San Diego, USA                   English Max Course (18 hrs/week)
16      San Diego, USA                 English Focus Course (12 hrs/week)
17      San Diego, USA                                        English Max
18      San Diego, USA                   English Max Course (18 hrs/week)
19      San Diego, USA                 English Focus Course (12 hrs/week)
20    Los Angeles, USA                    ESL Program (12 Weeks Minimum)
21         Boston, USA                                       Intensive 30
22         Boston, USA  Club 40+ (20 lessons per week plus afternoon a…
23      San Diego, USA                                       Intensive 30
24  New York City, USA                   One-to-One (20 lessons per week)
25         Boston, USA                    One-to-One (5 lessons per week)
26         Boston, USA                                         General 20
27  New York City, USA                   One-to-One (10 lessons per week)
28      San Diego, USA  Club 40+ (20 lessons per week plus afternoon a…
29      San Diego, USA                                         General 20

                                        InstituteName PriceFrom PriceTo  \
0                               ANNE'S Language House       354     687
1                                Metropolitan College       384     619
2                               ANNE'S Language House       384     717
3                                Metropolitan College       290     525
4     Language Studies International (LSI): Vancouver       431     759
5       Language Studies International (LSI): Toronto       495     823
6     Language Studies International (LSI): Vancouver       495     823
7       Language Studies International (LSI): Toronto       431     759
```

```
8     Language Studies International (LSI): Vancouver       465     794
9       Language Studies International (LSI): Toronto       708    1037
10      Language Studies International (LSI): Toronto       286     614
11    Language Studies International (LSI): Vancouver       708    1037
12    Language Studies International (LSI): Vancouver       286     614
13  International Language Academy of Canada Vanco…       503     870
14                   Bouchereau Lingua International       533     887
15  Connect English Language Institute- Mission Va…       390    None
16                         Connect English- La Jolla       365    None
17  Connect English Language Institute- San Diego …       390    None
18                         Connect English- La Jolla       390    None
19  Connect English Language Institute- Mission Va…       315    None
20                  American English Language School       335    None
21       Language Studies International (LSI): Boston       645     960
22       Language Studies International (LSI): Boston       840    1155
23    Language Studies International (LSI): San Diego       645     950
24     Language Studies International (LSI): New York      1955    2280
25       Language Studies International (LSI): Boston       575     890
26       Language Studies International (LSI): Boston       535     850
27     Language Studies International (LSI): New York      1055    1380
28    Language Studies International (LSI): San Diego       840    1240
29    Language Studies International (LSI): San Diego       535     935

   Address                                        urlcourse urlinstitute  \
0      NaN  /course/general-english-20-ge20-anne-s-languag…          NaN
1      NaN  /course/esl-private-lessons-metropolitan-colle…          NaN
2      NaN  /course/general-english-ge25-anne-s-language-h…          NaN
3      NaN  /course/intensive-english-metropolitan-college…          NaN
4      NaN  /course/general-20-language-studies-internatio…          NaN
5      NaN  /course/intensive-30-language-studies-internat…          NaN
6      NaN  /course/intensive-30-language-studies-internat…          NaN
7      NaN  /course/general-20-language-studies-internatio…          NaN
8      NaN  /course/intensive-25-language-studies-internat…          NaN
9      NaN  /course/club-40-20-lessons-per-week-plus-after…          NaN
10     NaN  /course/afternoon-10-language-studies-internat…          NaN
11     NaN  /course/club-40-20-lessons-per-week-plus-after…          NaN
12     NaN  /course/afternoon-10-language-studies-internat…          NaN
13     NaN  /course/general-english-intensive-internationa…          NaN
14     NaN  /course/super-intensive-english-bouchereau-lin…          NaN
15     NaN  /course/english-max-course-18-hrs-week-connect…          NaN
16     NaN  /course/english-focus-course-12-hrs-week-conne…          NaN
17     NaN  /course/english-max-connect-english-language-i…          NaN
18     NaN  /course/english-max-course-18-hrs-week-connect…          NaN
19     NaN  /course/english-focus-course-12-hrs-week-conne…          NaN
20     NaN  /course/esl-program-12-weeks-minimum-american-…          NaN
21     NaN  /course/intensive-30-language-studies-internat…          NaN
22     NaN  /course/club-40-20-lessons-per-week-plus-after…          NaN
```

```
23     NaN  /course/intensive-30-language-studies-internat…           NaN
24     NaN  /course/one-to-one-20-lessons-per-week-languag…           NaN
25     NaN  /course/one-to-one-5-lessons-per-week-language…           NaN
26     NaN  /course/general-20-language-studies-internatio…           NaN
27     NaN  /course/one-to-one-10-lessons-per-week-languag…           NaN
28     NaN  /course/club-40-20-lessons-per-week-plus-after…           NaN
29     NaN  /course/general-20-language-studies-internatio…           NaN

   latitude longitude Country
0       NaN       NaN  Canada
1       NaN       NaN  Canada
2       NaN       NaN  Canada
3       NaN       NaN  Canada
4       NaN       NaN  Canada
5       NaN       NaN  Canada
6       NaN       NaN  Canada
7       NaN       NaN  Canada
8       NaN       NaN  Canada
9       NaN       NaN  Canada
10      NaN       NaN  Canada
11      NaN       NaN  Canada
12      NaN       NaN  Canada
13      NaN       NaN  Canada
14      NaN       NaN  Canada
15      NaN       NaN     USA
16      NaN       NaN     USA
17      NaN       NaN     USA
18      NaN       NaN     USA
19      NaN       NaN     USA
20      NaN       NaN     USA
21      NaN       NaN     USA
22      NaN       NaN     USA
23      NaN       NaN     USA
24      NaN       NaN     USA
25      NaN       NaN     USA
26      NaN       NaN     USA
27      NaN       NaN     USA
28      NaN       NaN     USA
29      NaN       NaN     USA
```

The column **Address** and **urlInstitue** as you can see, have **NaN** values because the address and the url of the institute is not present in the scraped url.

The address can be found once the course is selected. A new page needs to be scraped for each one of the courses present in the dataframe

The url format to scrape data related to the course for all the rows is https://www.languageinternational.com/xxxxxxxxxx

Where ***xxxxxxxxxx*** will be replaced with the **urlcourse column** to scrape the address and url of the institute of each course

**Note:** the url of the institute is scraped because the data of the latitude and longitude is already present in the web site in the page of the institutes

[12]:
```
for index, row in df_courses.iterrows():

    urlcourse="https://www.languageinternational.com" + row['urlcourse']
    html_data_course  = requests.get(urlcourse).text
    soup_course = BeautifulSoup(html_data_course,"html5lib")

    for div in soup_course.find_all(attrs={'class': 'school_upper'}):
        df_courses.iloc[index]['Address'] = div.find_all('span')[0].text
        print('Address: {} '.format(div.find_all('span')[0].text))

        for div in soup_course.find_all(attrs={'class': 'pageheader_text'}):
            df_courses.iloc[index]['urlinstitute'] = div.find_all('a',␣
↪href=True)[0]['href']
            print('url of institue: {} '.format(div.find_all('a',␣
↪href=True)[0]['href']))
            print('--- Added to dataframe ---')

print('----------------')
print('--End of update-')
```

```
Address: 101 6th Avenue S.W., Suite 1250, Calgary, Alberta, AB T2P3P4, Canada
url of institue: /school/anne-s-language-house-64694
--- Added to dataframe ---
Address: 146 James Street South, Hamilton, Ontario,, Hamilton, Ontario L8P 3A2,
Canada
url of institue: /school/metropolitan-college-64750
--- Added to dataframe ---
Address: 101 6th Avenue S.W., Suite 1250, Calgary, Alberta, AB T2P3P4, Canada
url of institue: /school/anne-s-language-house-64694
--- Added to dataframe ---
Address: 146 James Street South, Hamilton, Ontario,, Hamilton, Ontario L8P 3A2,
Canada
url of institue: /school/metropolitan-college-64750
--- Added to dataframe ---
Address: 101-808 Nelson Street, Vancouver, BC V6Z 2H2, Canada
url of institue: /school/language-studies-international-lsi-vancouver-18606
--- Added to dataframe ---
Address: 1055 Yonge Street, Suite #210, Toronto, ON M4W 2L2, Canada
url of institue: /school/language-studies-international-lsi-toronto-18605
--- Added to dataframe ---
Address: 101-808 Nelson Street, Vancouver, BC V6Z 2H2, Canada
url of institue: /school/language-studies-international-lsi-vancouver-18606
```

