Final Assignment: Segmenting and Clustering Neighborhoods in Toronto, Canada

SUMMER TRAINING REPORT

Submitted by
Abdelkader Abdelrsoul

Introduction:

Final Assignment: Segmenting and Clustering Neighborhoods in Toronto, Canada

This notebook is to extract the list of postal codes of Toronto from this Wikipedia page and then cluster the neighborhoods located in one of Toronoto's boroughs based on the venues located in each neighborhood

Table of Content

Data where you describe the data that will be used to solve the problem and the source of the data

This notebook is to extract the list of postal codes of Toronto from this Wikipedia page and then cluster the neighborhoods located in one of Toronoto's boroughs based on the venues located in each neighborhood

Project Description

In this project I explore, segment, and cluster the neighborhoods in the city of Toronto. The neighborhood data though is not readily available on the internet.
For the Toronto neighborhood data, a Wikipedia

page: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M exists that has all the information we need to explore and cluster the neighborhoods in Toronto. We will have to scrape the Wikipedia page and wrangle the data, clean it, and then read it into a pandas dataframe so that it is in a structured format.

Once the data is in a structured format, we can start the analysis to explore and cluster the neighborhoods in the city of Toronto.
We build the code to scrape the following Wikipedia

page, https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M, in order to obtain the data that is in the table of postal codes and to transform the data into a pandas dataframe.

The dataframe will consist of three columns: PostalCode, Borough, and Neighborhood. We only process the cells that have an assigned borough. We ignore cells with a borough that is Not assigned. More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, you will notice that M5A is listed twice and has two neighborhoods: Harbourfront and Regent Park.

These two rows will be combined into one row with the neighborhoods separated with a comma. If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough. So for the 9th cell in the table on the Wikipedia page, the value of the Borough and the Neighborhood columns will be Queen's Park.

There are different website scraping libraries and packages in Python. One of the most common packages is BeautifulSoup and we will use it in this project. Package's main documentation

Methodology section :

In this course I have learned about the major steps involved in tackling a data science problem. - The major steps involved in practicing data science, from forming a concrete business or research problem, to collecting and analyzing data, to building a model, and understanding the feedback after model deployment. - How data scientists think!

Open Source tools for Data Science

In this course, I have learned about various open source tools for Data Science. • Skill Network Labs • Jupyter Notebooks • Apache Zeppelin Notebooks • Rstudio IDE • IBM Watson studio

Python for Data Science and AI

In this course I have learned about Python Basics like types, expressions, variables, string operations, lists, tuples, sets, dictionaries, Loops, objects and classes, file handling, pandas and numpy.

Databases and SQL for Data Science

In this course, I have learned about relational database concepts that helps to apply foundational knowledge of the SQL language, performing SQL access in a data science environment. The emphasis in this course is on hands-on and practical learning. I have also created some database instances in the cloud. I have done series of hands-on labs, practice building and running SQL queries in this lab. I have also learned how we can access databases from Jupyter notebooks using SQL and Python.

Data Analysis with Python

In this course I have learned about Importing Datasets, Cleaning the Data , Data frame manipulation, Summarizing the Data. It includes following parts: Data Analysis libraries, use of Pandas, Numpy and Scipy libraries to work with a sample dataset. I have used this library to load, manipulate, analyze, and visualize cool datasets

Data Visualization with Python

This course was all about several data visualization libraries in Python like Matplotlib, Seaborn, and Folium and how we can tell a compelling story by visualizing the data and findings from the data.

Machine Learning with Python

In this course I have learned about some of machine learning topics like supervised and unsupervised learning, classification, clustering and some Python libraries like Sci-kit learn and Scipy.

Applied Data Science Capstone

In this course I have learned about FourSquare API (It is a restful API to retrieve the data about venues in different neighborhoods around the world and I have applied this learnings to complete my Capstone Project