Capstone Project – Week 1 Battle of Neighborhoods

INTRODUCTION/BUSINESS PROBLEM

In this project, we will explore the option of opening an Indian Restaurant in the city of Calgary. This vibrant multi-cultural city is quit diverse in its population attracting people from a wide variety of ethnic backgrounds. Being Canada's primary oil and gas economic engine, Calgary attracts a diverse population of talent including South Asians often working in engineering or information technology. The city is however still relatively small and underdeveloped compared to metropolitan cities like Toronto, Montreal and Vancouver. Nevertheless, Calgary's population is expected to rise as the city matures and diversifies its economy and demographics. The primary object of this Capstone project is to propose a neighbourhood in Calgary for an individual or a business franchise owner, who is considering opening an Indian Restaurant.

DATA & METHODOLOGY

In first part of this project, we will explore and develop insight on Toronto and its neighbourhoods. We will extract Toronto postal codes and respective neighbourhoods by using Python's Beautifulsoup package on Toronto's Wikipedia page. Neighbourhood names will then be passed into Python's geocoder package to obtain location latitude and longitude coordinates. We will then start exploring Toronto's neighbourhoods such as Thorncliffe Park in East York, which is predominantly known as the immigrant hub of the greater Toronto area. Neighbourhood venues data will be acquired through API connection to well-established location based service, Foursquaures. Similar neighbourhoods will then be clustered based using clustering algorithm such as K-means. To apply machine learning models effectively, we will take mean frequency of occurrence for each venue category for their respective neighbourhoods. A predictive modeling technique such as support vector machines (SVM) will be trained and fitted on the mean frequency dataset to help us predict neighbourhoods with high frequency occurrence for an Indian Restaurant in any city.

In the second part of this project, we will again extract neighbourhoods Calgary's Wikipedia page through Python's beautifulsoup package. Unlike the Toronto dataset, the location data for each neighbourhood will also be extracted from Wikipedia and stored as csv file for further analysis. In order for the SVM model to apply effectively, we will ensure both Toronto and Calgary data frames contain the same venues on which the original model was trained. This is to ensure that the number of features (or venue categories) are identical in both the trained (Toronto) and test (Calgary) data sets. The machine-learning model from Toronto will then be applied to Calgary data set to determine possible neighbourhoods where an Indian Restaurant is likely to succeed.