IBM DATA SCINCE PROFESSIONAL CERTIFICATE

APPLIED DATA SCIENCE CAPSTONE PROJECT

PART I

AUGUST 16, 2021

THE BATTLE OF NEIGHBORHOODS

RISHI SAXENA

DATA

The most important element of a machine learning model is the quality of the training data. As the famous saying goes, *"Garbage in, garbage out!"*. This is particularly true in the field of machine learning and data science. If the data used to train the model is not diverse enough, or lacks the desired accuracy, the model would underfit and unreservedly fail in delivering factual results. Consequently, if the train data is excessive, the model would overfit the data and be unusable to predict anything outside of the train dataset.

While this program can be easily tailored to work on any combination of two cities across North America, provided the spatial data is available, the two cities used in this initial demonstration of the program are New York and Toronto. In order to create this model, the data should be able to offer a detailed overview of the venues and amenities for every neighborhood in the city. A key player that offers a free to use, publicly available online tool for this service is FourSquare. FourSquare is an American technology company that provides a location tracking platform for a several businesses and consumer products.

The FourSquare API provides a detailed overlook of the amenities and services available at a given location. However, this location has be provided in the form of spatial coordinates, i.e., latitudes and longitudes. This implies that the program would need another source of data that can provide the coordinate for every neighborhood in the cities of interest. Effort is put into ensuring that the spatial data used for this program is accurate and reliable. The two primary sources for data are –

1. New York University – Spatial Data Repository (https://geo.nyu.edu/)
2. Wikipedia (https://wikipedia.org)

Unfortunately, this is the bottleneck for this program. If the coordinate data for a city is unavailable, the program cannot provide the necessary inputs to FourSquare API and fetch the results. Another liability of this program is that publicly available, free data would always have its limitations. On the bright side, FourSquare is a reliable and accurate partner that makes this process viable, despite all challenges.