```
--- Added to dataframe ---
Address: 1055 Yonge Street, Suite #210, Toronto, ON M4W 2L2, Canada
url of institue: /school/language-studies-international-lsi-toronto-18605
--- Added to dataframe ---
Address: 101-808 Nelson Street, Vancouver, BC V6Z 2H2, Canada
url of institue: /school/language-studies-international-lsi-vancouver-18606
--- Added to dataframe ---
Address: 1055 Yonge Street, Suite #210, Toronto, ON M4W 2L2, Canada
url of institue: /school/language-studies-international-lsi-toronto-18605
--- Added to dataframe ---
Address: 1055 Yonge Street, Suite #210, Toronto, ON M4W 2L2, Canada
url of institue: /school/language-studies-international-lsi-toronto-18605
--- Added to dataframe ---
Address: 101-808 Nelson Street, Vancouver, BC V6Z 2H2, Canada
url of institue: /school/language-studies-international-lsi-vancouver-18606
--- Added to dataframe ---
Address: 101-808 Nelson Street, Vancouver, BC V6Z 2H2, Canada
url of institue: /school/language-studies-international-lsi-vancouver-18606
--- Added to dataframe ---
Address: 1199 West Pender Street, Vancouver, British Columbia V6E 2R1, Canada
url of institue: /school/international-language-academy-of-canada-
vancouver-63381
--- Added to dataframe ---
Address: 70 Notre Dame West, Suite 400, Montréal, Quebec H2Y 1S6, Canada
url of institue: /school/bouchereau-lingua-international-63865
--- Added to dataframe ---
Address: 4560 Alvarado Canyon Road, Suite 2B, San Diego, California 92120, USA
url of institue: /school/connect-english-language-institute-mission-valley-63248
--- Added to dataframe ---
Address: 5090 Shoreham Place, Suite 206, San Diego, California 92037, USA
url of institue: /school/connect-english-la-jolla-64046
--- Added to dataframe ---
Address: 3565 Del Rey St., Suite 300, San Diego, California 92109, USA
url of institue: /school/connect-english-language-institute-san-diego-pacific-
beach-campus-64605
--- Added to dataframe ---
Address: 5090 Shoreham Place, Suite 206, San Diego, California 92037, USA
url of institue: /school/connect-english-la-jolla-64046
--- Added to dataframe ---
Address: 4560 Alvarado Canyon Road, Suite 2B, San Diego, California 92120, USA
url of institue: /school/connect-english-language-institute-mission-valley-63248
--- Added to dataframe ---
Address: 3230 E. Imperial Hwy, Suite 301, Brea, California 92821, USA
url of institue: /school/american-english-language-school-68322
--- Added to dataframe ---
Address: 105 Beach Street, Boston, MA 02111, USA
url of institue: /school/language-studies-international-lsi-boston-68
--- Added to dataframe ---
```

```
Address: 105 Beach Street, Boston, MA 02111, USA
url of institue: /school/language-studies-international-lsi-boston-68
--- Added to dataframe ---
Address: 1706 5th Avenue, San Diego, CA 92101, USA
url of institue: /school/language-studies-international-lsi-san-diego-231
--- Added to dataframe ---
Address: 40 Rector Street, 10th Floor, Suite 1000, New York, NY 10006, USA
url of institue: /school/language-studies-international-lsi-new-york-232
--- Added to dataframe ---
Address: 105 Beach Street, Boston, MA 02111, USA
url of institue: /school/language-studies-international-lsi-boston-68
--- Added to dataframe ---
Address: 105 Beach Street, Boston, MA 02111, USA
url of institue: /school/language-studies-international-lsi-boston-68
--- Added to dataframe ---
Address: 40 Rector Street, 10th Floor, Suite 1000, New York, NY 10006, USA
url of institue: /school/language-studies-international-lsi-new-york-232
--- Added to dataframe ---
Address: 1706 5th Avenue, San Diego, CA 92101, USA
url of institue: /school/language-studies-international-lsi-san-diego-231
--- Added to dataframe ---
Address: 1706 5th Avenue, San Diego, CA 92101, USA
url of institue: /school/language-studies-international-lsi-san-diego-231
--- Added to dataframe ---
----------------
--End of update-
```

[13]:
```
df_courses.head()
```

[13]:
```
            Location                 CourseName  \
0    Calgary, Canada  General English 20 (GE20)   
1   Hamilton, Canada        ESL Private Lessons   
2    Calgary, Canada     General English (GE25)   
3   Hamilton, Canada          Intensive English   
4  Vancouver, Canada                 General 20   

                                    InstituteName PriceFrom PriceTo  \
0                            ANNE'S Language House       354     687   
1                              Metropolitan College       384     619   
2                            ANNE'S Language House       384     717   
3                              Metropolitan College       290     525   
4  Language Studies International (LSI): Vancouver       431     759   

                                           Address  \
0  101 6th Avenue S.W., Suite 1250, Calgary, Albe…   
1  146 James Street South, Hamilton, Ontario,, Ha…   
2  101 6th Avenue S.W., Suite 1250, Calgary, Albe…   
```

```
3  146 James Street South, Hamilton, Ontario,, Ha…
4  101-808 Nelson Street, Vancouver, BC V6Z 2H2, …

                                           urlcourse  \
0  /course/general-english-20-ge20-anne-s-languag…
1  /course/esl-private-lessons-metropolitan-colle…
2  /course/general-english-ge25-anne-s-language-h…
3  /course/intensive-english-metropolitan-college…
4  /course/general-20-language-studies-internatio…

                                        urlinstitute latitude longitude  \
0                /school/anne-s-language-house-64694      NaN       NaN
1                 /school/metropolitan-college-64750      NaN       NaN
2                /school/anne-s-language-house-64694      NaN       NaN
3                 /school/metropolitan-college-64750      NaN       NaN
4  /school/language-studies-international-lsi-van…      NaN       NaN

  Country
0  Canada
1  Canada
2  Canada
3  Canada
4  Canada
```

At this point, in the dataframe, the columns *address* and *urlinstitute* are already filled up with the scraped data

With this data ready, we can proceed to get the coordinates from the website as well. The coordinates will be added to the dataframe

**Note:** in this step, a regular expression is used to extract a portion of the entire text where the coordinates are

[14]:

```python
for index, row in df_courses.iterrows():
    urlinstitute="https://www.languageinternational.com" + row['urlinstitute']
    html_data_institute  = requests.get(urlinstitute).text

    # Create a pattern to match names
    pattern = re.compile(r'(?=latitude)(.*)(?=url)', flags = re.M)
    # Find all occurrences of the pattern
    result = pattern.findall(html_data_institute)
    # Let's clean the text matched
    mapping = {'latitude':'', 'longitude':'', '"':'', ':':''}
    for k, v in mapping.items():
        result[0] = result[0].replace(k, v)

    # Now, a split is used to assign latitude and longitude variables
    latitude = result[0].split(',')[0]
    df_courses.iloc[index]['latitude'] = result[0].split(',')[0]

    longitude = result[0].split(',')[1]
    df_courses.iloc[index]['longitude'] = result[0].split(',')[1]

    print(latitude, longitude)
    print('--- Lat and Lon added ---')

print('----------------')
print('--End of update-')
```

```
51.04715 -114.06334
--- Lat and Lon added ---
43.25243 -79.8713
--- Lat and Lon added ---
51.04715 -114.06334
--- Lat and Lon added ---
43.25243 -79.8713
--- Lat and Lon added ---
49.28004 -123.12491
--- Lat and Lon added ---
43.67851 -79.38973
--- Lat and Lon added ---
49.28004 -123.12491
--- Lat and Lon added ---
43.67851 -79.38973
--- Lat and Lon added ---
49.28004 -123.12491
--- Lat and Lon added ---
43.67851 -79.38973
--- Lat and Lon added ---
43.67851 -79.38973
--- Lat and Lon added ---
49.28004 -123.12491
--- Lat and Lon added ---
49.28004 -123.12491
--- Lat and Lon added ---
49.28836 -123.12259
--- Lat and Lon added ---
45.466221 -74.077549
--- Lat and Lon added ---
32.78098 -117.09628
--- Lat and Lon added ---
32.85252 -117.18622
--- Lat and Lon added ---
32.80372 -117.21411
--- Lat and Lon added ---
32.85252 -117.18622
```

```
--- Lat and Lon added ---
32.78098 -117.09628
--- Lat and Lon added ---
33.90986 -117.85417
--- Lat and Lon added ---
42.35072 -71.05826
--- Lat and Lon added ---
42.35072 -71.05826
--- Lat and Lon added ---
32.72327 -117.16051
--- Lat and Lon added ---
40.70863 -74.01466
--- Lat and Lon added ---
42.35072 -71.05826
--- Lat and Lon added ---
42.35072 -71.05826
--- Lat and Lon added ---
40.70863 -74.01466
--- Lat and Lon added ---
32.72327 -117.16051
--- Lat and Lon added ---
32.72327 -117.16051
--- Lat and Lon added ---
----------------
--End of update-
```

**The dataframe now is complete!**

```
[15]: df_courses.head()
```

```
[15]:             Location                 CourseName  \
      0    Calgary, Canada  General English 20 (GE20)
      1   Hamilton, Canada        ESL Private Lessons
      2    Calgary, Canada     General English (GE25)
      3   Hamilton, Canada          Intensive English
      4  Vancouver, Canada                 General 20

                                        InstituteName PriceFrom PriceTo  \
      0                           ANNE'S Language House       354     687
      1                             Metropolitan College       384     619
      2                           ANNE'S Language House       384     717
      3                             Metropolitan College       290     525
      4  Language Studies International (LSI): Vancouver       431     759

                                                Address  \
      0  101 6th Avenue S.W., Suite 1250, Calgary, Albe…
      1  146 James Street South, Hamilton, Ontario,, Ha…
      2  101 6th Avenue S.W., Suite 1250, Calgary, Albe…
```

