The Battle of The Neighborhoods

Aditya Gouroju

April 2020

1 Introduction

1.1 Background

Many people move from one city to another or maybe from one country to another due to their personal or professional causes.As Different cities in the world are filled with numerous kinds of venues that in turn define the cultures of the cities. A city not only differs from another by means of global positioning it also depends upon the interests of the people living in it .Despite the dissimilarities of being different cities it is possible to group the neighborhoods and segment the neighborhoods with the venue categories of personal choices to decide the appropriate neighborhood while moving to the city

1.2 Problem

Finding identical neighborhoods in different cities in order to help provide a perception of similar neighborhoods which may provide with a great deal of insights in order to make a decision of choosing a neighborhood that is far away, yet somewhat feels like home.

1.3 Interest

People who is moving to far away cities would be very much interested in finding out the most appropriate neighborhood they can move to which is similar to the one in which they are currently residing in

2 Data Acquisition and Cleaning

2.1 Data sources

In this project we are going to see how similar or dissimilar the neighborhoods of Toronto and New York cities are.We can get the data of Neighborhoods and its geographical data i.e, their latitudes and longitudes of Toronto by scraping from here,and of New York from here.we have to get all the data of the neighborhoods venues by Foursquare API's

2.2 Data Cleaning

The first data source in the described link is in .json format.Upon examining the data and further formatting of the .json data finally we can convert it into a dataframe that consists of 4 columns, namely: Borough, Neighborhood, Latitude and Longitude by using Pandas Library. The second data source is a Wikipedia page that contains Postcode of the city of Toronto in a wikitable. we can scrape the page using bs4(Beautiful Soup) library and retrieve the required table as a dataframe using Pandas library .After going through a few more steps, the dataframe was obtained which consists of: PostalCode,Borough and Neighborhood.In this Dataframe the rows with Borough's which are unassigned must be dropped as they will be no use for us.After that if if there are any unassigned neighborhoods we will use them with the same names of their Borough's.

2.3 Feature Selection

Now that we have obtained the different neighborhoods and their respective geometric coordinates for the city of New York and Toronto, To dive further to the problem we are going to need the data of the venues of the respective neighborhoods.For that we are going to need the Foursquare API. Foursquare API provides with an access to an enormous database consisting of venues from all around the world including rich variety of information such as addresses, tips, photos and comments. Having signed up for a Foursquare developer, using the Client ID and Client Secret, it is possible to make API requests in order in order to retrieve venue information. By feeding a function with Neighborhood name and its geometric coordinates, using Foursquare API different venues (Restaurants, Coffee shops, etc) were extracted. After performing One-HotEncoding and grouping together the rows by neighborhoods, the NY dataset and Toronto dataset were combined into a single dataframe in order to perform clustering operation.