Clustering Neighborhoods for supporting tourists decisions

Thiago F. C. de Oliveira

July, 2020

Abstract

This is a report of the Applied Data Science Capstone, the last course of the IBM Data Science Professional Certificate. I used data from Foursquare API and transitland to cluster Toronto's neighborhoods in order to help tourists to decide which neighborhood to stay during a vacation, based on venues price, mobilitiby and proximity of the tourist attractions.

1 Introduction

1.1 Backgroud

We all know that planning a trip can be a quite difficult task. You have to look for plane tickets, passports, schedule and for a place to stay. We also know that frequently things can get out of control, even though you did the best travel itinerary. And when you loose control, it's possible that you loose a lot of money and time too. And all that was supposed to be about relaxing, becomes a nightmare.

You may have found the cheapest bedroom in Airbnb, or the most confortable hotel. But before you check in, you realize that is far away from everything. Or that all the restaurantes around are too expensive, or even that there isn't a bus stop around and you'll spent a lot of money in cabs.

But what if you can get all the information about the neighborhood you will be staying? How easy is to get the subway, or how many options would you have to eat if you decide to take a walk? Wouldn't it be great to have all neighborhoods of your destination clustered according to its characteriscs?

The objective of this work is to provide a cluster analysis of Toronto's neighborhoods in order to support tourists decision in choosing a place to stay in a vacation travel.

1.2 Problem

Gathering data from Foursquare API, containing venues in the neighborhood, venues category and also venues price. I'll also get the location of all subway and bus stations and tourist attractions. With all this information, I'll apply k-Means clustering and label neighborhoods according to price, venues variability, mobility and proximity to tourist attractions.

1.3 Interest

This work can give insight to any traveler going to Toronto and can be also be interesting to Travel Agencies, improving costumer satisfaction in choosing a place to stay in a vacation. With the right data, this work can be applied to any city.

2 Data

2.1 Data Sources

Three different sources will be required for this project. With all of them gathered together, it will be possible to assemble a model to use in the k-Means clustering algorithm.

First, the list with all postal codes, boroughs and neighborhoodsd of Toronto. This list is avaiable in wikipedia, and the table with all this information will be scraped with BeautifulSoup library.

After this, data will be acquiered by the Foursquare API. I'll get all venues related with tourism activities: restaurants, malls, tourist attractions, beachs and parks. I'll also use premium calls in the Foursquare API to retrieve venues prices.

Finally, I'll use transitland database to get all subways and bus station, in order to measure the mobility of each neighborhood.

3 Exploratory Analysis

4 Results and Discussion

5 Conclusion

6 Future Directions

References