```
3  146 James Street South, Hamilton, Ontario,, Ha…
4  101-808 Nelson Street, Vancouver, BC V6Z 2H2, …

                                           urlcourse  \
0  /course/general-english-20-ge20-anne-s-languag…
1  /course/esl-private-lessons-metropolitan-colle…
2  /course/general-english-ge25-anne-s-language-h…
3  /course/intensive-english-metropolitan-college…
4  /course/general-20-language-studies-internatio…

                                        urlinstitute  latitude   longitude  \
0                /school/anne-s-language-house-64694  51.04715  -114.06334
1                 /school/metropolitan-college-64750  43.25243    -79.8713
2                /school/anne-s-language-house-64694  51.04715  -114.06334
3                 /school/metropolitan-college-64750  43.25243    -79.8713
4  /school/language-studies-international-lsi-van…  49.28004  -123.12491

  Country
0  Canada
1  Canada
2  Canada
3  Canada
4  Canada
```

## 5 Institutes visualization and Data

Looking at the data, it is suitable to have a child dataframe with the institutes with their respectives coordinates to avoid duplicates. This happens because it can be that a institute can offer more than one course.

[16]:

```
df_institutes = df_courses[['InstituteName','latitude','longitude', 'Country']].
 ↪drop_duplicates(subset=['InstituteName','latitude','longitude', 'Country'],␣
 ↪keep='last').reset_index()
df_institutes
```

[16]:

```
    index                                      InstituteName   latitude  \
0       2                              ANNE'S Language House   51.04715
1       3                               Metropolitan College   43.25243
2      10      Language Studies International (LSI): Toronto   43.67851
3      12    Language Studies International (LSI): Vancouver   49.28004
4      13  International Language Academy of Canada Vanco…   49.28836
5      14                    Bouchereau Lingua International  45.466221
6      17  Connect English Language Institute- San Diego …   32.80372
7      18                          Connect English- La Jolla   32.85252
8      19  Connect English Language Institute- Mission Va…   32.78098
9      20                   American English Language School   33.90986
10     26       Language Studies International (LSI): Boston   42.35072
```

```
11     27     Language Studies International (LSI): New York   40.70863
12     29    Language Studies International (LSI): San Diego   32.72327

     longitude Country
0   -114.06334  Canada
1     -79.8713  Canada
2    -79.38973  Canada
3   -123.12491  Canada
4   -123.12259  Canada
5   -74.077549  Canada
6   -117.21411     USA
7   -117.18622     USA
8   -117.09628     USA
9   -117.85417     USA
10   -71.05826     USA
11   -74.01466     USA
12  -117.16051     USA
```

[17]:

```
df_institutes["latitude"] = df_institutes["latitude"].astype("float")
df_institutes["longitude"] = df_institutes["longitude"].astype("float")

print(df_institutes.dtypes)
```

```
index              int64
InstituteName     object
latitude         float64
longitude        float64
Country           object
dtype: object
```

[18]:

```
print('--> We can observe now that there are ', df_institutes.shape[0], '␣
 ↪different institutes that offer the best 30 courses in Canada and United␣
 ↪States <--')
```

```
--> We can observe now that there are  13  different institutes that offer the
best 30 courses in Canada and United States <--
```

[19]:

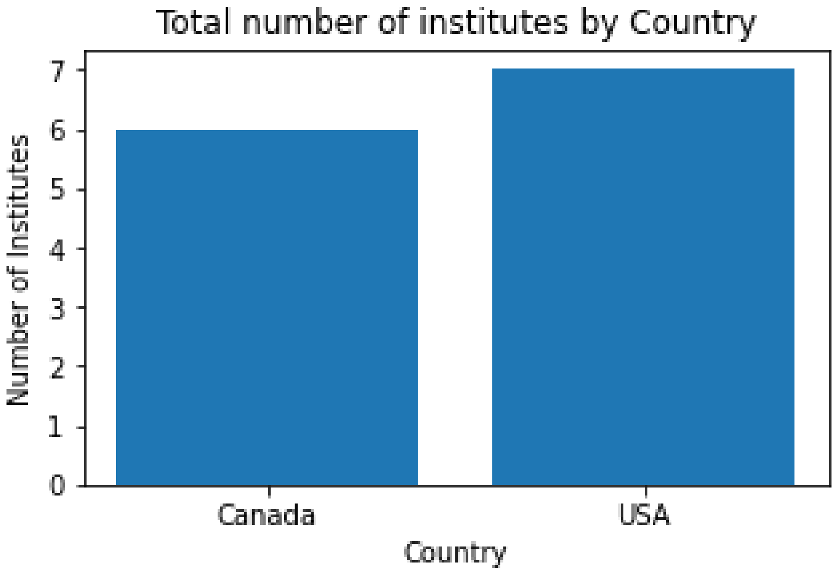
```
bar_data = df_institutes.groupby(['Country'])['InstituteName'].count().
 ↪reset_index()
bar_data
```

[19]:

```
  Country  InstituteName
0  Canada              6
1     USA              7
```

[20]:

```
fig, ax = plt.subplots(figsize=(5, 3))

ax.bar(bar_data['Country'],bar_data['InstituteName'])
ax.set_xlabel('Country')
ax.set_ylabel('Number of Institutes')

plt.title('Total number of institutes by Country')
plt.show()
print('Done!')
```



```
Done!
```

Let's get the geographical coordinates of North America.

[21]:
```
address = 'North America'

geolocator = Nominatim(user_agent="northamerica_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude

print('The geograpical coordinate of North America are {}, {}.'.
 ↪format(latitude, longitude))
```

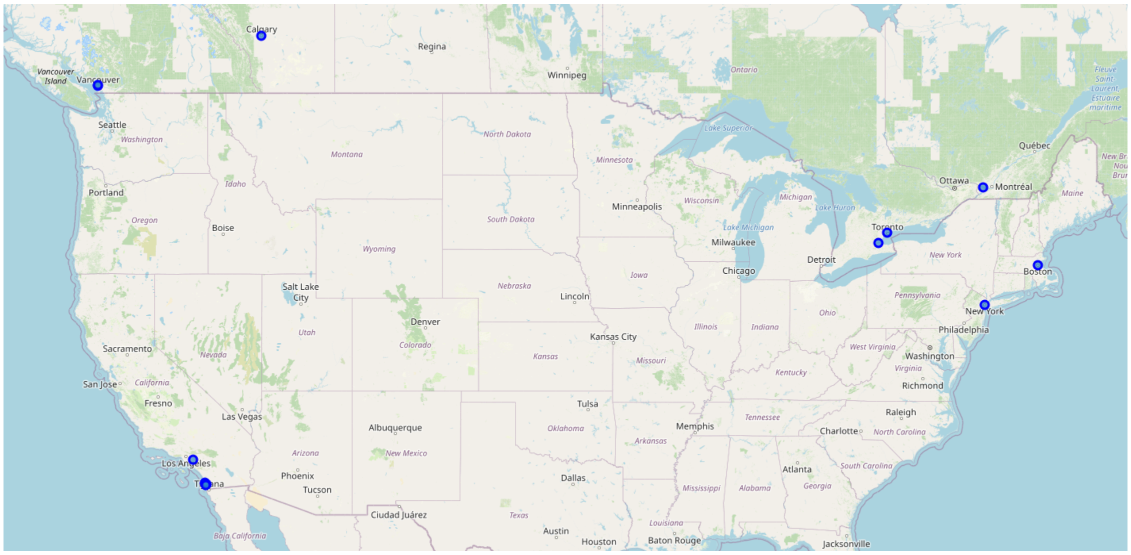
```
The geograpical coordinate of North America are 51.0000002, -109.0.
```

**Create a map of North America with the institutes superimposed on top**

[23]:
```
courses_map = folium.Map(location=[latitude, longitude], zoom_start=4)
```

```
for lat, lng, institute, country in zip(df_institutes['latitude'],␣
 ↪df_institutes['longitude'], df_institutes['InstituteName'],␣
 ↪df_institutes['Country']):
    label = institute
    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(courses_map)

# display map
courses_map
#Image(filename='mapInstitutes.png')
```

[23]:



### Define Foursquare Credentials and Version

Next, we are going to start utilizing the Foursquare API to explore the neighborhoods and segment them.

[24]:

```
CLIENT_ID = 'IJFZODC4A2GVGVYMK0QCF413Q114Y1GSJX3KP0T44C4C3JGI' # your␣
 ↪Foursquare ID
CLIENT_SECRET = 'E50AREFRWDXT1FY3SAYQJZEAGNX4S4VUTI01N4WNELI4HF3J' # your␣
 ↪Foursquare Secret
VERSION = '20180605' # Foursquare API version
```

```
LIMIT = 100 # A default Foursquare API limit value

print('Your credentails:')
print('CLIENT_ID: ' + CLIENT_ID)
print('CLIENT_SECRET:' + CLIENT_SECRET)
```

```
Your credentails:
CLIENT_ID: IJFZ0DC4A2GVGVYMK0QCF413Q114Y1GSJX3KP0T44C4C3JGI
CLIENT_SECRET:E50AREFRWDXT1FY3SAYQJZEAGNX4S4VUTI01N4WNELI4HF3J
```

**Let's create a function to get the data from the venues of the entire list of institutes**

[25]:

