Capstone Project

The battle of Neighborhoods(week1)

Part 2: Data Description:

2.1 Description:

This project will rely mainly on the publicly available data from Wikipedia as well as Foursquare.

2.1.1 Dataset 1: List of postal codes of Toronto along with the boroughs and neighborhoods

Since we focus on Toronto in this project, we will be looking to procure all the demographic information related to city, including all the boroughs and neighborhoods of Toronto along with their associated zip code.

To do this, we rely on the publicly available Wikipedia page for the same, titled List of postal codes of Canada: M (Link can found here). As can be seen from the description of the page:

"Postal codes beginning with M are located within the city of Toronto in the province of Ontario."

We will be scraping this data from the Wikipedia page with the help of Python's pandas and Wikipedia packages.

2.1.2 Dataset 2: Geographical co-ordinates of the neighborhoods

In order to plot the neighborhoods on the map, we will also be using the geographical co-ordinates of the neighborhoods of Toronto. Although this data can be obtained using the Google Maps Geocoding API, given the unreliability of the package we use data from the following source: http://cocl.us/Geospatial_data

2.1.3 Dataset 3: Data called from through Foursquare API

In this project, we will be leveraging the Foursquare API to obtain the geographical location data of various neighborhoods of Toronto. This data will be used to explore the popular and commonly visited venues in each of the neighborhoods, which will help us to identify the best possible location for our client's Indian restaurant.

We will be identifying top 10 popular venues in each of the areas to satisfy the client's requirement that our locality should not have a restaurant in the top 2 commonly visited venues. Due to the API restrictions set by Foursquare, our search for the number of venues would be limited only to 100.

2.1.4 Dataset 4:Data Pre-processing

After scraping the initial data from Wikipedia, there were some improvements required before it could be used for analysis. Some modifications that were made to it were:

- Dropping all the rows from the derived table where boroughs were not assigned
- Combining different neighborhoods with the same postcode
- For neighborhoods with no name, assigning a borough name to it for the purposes of simplicity

- Concatenating the geospatial co-ordinates obtained from Dataset 2 to this dataframe.