Food around the Universities in Canada

Applied Data Science Capstone
Coursera Capstone Project - The Battle of Neighborhoods

Ripple Shi

5/20/2020

This report is drafted for the capstone project of the specialization Applied Data Science Capstone offered on Coursera by IBM. In this report, we will explore the recommended food options near the universities in Canada using the machine learning algorithm.

Table of Contents

1. INTRODUCTION

I believe food is an important part of life. Food provides you with nutrition, energy and ideally, satisfaction. As someone who will soon enroll in a university in Canada, I am really curious about what types of food I can get there. I was born and raised in Asia, where the diets are quite different from those in North America, so it will be a great relief if I know whether I could easily find my familiar types of meals nearby.

I am sure this is also a concern for others who are seeking to study, work or live in a different country. However, limit to the scale of the project, we will only explore different food suppliers around Canadian universities. They could be restaurants, café or any other venues that could satisfy people's needs in food.

I hope this project could provide the readers with some insights on this subject. Although here we will focus on food, universities and Canada, I think the idea behind this project can also be applied to any similar intention.

To carry out the project, we are going to rely on the recommendations provided by Foursquare's API. By specifying the coordinate of the university, Foursquare will return us some recommended venues that are in the food section within the limit and distance we set. Meanwhile, the category of a venue will also be returned, so we could use that to guess the main cuisine of the venue.

Also, we will use K-Means, a clustering algorithm, to cluster the universities in groups and explore the features of each group. Hopefully, we could get some interesting findings out of that.

2. DATA

The data used in this project come from three sources.

We start by getting a list of universities in Canada from Wikipedia[1]. Although we are not sure whether the list is complete, it should be enough to represent the population we are interested in. Here we define the universities included in this list to be the study object of this project. There are 91 universities contained in the list, most of them are public universities, 16 are private universities. The list also provides the provinces and the cities where the universities locate. This information will play an important role in determining the location of the universities.

Next, we will use the name and the province of the universities to get their coordinates. We obtain the latitude and longitude of each university using the dataset on http://py4edata.dr-chuck.net/. This is a subset of data from the Google Geocoding API, established by Dr. Charles R. Severance from University of Michigan. This data set is built to facilitate the study of Python courses taught by Dr. Chuck. Please note that this dataset is not my first choice to get the coordinates. Recall that we only have the name of a university and the city and the province where it locates. Due to the limits of the available geocoding APIs, I cannot get the coordinates of the universities only using that information as the parameters. It turned out that the data set built by Dr. Chuck is the only option I am aware of that could help me attain my goal. We will further explain this problem in the Data Preprocessing section afterwards. Anyway, the coordinates retrieving from this data set allow us to specify the location in the search queries of Foursquare.

Finally, we use Foursquare's API to get the recommendations in the food section and do the analysis. According to the documentation of Foursquare's API, by using the endpoint "explore" we could get a list of recommended venues near the current location. The list includes much information, but we only need the venue name and the venue type. This information will be enough for us to summarize what types of venues we could find around the universities. Based on that, we will build our variables of what percent a venue category is taking among all the recommended venues that meet the conditions we set. These self-created variables will be used to train a model using K-Means algorithm to get the clusters of the universities.

[1] The link of the page is https://en.wikipedia.org/wiki/List_of_universities_in_Canada.