```
def getNearbyVenues(names, latitudes, longitudes, radius=500):

    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print(name)

        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/explore?
 ↪&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            LIMIT)

        # make the GET request
        results = requests.get(url).json()["response"]['groups'][0]['items']

        # return only relevant information for each nearby venue
        venues_list.append([(
            name,
            lat,
            lng,
            v['venue']['name'],
            v['venue']['location']['lat'],
            v['venue']['location']['lng'],
            v['venue']['categories'][0]['name']) for v in results])

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item␣
 ↪in venue_list])
    nearby_venues.columns = ['InstituteName',
                  'InstituteLatitude',
                  'InstituteLongitude',
                  'Venue',
                  'VenueLatitude',
                  'VenueLongitude',
                  'VenueCategory']

    return(nearby_venues)
```

**Let's run the above function on each institute and create a new dataframe called institutes_venues**

[26]:

```
# type your answer here
institutes_venues = getNearbyVenues(names=df_institutes['InstituteName'],
                                   latitudes=df_institutes['latitude'],
                                   longitudes=df_institutes['longitude']
                                  )
```

```
ANNE'S Language House
Metropolitan College
Language Studies International (LSI): Toronto
Language Studies International (LSI): Vancouver
International Language Academy of Canada Vancouver
Bouchereau Lingua International
Connect English Language Institute- San Diego (Pacific Beach Campus)
Connect English- La Jolla
Connect English Language Institute- Mission Valley
American English Language School
Language Studies International (LSI): Boston
Language Studies International (LSI): New York
Language Studies International (LSI): San Diego
```

**Let's check the size of the resulting dataframe**

[27]:

```
print(institutes_venues.shape)
institutes_venues.head()
```

```
(515, 7)
```

[27]:

```
            InstituteName  InstituteLatitude  InstituteLongitude  \
0  ANNE'S Language House           51.04715          -114.06334
1  ANNE'S Language House           51.04715          -114.06334
2  ANNE'S Language House           51.04715          -114.06334
3  ANNE'S Language House           51.04715          -114.06334
4  ANNE'S Language House           51.04715          -114.06334

                            Venue  VenueLatitude  VenueLongitude  \
0           The Palomino Smokehouse      51.046435     -114.063410
1                             Blink      51.045422     -114.063733
2  Phil & Sebastian Coffee Roasters      51.045619     -114.063324
3                 Over Easy Breakfast      51.048561     -114.065917
4               Hyatt Regency Calgary      51.046373     -114.062583
```

```
         VenueCategory
0  American Restaurant
1           Restaurant
2          Coffee Shop
3       Breakfast Spot
4                Hotel
```

Let's check how many venues were returned for each institute

```
[28]: institutes_venues.groupby('InstituteName')['InstituteName'].count()
```

```
[28]: InstituteName
ANNE'S Language House                                                      41
American English Language School                                           21
Bouchereau Lingua International                                             5
Connect English Language Institute- Mission Valley                         16
Connect English Language Institute- San Diego (Pacific Beach Campus)       13
Connect English- La Jolla                                                   6
International Language Academy of Canada Vancouver                         45
Language Studies International (LSI): Boston                               82
Language Studies International (LSI): New York                            100
Language Studies International (LSI): San Diego                            23
Language Studies International (LSI): Toronto                              42
Language Studies International (LSI): Vancouver                            89
Metropolitan College                                                       32
Name: InstituteName, dtype: int64
```

**Let's find out how many unique categories can be curated from all the returned venues**

```
[33]: print('There are {} uniques categories.'.
       ↪format(len(institutes_venues['VenueCategory'].unique())))
```

```
There are 151 uniques categories.
```

## 6 Preproccesing steps: one hot encoding

```
[34]: institutes_onehot = pd.get_dummies(institutes_venues[['VenueCategory']],␣
       ↪prefix="", prefix_sep="")

      # add institute column back to dataframe
      institutes_onehot['InstituteName'] = institutes_venues['InstituteName']

      # move InstituteName column to the first column
      fixed_columns = [institutes_onehot.columns[-1]] + list(institutes_onehot.
       ↪columns[:-1])
      institutes_onehot = institutes_onehot[fixed_columns]
```

```
institutes_onehot.head()
```

[34]:
```
           InstituteName  Accessories Store  American Restaurant  Art Gallery  \
0  ANNE'S Language House                  0                    1            0
1  ANNE'S Language House                  0                    0            0
2  ANNE'S Language House                  0                    0            0
3  ANNE'S Language House                  0                    0            0
4  ANNE'S Language House                  0                    0            0

   Asian Restaurant  Athletics & Sports  Auditorium  Auto Dealership  \
0                 0                   0           0                0
1                 0                   0           0                0
2                 0                   0           0                0
3                 0                   0           0                0
4                 0                   0           0                0

   BBQ Joint  Bakery  …  Toy / Game Store  Train Station  Tree  \
0          0       0  …                 0              0     0
1          0       0  …                 0              0     0
2          0       0  …                 0              0     0
3          0       0  …                 0              0     0
4          0       0  …                 0              0     0

   Vegetarian / Vegan Restaurant  Vietnamese Restaurant  Waterfront  Wine Bar  \
0                              0                      0           0         0
1                              0                      0           0         0
2                              0                      0           0         0
3                              0                      0           0         0
4                              0                      0           0         0

   Wine Shop  Women's Store  Yoga Studio
0          0              0            0
1          0              0            0
2          0              0            0
3          0              0            0
4          0              0            0

[5 rows x 152 columns]
```

**Next, let's group rows by institute and by taking the mean of the frequency of occurrence of each category**

[35]:
```
institutes_grouped = institutes_onehot.groupby('InstituteName').mean().
 ↪reset_index()
institutes_grouped
```

```
[35]:                                         InstituteName  Accessories Store  \
      0                               ANNE'S Language House               0.00
      1                    American English Language School               0.00
      2                     Bouchereau Lingua International               0.00
      3   Connect English Language Institute- Mission Va…               0.00
      4   Connect English Language Institute- San Diego …               0.00
      5                           Connect English- La Jolla               0.00
      6   International Language Academy of Canada Vanco…               0.00
      7        Language Studies International (LSI): Boston               0.00
      8      Language Studies International (LSI): New York               0.01
      9     Language Studies International (LSI): San Diego               0.00
      10      Language Studies International (LSI): Toronto               0.00
      11    Language Studies International (LSI): Vancouver               0.00
      12                               Metropolitan College               0.00

          American Restaurant  Art Gallery  Asian Restaurant  Athletics & Sports  \
      0              0.048780     0.024390          0.000000             0.00000
      1              0.047619     0.000000          0.000000             0.00000
      2              0.000000     0.000000          0.000000             0.00000
      3              0.000000     0.000000          0.000000             0.00000
      4              0.000000     0.000000          0.000000             0.00000
      5              0.000000     0.000000          0.000000             0.00000
      6              0.066667     0.000000          0.000000             0.00000
      7              0.024390     0.000000          0.060976             0.00000
      8              0.020000     0.000000          0.000000             0.00000
      9              0.043478     0.000000          0.000000             0.00000
      10             0.023810     0.000000          0.000000             0.02381
      11             0.000000     0.011236          0.000000             0.00000
      12             0.000000     0.000000          0.031250             0.00000

          Auditorium  Auto Dealership  BBQ Joint    Bakery  …  Toy / Game Store  \
      0         0.00           0.0000    0.00000  0.000000  …          0.000000
      1         0.00           0.0000    0.00000  0.000000  …          0.000000
      2         0.00           0.0000    0.00000  0.000000  …          0.000000
      3         0.00           0.0625    0.00000  0.000000  …          0.000000
      4         0.00           0.0000    0.00000  0.000000  …          0.000000
      5         0.00           0.0000    0.00000  0.000000  …          0.000000
      6         0.00           0.0000    0.00000  0.022222  …          0.000000
      7         0.00           0.0000    0.00000  0.097561  …          0.000000
      8         0.01           0.0000    0.01000  0.000000  …          0.000000
      9         0.00           0.0000    0.00000  0.000000  …          0.000000
      10        0.00           0.0000    0.02381  0.023810  …          0.000000
      11        0.00           0.0000    0.00000  0.044944  …          0.022472
      12        0.00           0.0000    0.00000  0.000000  …          0.000000

          Train Station  Tree  Vegetarian / Vegan Restaurant  Vietnamese Restaurant  \
      0          0.0000  0.00                        0.00000               0.024390
```

```
1          0.0000  0.00                        0.00000               0.000000
2          0.0000  0.00                        0.00000               0.000000
3          0.0625  0.00                        0.00000               0.000000
4          0.0000  0.00                        0.00000               0.000000
5          0.0000  0.00                        0.00000               0.000000
6          0.0000  0.00                        0.00000               0.000000
7          0.0000  0.00                        0.02439               0.012195
8          0.0000  0.01                        0.00000               0.000000
9          0.0000  0.00                        0.00000               0.000000
10         0.0000  0.00                        0.00000               0.000000
11         0.0000  0.00                        0.00000               0.000000
12         0.0000  0.00                        0.00000               0.000000

