Capstone Project

The Battle of Neighborhoods

(Week 1)

2. Data

A description of the data and how it will be used to solve the problem:

a) Data used for the project are as following:

For New York City:

- New York City dataset:
  https://geo.nyu.edu/catalog/nyu_2451_34572
- New York City latitude and longitude:
  Geolibrary
- Explore neighborhoods in New York City:
  Foursquare API

For Toronto:

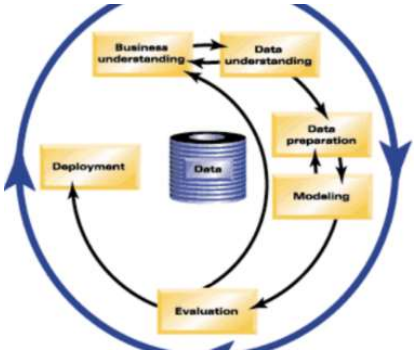
- Toronto Dataset (Toronto's zip codes, boroughs and neighborhood names):
  https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.
- Toronto latitude and longitude information:
  http://cocl.us/Geospatial_data
- Explore neighborhoods in Toronto:
  Foursquare API¶

b) The following tools were used:

- Pandas: Data Analysis
- NumPy: Data handling (vectors)
- JSON: Handling of JSON files
- Geopy: Retrieving location data
- Matplotlib: Data Visualization/Plotting
- Sklearn: Machine Learning; k-Mean calculation
- Folium: Map Rendering

c) Methodology:

We will use for our project some parts of the methodology described in the IBM Data science course as seen below in the figure.



After defining the Business understanding and the problem, we need to understand what data are needed to answer the question and make sure we select the right cohort of data and they are representative to solve the problem. Then the data can be downloaded.

The following data preparation process includes also data transformation and cleaning like renaming of columns, changing datatypes, deleting not required columns.

In our case as a next step data are modeled by combining data and adding latitude as well as longitude values to the respective neighborhoods.
Segmentation of the neighborhoods is the next critical step.
The most important part is to explore each neighborhood using Foursquare to list the venues with name and category per borough.
After analyzing the different neighborhoods also by visualization, clustering using k-means methodology is performed.

In short, the following key steps are done in this project to answer the business problem:

1. Data download
2. Data transformation
3. Adding latitude/longitude
4. Segmentation and Explore neighborhoods
5. Data analyzing
6. Clustering using k-means