The Battle of Neighborhoods: New York vs Toronto

Hanning Gu

1 January 2020

Introduction and Problem

As an international centre of finance, culture, arts and business and the most populous city of their respective countries, both New York and Toronto are recognized as one of the most cosmopolitan and multicultural cities in the world. Therefore, they have attracted tons of tourists and migrants every year. Despite their similarities, their differences have also been discussed. Which city would you consider as a **migrant**? Where city is a better choice to open a restaurant of your own cuisine? With the above questions in mind, I will conduct a comparison study of New York (Manhattan) and Toronto city using the techniques I learnt in the IBM data science course on Coursera. This report can also help **travellers** who are interested in visiting these two cities.

Data acquisition and pre-processing

Information of neighbourhoods of New York and Toronto were directly obtained from https://geo.nyu.edu/catalog/nyu_2451_34572 and https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. Then, data of venues of interests of these neighbourhood areas were scraped from Foursquare location data. The venue data not only contains information such as location (i.e. latitude, longitude) but also the category of the venue, which will provide the most important information on differentiating neighbourhoods.

Firstly, neighbourhood data of New York city was loaded to a table with information such as Borough name, Neighbourhood name and its location (Latitude and Longitude). For the sake of simplicity, I selected suburbs in Manhattan for further investigation. Next, for each neighbourhood in

Manhattan, we scraped data of up to 100 venues within a disk centred at its given coordinate with radius=500 from Foursquare. The data collected also tells which category every venue belongs to. Then I grouped all venue entries by their neighbourhoods to have aggregated information of each neighbourhood.

Secondly, neighbourhood data of City of Toronto were processed in a similar way except that I removed cells with missing ‘Borough’, merged neighbourhoods with same postcode and named neighbourhood as their boroughs if their names are missing. In the end, venues with missing category were removed.

Lastly, the two data frames of two cities were merged into one data frame for machine learning which will be discussed later.