    Waterfront  Wine Bar  Wine Shop  Women's Store  Yoga Studio
0     0.000000   0.00000   0.000000       0.000000     0.000000
1     0.000000   0.00000   0.000000       0.047619     0.047619
2     0.000000   0.00000   0.000000       0.000000     0.000000
3     0.000000   0.00000   0.000000       0.000000     0.000000
4     0.000000   0.00000   0.000000       0.000000     0.000000
5     0.000000   0.00000   0.000000       0.000000     0.000000
6     0.022222   0.00000   0.000000       0.000000     0.000000
7     0.000000   0.00000   0.000000       0.000000     0.012195
8     0.000000   0.00000   0.030000       0.020000     0.000000
9     0.000000   0.00000   0.000000       0.000000     0.000000
10    0.000000   0.02381   0.000000       0.000000     0.023810
11    0.000000   0.00000   0.011236       0.000000     0.011236
12    0.000000   0.00000   0.000000       0.000000     0.000000

[13 rows x 152 columns]
```

**Let's print each institute along with the top 5 most common venues**

```
[36]: num_top_venues = 5

for institute in institutes_grouped['InstituteName']:
    print("----"+institute+"----")
    temp = institutes_grouped[institutes_grouped['InstituteName'] == institute].
 ↪T.reset_index()
    temp.columns = ['venue','freq']
    temp = temp.iloc[1:]
    temp['freq'] = temp['freq'].astype(float)
    temp = temp.round({'freq': 2})
    print(temp.sort_values('freq', ascending=False).reset_index(drop=True).
 ↪head(num_top_venues))
    print('\n')
```

```
----ANNE'S Language House----
                   venue  freq
```

```
0            Restaurant  0.12
1                 Hotel  0.10
2  Performing Arts Venue  0.07
3            Steakhouse  0.05
4           Coffee Shop  0.05


----American English Language School----
          venue  freq
0   Coffee Shop  0.10
1   Yoga Studio  0.05
2  Burger Joint  0.05
3    Food Truck  0.05
4   Gas Station  0.05


----Bouchereau Lingua International----
                        venue  freq
0                Home Service   0.2
1                       Diner   0.2
2              Chocolate Shop   0.2
3                 Supermarket   0.2
4  Construction & Landscaping   0.2


----Connect English Language Institute- Mission Valley----
                   venue  freq
0                    Pub  0.06
1    Martial Arts School  0.06
2            Flower Shop  0.06
3  Performing Arts Venue  0.06
4    Rental Car Location  0.06


----Connect English Language Institute- San Diego (Pacific Beach Campus)----
               venue  freq
0       Intersection  0.15
1  Convenience Store  0.15
2              Hotel  0.15
3    Thai Restaurant  0.08
4          Nightclub  0.08


----Connect English- La Jolla----
                 venue  freq
0                 Park  0.33
1       Scenic Lookout  0.33
2          Golf Course  0.17
```

```
3  Sporting Goods Shop  0.17
4         Neighborhood  0.00


----International Language Academy of Canada Vancouver----
                 venue  freq
0                Hotel  0.11
1  American Restaurant  0.07
2           Restaurant  0.07
3       Cosmetics Shop  0.04
4   Miscellaneous Shop  0.04


----Language Studies International (LSI): Boston----
                venue  freq
0              Bakery  0.10
1  Chinese Restaurant  0.10
2    Asian Restaurant  0.06
3         Coffee Shop  0.05
4          Food Truck  0.05


----Language Studies International (LSI): New York----
           venue  freq
0           Park  0.08
1    Pizza Place  0.06
2    Coffee Shop  0.04
3  Memorial Site  0.04
4          Hotel  0.04


----Language Studies International (LSI): San Diego----
                       venue  freq
0                       Park  0.09
1                        Spa  0.09
2                      Hotel  0.09
3  Middle Eastern Restaurant  0.04
4                   Boutique  0.04


----Language Studies International (LSI): Toronto----
                venue  freq
0  Italian Restaurant  0.07
1         Coffee Shop  0.07
2                Café  0.05
3                 Spa  0.05
4                Park  0.05
```

```
----Language Studies International (LSI): Vancouver----
                 venue  freq
0                Hotel  0.06
1   Mexican Restaurant  0.04
2               Bakery  0.04
3  Japanese Restaurant  0.04
4         Concert Hall  0.03


----Metropolitan College----
                venue  freq
0                 Pub  0.22
1         Coffee Shop  0.09
2  Italian Restaurant  0.06
3                Park  0.06
4      Sandwich Place  0.06
```

### Let's put that into a *pandas* dataframe

First, let's write a function to sort the venues in descending order.

[37]:

```
def return_most_common_venues(row, num_top_venues):
    row_categories = row.iloc[1:]
    row_categories_sorted = row_categories.sort_values(ascending=False)

    return row_categories_sorted.index.values[0:num_top_venues]
```

Now let's create the new dataframe and display the top 10 venues for each neighborhood.

[38]:

```
num_top_venues = 5

indicators = ['st', 'nd', 'rd']

# create columns according to number of top venues
columns = ['InstituteName']
for ind in np.arange(num_top_venues):
    try:
        columns.append('{}{} Most Common Venue'.format(ind+1, indicators[ind]))
    except:
        columns.append('{}th Most Common Venue'.format(ind+1))

# create a new dataframe
institutes_venues_sorted = pd.DataFrame(columns=columns)
institutes_venues_sorted['InstituteName'] = institutes_grouped['InstituteName']

for ind in np.arange(institutes_grouped.shape[0]):
    institutes_venues_sorted.iloc[ind, 1:] =␣
↪return_most_common_venues(institutes_grouped.iloc[ind, :], num_top_venues)

institutes_venues_sorted
```

[38]:

```
                                        InstituteName 1st Most Common Venue  \
0                               ANNE'S Language House            Restaurant
1                    American English Language School           Coffee Shop
2                     Bouchereau Lingua International                 Diner
3   Connect English Language Institute- Mission Va…   Martial Arts School
4   Connect English Language Institute- San Diego …     Convenience Store
5                           Connect English- La Jolla                  Park
6   International Language Academy of Canada Vanco…                 Hotel
7        Language Studies International (LSI): Boston                Bakery
8      Language Studies International (LSI): New York                  Park
9     Language Studies International (LSI): San Diego                 Hotel
10      Language Studies International (LSI): Toronto    Italian Restaurant
11    Language Studies International (LSI): Vancouver                 Hotel
12                               Metropolitan College                   Pub

   2nd Most Common Venue  3rd Most Common Venue 4th Most Common Venue  \
0                  Hotel  Performing Arts Venue           Coffee Shop
1            Yoga Studio   Gym / Fitness Center            Food Truck
2           Home Service            Supermarket        Chocolate Shop
3                    Gym           Climbing Gym        Clothing Store
4                  Hotel           Intersection       Thai Restaurant
5         Scenic Lookout    Sporting Goods Shop           Golf Course
6             Restaurant    American Restaurant        Cosmetics Shop
7     Chinese Restaurant       Asian Restaurant           Coffee Shop
8            Pizza Place            Coffee Shop         Memorial Site
9                   Park                    Spa              Boutique
10           Coffee Shop                   Park                   Spa
11                Bakery    Japanese Restaurant    Mexican Restaurant
12           Coffee Shop         Sandwich Place    Italian Restaurant

         5th Most Common Venue
0                      Theater
1                  Gas Station
2   Construction & Landscaping
3          Rental Car Location
4         Marijuana Dispensary
5                  Flower Shop
6                   Steakhouse
7                   Food Truck
8                        Hotel
9                Deli / Bodega
10                        Café
```

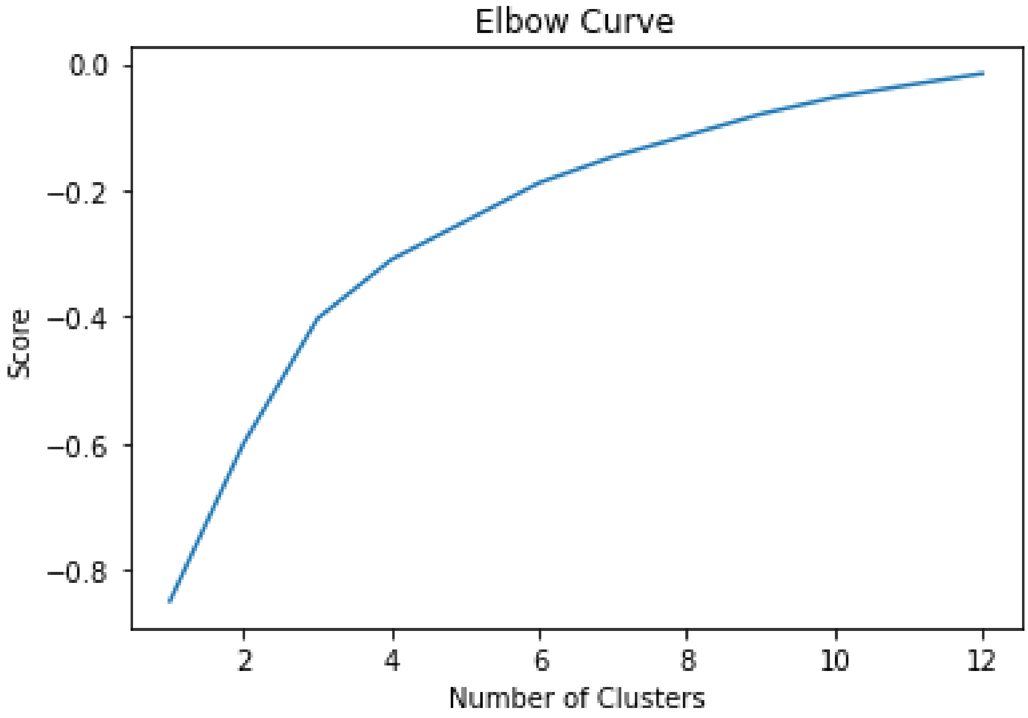
```
11            French Restaurant
12                         Park
```

## 7 Methodology: K-means

Run *k*-means to cluster the institutes into clusters.

[39]:

```
# Create range of K values

X = institutes_grouped_clustering = institutes_grouped.drop('InstituteName', 1)

Nc = range(1, 13)

#Create instances
kmeans = [KMeans(n_clusters=i, n_init = 12) for i in Nc]

#Fit the model and get scores for different K values
score = [kmeans[i].fit(X).score(X) for i in range(len(kmeans))]

#Plot the Elbow curve
plt.plot(Nc,score)
plt.xlabel('Number of Clusters')
plt.ylabel('Score')
plt.title('Elbow Curve')
plt.show()
```



[40]:

```
# Let's select k as 3 because is where an elbow can be identified in the above␣
 ↪graph
kclusters = 3

institutes_grouped_clustering = institutes_grouped.drop('InstituteName', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).
 ↪fit(institutes_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

[40]:

```
array([0, 0, 1, 0, 0, 2, 0, 0, 0, 0])
```

Let's create a new dataframe that includes the cluster as well as the top 10 venues for each institute.

[44]:

```
# add clustering labels
institutes_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)

institutes_merged = df_institutes

# merge institutes_grouped with manhattan_data to add latitude/longitude for␣
 ↪each neighborhood
institutes_merged = institutes_merged.join(institutes_venues_sorted.
 ↪set_index('InstituteName'), on='InstituteName')

institutes_merged.drop(['index'], axis=1, inplace=True) # check the last␣
 ↪columns!
institutes_merged
```

[44]:

```
                                         InstituteName   latitude   longitude  \
0                              ANNE'S Language House  51.047150 -114.063340
1                                Metropolitan College  43.252430  -79.871300
2        Language Studies International (LSI): Toronto  43.678510  -79.389730
3      Language Studies International (LSI): Vancouver  49.280040 -123.124910
4    International Language Academy of Canada Vanco…  49.288360 -123.122590
5                    Bouchereau Lingua International  45.466221  -74.077549
6    Connect English Language Institute- San Diego …  32.803720 -117.214110
7                             Connect English- La Jolla  32.852520 -117.186220
8    Connect English Language Institute- Mission Va…  32.780980 -117.096280
9                        American English Language School  33.909860 -117.854170
10        Language Studies International (LSI): Boston  42.350720  -71.058260
11      Language Studies International (LSI): New York  40.708630  -74.014660
12     Language Studies International (LSI): San Diego  32.723270 -117.160510
```

```
   Country  Cluster Labels 1st Most Common Venue 2nd Most Common Venue  \
0   Canada               0            Restaurant                 Hotel
1   Canada               0                   Pub           Coffee Shop
2   Canada               0    Italian Restaurant           Coffee Shop
3   Canada               0                 Hotel                Bakery
4   Canada               0                 Hotel            Restaurant
5   Canada               1                 Diner          Home Service
6      USA               0     Convenience Store                 Hotel
7      USA               2                  Park        Scenic Lookout
8      USA               0   Martial Arts School                   Gym
9      USA               0           Coffee Shop           Yoga Studio
10     USA               0                Bakery    Chinese Restaurant
11     USA               0                  Park           Pizza Place
12     USA               0                 Hotel                  Park

    3rd Most Common Venue 4th Most Common Venue       5th Most Common Venue
0   Performing Arts Venue           Coffee Shop                     Theater
1          Sandwich Place    Italian Restaurant                        Park
2                    Park                   Spa                        Café
3     Japanese Restaurant    Mexican Restaurant           French Restaurant
4     American Restaurant        Cosmetics Shop                  Steakhouse
5             Supermarket        Chocolate Shop  Construction & Landscaping
6            Intersection       Thai Restaurant        Marijuana Dispensary
7     Sporting Goods Shop           Golf Course                 Flower Shop
8            Climbing Gym        Clothing Store         Rental Car Location
9    Gym / Fitness Center            Food Truck                 Gas Station
10       Asian Restaurant           Coffee Shop                  Food Truck
11            Coffee Shop         Memorial Site                       Hotel
12                    Spa              Boutique               Deli / Bodega
```

Finally, let's visualize the resulting clusters

[45]:

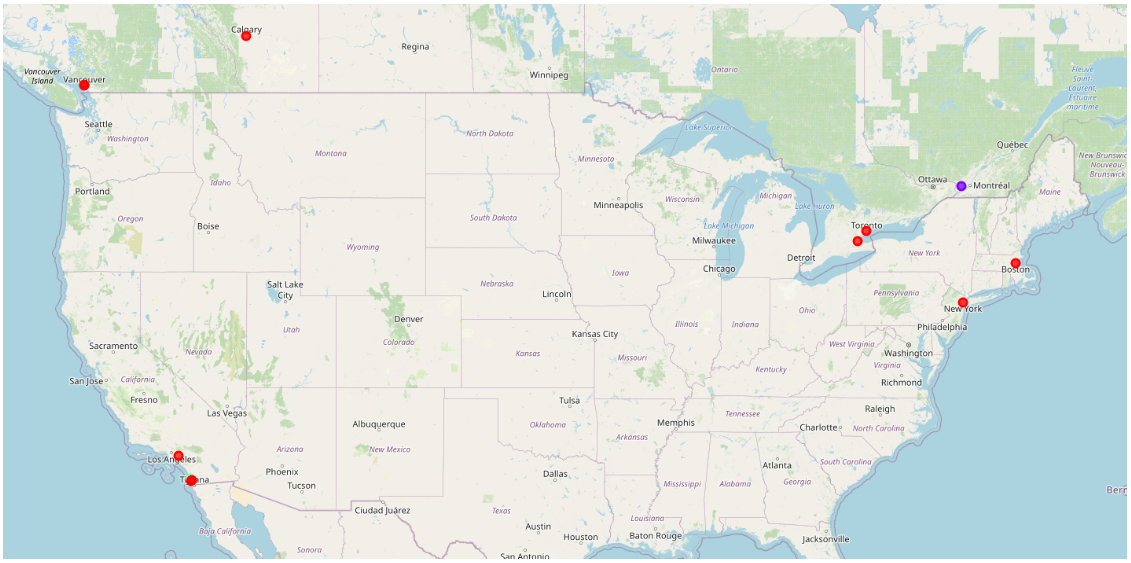
```
institutes_merged["latitude"] = institutes_merged["latitude"].astype("float")
institutes_merged["longitude"] = institutes_merged["longitude"].astype("float")
```

[46]:

```
# create map
map_clusters = folium.Map(location=[latitude, longitude], zoom_start=4)

# set color scheme for the clusters
x = np.arange(kclusters)
ys = [i + x + (i*x)**2 for i in range(kclusters)]
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, poi, cluster in zip(institutes_merged['latitude'],␣
↪institutes_merged['longitude'], institutes_merged['InstituteName'],␣
↪institutes_merged['Cluster Labels']):
    label = folium.Popup(str(poi) + ' 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
#Image(filename='mapClusters.png')
```

[46]:



# 8 Results

Now, is possible to examine each cluster and determine the discriminating venue categories that distinguish each cluster.

[47]:
```
institutes_merged.loc[institutes_merged['Cluster Labels'] == 0,␣
↪institutes_merged.columns[[1] + list(range(5, institutes_merged.shape[1]))]]
```

[47]:
```
   latitude 1st Most Common Venue 2nd Most Common Venue  \
0  51.04715            Restaurant                 Hotel
1  43.25243                   Pub           Coffee Shop
2  43.67851    Italian Restaurant           Coffee Shop
```

```
3   49.28004                 Hotel                Bakery
4   49.28836                 Hotel            Restaurant
6   32.80372     Convenience Store                 Hotel
8   32.78098   Martial Arts School                   Gym
9   33.90986           Coffee Shop           Yoga Studio
10  42.35072                Bakery    Chinese Restaurant
11  40.70863                  Park           Pizza Place
12  32.72327                 Hotel                  Park

    3rd Most Common Venue 4th Most Common Venue 5th Most Common Venue
0   Performing Arts Venue           Coffee Shop               Theater
1          Sandwich Place    Italian Restaurant                  Park
2                    Park                   Spa                  Café
3     Japanese Restaurant    Mexican Restaurant     French Restaurant
4     American Restaurant        Cosmetics Shop            Steakhouse
6            Intersection       Thai Restaurant  Marijuana Dispensary
8            Climbing Gym        Clothing Store   Rental Car Location
9    Gym / Fitness Center            Food Truck           Gas Station
10       Asian Restaurant           Coffee Shop            Food Truck
11            Coffee Shop         Memorial Site                 Hotel
12                    Spa              Boutique         Deli / Bodega
```

[48]:

```
institutes_merged.loc[institutes_merged['Cluster Labels'] == 1,␣
↪institutes_merged.columns[[1] + list(range(5, institutes_merged.shape[1]))]]
```

[48]:

```
    latitude 1st Most Common Venue 2nd Most Common Venue  \
5  45.466221                 Diner          Home Service

  3rd Most Common Venue 4th Most Common Venue       5th Most Common Venue
5           Supermarket        Chocolate Shop  Construction & Landscaping
```

[49]:

```
institutes_merged.loc[institutes_merged['Cluster Labels'] == 2,␣
↪institutes_merged.columns[[1] + list(range(5, institutes_merged.shape[1]))]]
```

[49]:

```
   latitude 1st Most Common Venue 2nd Most Common Venue 3rd Most Common Venue  \
7  32.85252                  Park        Scenic Lookout   Sporting Goods Shop

  4th Most Common Venue 5th Most Common Venue
7           Golf Course           Flower Shop
```

The 3 clusters according to the most common categories of venues for each group could be classified as:

| Name | Cluster |
|---|---|
| LEISURE, STUDENTS, TOURISTS | 0 |
| FAMILIAR | 1 |
| NATURE LOVERS | 2 |

Now that each institute has its own cluster, would be great to have the information about the courses that the institutes offers with their respective prices

This would help in some way to decide which course could be the best fit taking into account the venues, type of course and prices. The decision is up to the "new student"

```
[50]: institutes_all = df_courses.drop(columns=['latitude', 'longitude', 'Country',␣
       ↪'urlcourse', 'urlinstitute']) # Let's drop some columns because they are␣
       ↪present in the table that is going to be merged with

      institutes_all = institutes_merged.join(institutes_all.
       ↪set_index('InstituteName'), on='InstituteName').drop(columns=['latitude',␣
       ↪'longitude']).reset_index().drop(columns=['index']) # Here is where the␣
       ↪merge is done

      institutes_all['PriceTo']=np.
       ↪where(institutes_all['PriceTo']=='None',institutes_all['PriceFrom'],␣
       ↪institutes_all['PriceTo']) # Let¿s replace the none values of the PriceTo␣
       ↪column with the PriceFrom column
```

```
[51]: institutes_all["PriceTo"] = institutes_all["PriceTo"].astype("float")
      institutes_all["PriceFrom"] = institutes_all["PriceFrom"].astype("float")

      institutes_all
```

```
[51]:                                         InstituteName Country  Cluster Labels  \
      0                              ANNE'S Langauge House  Canada               0
      1                              ANNE'S Langauge House  Canada               0
      2                               Metropolitan College  Canada               0
      3                               Metropolitan College  Canada               0
      4        Language Studies International (LSI): Toronto  Canada               0
      5        Language Studies International (LSI): Toronto  Canada               0
      6        Language Studies International (LSI): Toronto  Canada               0
      7        Language Studies International (LSI): Toronto  Canada               0
      8      Language Studies International (LSI): Vancouver  Canada               0
      9      Language Studies International (LSI): Vancouver  Canada               0
      10     Language Studies International (LSI): Vancouver  Canada               0
      11     Language Studies International (LSI): Vancouver  Canada               0
      12     Language Studies International (LSI): Vancouver  Canada               0
      13   International Language Academy of Canada Vanco…  Canada               0
      14                       Bouchereau Lingua International  Canada               1
      15   Connect English Language Institute- San Diego …     USA               0
      16                             Connect English- La Jolla     USA               2
      17                             Connect English- La Jolla     USA               2
      18   Connect English Language Institute- Mission Va…     USA               0
      19   Connect English Language Institute- Mission Va…     USA               0
      20                        American English Language School     USA               0
      21          Language Studies International (LSI): Boston     USA               0
```

```
22       Language Studies International (LSI): Boston     USA               0
23       Language Studies International (LSI): Boston     USA               0
24       Language Studies International (LSI): Boston     USA               0
25     Language Studies International (LSI): New York     USA               0
26     Language Studies International (LSI): New York     USA               0
27    Language Studies International (LSI): San Diego     USA               0
28    Language Studies International (LSI): San Diego     USA               0
29    Language Studies International (LSI): San Diego     USA               0

   1st Most Common Venue 2nd Most Common Venue  3rd Most Common Venue  \
0             Restaurant                 Hotel  Performing Arts Venue
1             Restaurant                 Hotel  Performing Arts Venue
2                    Pub           Coffee Shop         Sandwich Place
3                    Pub           Coffee Shop         Sandwich Place
4     Italian Restaurant           Coffee Shop                   Park
5     Italian Restaurant           Coffee Shop                   Park
6     Italian Restaurant           Coffee Shop                   Park
7     Italian Restaurant           Coffee Shop                   Park
8                  Hotel                Bakery    Japanese Restaurant
9                  Hotel                Bakery    Japanese Restaurant
10                 Hotel                Bakery    Japanese Restaurant
11                 Hotel                Bakery    Japanese Restaurant
12                 Hotel                Bakery    Japanese Restaurant
13                 Hotel            Restaurant    American Restaurant
14                 Diner          Home Service            Supermarket
15     Convenience Store                 Hotel           Intersection
16                  Park        Scenic Lookout    Sporting Goods Shop
17                  Park        Scenic Lookout    Sporting Goods Shop
18   Martial Arts School                   Gym           Climbing Gym
19   Martial Arts School                   Gym           Climbing Gym
20           Coffee Shop           Yoga Studio   Gym / Fitness Center
21                Bakery    Chinese Restaurant       Asian Restaurant
22                Bakery    Chinese Restaurant       Asian Restaurant
23                Bakery    Chinese Restaurant       Asian Restaurant
24                Bakery    Chinese Restaurant       Asian Restaurant
25                  Park           Pizza Place            Coffee Shop
26                  Park           Pizza Place            Coffee Shop
27                 Hotel                  Park                    Spa
28                 Hotel                  Park                    Spa
29                 Hotel                  Park                    Spa

   4th Most Common Venue       5th Most Common Venue            Location  \
0            Coffee Shop                     Theater     Calgary, Canada
1            Coffee Shop                     Theater     Calgary, Canada
2     Italian Restaurant                        Park    Hamilton, Canada
3     Italian Restaurant                        Park    Hamilton, Canada
4                    Spa                        Café     Toronto, Canada
```

```
5                    Spa                        Café     Toronto, Canada
6                    Spa                        Café     Toronto, Canada
7                    Spa                        Café     Toronto, Canada
8     Mexican Restaurant           French Restaurant   Vancouver, Canada
9     Mexican Restaurant           French Restaurant   Vancouver, Canada
10    Mexican Restaurant           French Restaurant   Vancouver, Canada
11    Mexican Restaurant           French Restaurant   Vancouver, Canada
12    Mexican Restaurant           French Restaurant   Vancouver, Canada
13        Cosmetics Shop                  Steakhouse   Vancouver, Canada
14        Chocolate Shop  Construction & Landscaping    Montreal, Canada
15       Thai Restaurant        Marijuana Dispensary      San Diego, USA
16           Golf Course                 Flower Shop      San Diego, USA
17           Golf Course                 Flower Shop      San Diego, USA
18        Clothing Store         Rental Car Location      San Diego, USA
19        Clothing Store         Rental Car Location      San Diego, USA
20            Food Truck                 Gas Station    Los Angeles, USA
21           Coffee Shop                  Food Truck         Boston, USA
22           Coffee Shop                  Food Truck         Boston, USA
23           Coffee Shop                  Food Truck         Boston, USA
24           Coffee Shop                  Food Truck         Boston, USA
25         Memorial Site                       Hotel  New York City, USA
26         Memorial Site                       Hotel  New York City, USA
27              Boutique               Deli / Bodega      San Diego, USA
28              Boutique               Deli / Bodega      San Diego, USA
29              Boutique               Deli / Bodega      San Diego, USA

                                           CourseName  PriceFrom  PriceTo  \
0                           General English 20 (GE20)      354.0    687.0   
1                              General English (GE25)      384.0    717.0   
2                                 ESL Private Lessons      384.0    619.0   
3                                   Intensive English      290.0    525.0   
4                                        Intensive 30      495.0    823.0   
5                                          General 20      431.0    759.0   
6   Club 40+ (20 lessons per week plus afternoon a…      708.0   1037.0   
7                                        Afternoon 10      286.0    614.0   
8                                          General 20      431.0    759.0   
9                                        Intensive 30      495.0    823.0   
10                                       Intensive 25      465.0    794.0   
11  Club 40+ (20 lessons per week plus afternoon a…      708.0   1037.0   
12                                       Afternoon 10      286.0    614.0   
13                        General English - Intensive      503.0    870.0   
14                            Super Intensive English      533.0    887.0   
15                                        English Max      390.0    390.0   
16                 English Focus Course (12 hrs/week)      365.0    365.0   
17                   English Max Course (18 hrs/week)      390.0    390.0   
18                   English Max Course (18 hrs/week)      390.0    390.0   
19                 English Focus Course (12 hrs/week)      315.0    315.0   
```

```
20                     ESL Program (12 Weeks Minimum)      335.0    335.0
21                                       Intensive 30      645.0    960.0
22  Club 40+ (20 lessons per week plus afternoon a…      840.0   1155.0
23                    One-to-One (5 lessons per week)      575.0    890.0
24                                         General 20      535.0    850.0
25                   One-to-One (20 lessons per week)     1955.0   2280.0
26                   One-to-One (10 lessons per week)     1055.0   1380.0
27                                       Intensive 30      645.0    950.0
28  Club 40+ (20 lessons per week plus afternoon a…      840.0   1240.0
29                                         General 20      535.0    935.0

                                              Address
0   101 6th Avenue S.W., Suite 1250, Calgary, Albe…
1   101 6th Avenue S.W., Suite 1250, Calgary, Albe…
2   146 James Street South, Hamilton, Ontario,, Ha…
3   146 James Street South, Hamilton, Ontario,, Ha…
4   1055 Yonge Street, Suite #210, Toronto, ON M4W…
5   1055 Yonge Street, Suite #210, Toronto, ON M4W…
6   1055 Yonge Street, Suite #210, Toronto, ON M4W…
7   1055 Yonge Street, Suite #210, Toronto, ON M4W…
8   101-808 Nelson Street, Vancouver, BC V6Z 2H2, …
9   101-808 Nelson Street, Vancouver, BC V6Z 2H2, …
10  101-808 Nelson Street, Vancouver, BC V6Z 2H2, …
11  101-808 Nelson Street, Vancouver, BC V6Z 2H2, …
12  101-808 Nelson Street, Vancouver, BC V6Z 2H2, …
13  1199 West Pender Street, Vancouver, British Co…
14  70 Notre Dame West, Suite 400, Montréal, Quebe…
15  3565 Del Rey St., Suite 300, San Diego, Califo…
16  5090 Shoreham Place, Suite 206, San Diego, Cal…
17  5090 Shoreham Place, Suite 206, San Diego, Cal…
18  4560 Alvarado Canyon Road, Suite 2B, San Diego…
19  4560 Alvarado Canyon Road, Suite 2B, San Diego…
20  3230 E. Imperial Hwy, Suite 301, Brea, Califor…
21            105 Beach Street, Boston, MA 02111, USA
22            105 Beach Street, Boston, MA 02111, USA
23            105 Beach Street, Boston, MA 02111, USA
24            105 Beach Street, Boston, MA 02111, USA
25  40 Rector Street, 10th Floor, Suite 1000, New …
26  40 Rector Street, 10th Floor, Suite 1000, New …
27          1706 5th Avenue, San Diego, CA 92101, USA
28          1706 5th Avenue, San Diego, CA 92101, USA
29          1706 5th Avenue, San Diego, CA 92101, USA
```

### Let's calculate the average price from and to and the general as well

[52]:

```
institutes_prices = institutes_all.groupby(['InstituteName'])['Cluster Labels',␣
↪'PriceFrom', 'PriceTo'].mean().reset_index() #Let's calculate the mean of␣
↪the prices by column
institutes_prices['PriceAvg'] = (institutes_prices['PriceFrom'] +␣
↪institutes_prices['PriceTo']) / 2 #Let's calculate the mean of the prices by␣
↪row

institutes_prices = institutes_prices.sort_values('PriceAvg', ascending=False).
↪reset_index(drop=True)

institutes_prices['Cluster Labels'] = np.where(institutes_prices['Cluster␣
↪Labels'] == 0,'LEISURE, STUDENTS, TOURISTS', np.
↪where(institutes_prices['Cluster Labels'] == 1,'FAMILIAR', 'NATURE LOVERS'))

institutes_prices
```

```
<ipython-input-52-a505625daf17>:1: FutureWarning: Indexing with multiple keys
(implicitly converted to a tuple of keys) will be deprecated, use a list
instead.
  institutes_prices = institutes_all.groupby(['InstituteName'])['Cluster
Labels', 'PriceFrom', 'PriceTo'].mean().reset_index() #Let's calculate the mean
of the prices by column
```

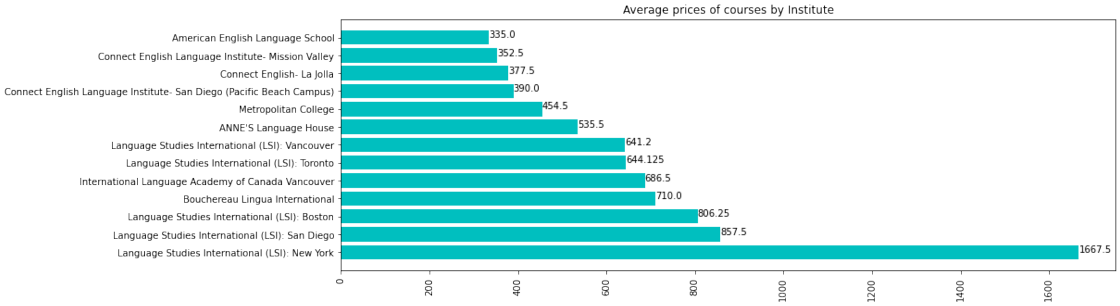
[52]:

```
                                        InstituteName  \
0      Language Studies International (LSI): New York
1     Language Studies International (LSI): San Diego
2        Language Studies International (LSI): Boston
3                     Bouchereau Lingua International
4    International Language Academy of Canada Vanco…
5       Language Studies International (LSI): Toronto
6     Language Studies International (LSI): Vancouver
7                               ANNE'S Language House
8                                Metropolitan College
9    Connect English Language Institute- San Diego …
10                          Connect English- La Jolla
11   Connect English Language Institute- Mission Va…
12                     American English Language School

                 Cluster Labels    PriceFrom      PriceTo  PriceAvg
0   LEISURE, STUDENTS, TOURISTS  1505.000000  1830.000000  1667.500
1   LEISURE, STUDENTS, TOURISTS   673.333333  1041.666667   857.500
2   LEISURE, STUDENTS, TOURISTS   648.750000   963.750000   806.250
3                      FAMILIAR   533.000000   887.000000   710.000
4   LEISURE, STUDENTS, TOURISTS   503.000000   870.000000   686.500
5   LEISURE, STUDENTS, TOURISTS   480.000000   808.250000   644.125
6   LEISURE, STUDENTS, TOURISTS   477.000000   805.400000   641.200
7   LEISURE, STUDENTS, TOURISTS   369.000000   702.000000   535.500
```

```
8   LEISURE, STUDENTS, TOURISTS   337.000000   572.000000   454.500
9   LEISURE, STUDENTS, TOURISTS   390.000000   390.000000   390.000
10                 NATURE LOVERS   377.500000   377.500000   377.500
11  LEISURE, STUDENTS, TOURISTS   352.500000   352.500000   352.500
12  LEISURE, STUDENTS, TOURISTS   335.000000   335.000000   335.000
```

[53]:

```
fig, ax = plt.subplots(figsize=(15, 5))

plt.title('Average prices of courses by Institute')
plt.xticks(rotation=90)

ax.barh(institutes_prices['InstituteName'],institutes_prices['PriceAvg'],␣
 ↪color=['c'])

for index, value in enumerate(institutes_prices['PriceAvg']):
    plt.text(value, index, str(value))

plt.show()
print('Done!')
```



```
Done!
```

## 9 Discussion

This analysis shows that from the 30 courses selected from Canada and USA, 13 institutes are offering them in different locations of the countries. The locations are actually quite far one from each other (except for San Diego, USA where 4 courses are relatively closer). In this sense, takes a lot of importance the decision of selecting which course to take according to the personal "taste"

The venues around the institutes, after getting the data from the Foursquare API, **151** unique categories were identified. It is good to know that the radius that was selected for the API was 500 meters that is a distance considerably closer from the institutes.

The point here is that there are only 3 clusters created and one of them has the majority of institutes. After looking at the categories of the venues from each cluster, the possible classification can be:

- LEISURE, STUDENTS, TOURISTS

- FAMILIAR
- NATURE LOVERS

The variety of categories of venues around the main one **LEISURE, STUDENTS, TOURISTS** let us assume that the public could be really diverse whereas in the other clusters, a more specific "taste" is found. It means, that for people for example that is thinking to travel with their families or that likes a lot the nature could be quite easier to take a decision. However, it is good to highlight that according to the locations of the institutes, there is a big opportunity of improvement of this model in terms of adding different variables that includes for example weather, temperature or cost of living. I would say that this is just a general idea of what to expect around the locations of study.

Finally, the costs should be taken into account because the average of the courses are between 335 USD and 1667 USD which let us think that other variables such as reviews, ratings, comments or duration would be great to include to have an explanation of why the prices are so diverse.

## 10 Conclusion

This can be the first approach and as it is stated above, there is some additional information that can be useful to allow the "new student" to take a better decision. The good point here, is that, we already have a general view of the courses offered by the best institues according to the wbesite https://www.languageinternational.com/ clustered by the categories of venues found around them.

[ ]:

```
```