Clustering Similar Neighborhoods in Different Cities

Phan Nhat Hoang Nguyen

Feb, 2020

I. INTRODUCTION

1. Background

   Cities around the world are filled with many kinds of venues that define the cultures of the cities. A city's inhabitants or services of tourists have put a significant mark on differentiating it from another, not only the mean of global position. Despite of having many different features, it is somehow possible for us to group different venues which has the similar kinds of neighborhoods in different cities. Having grouped together similar kind of neighborhoods may help people make a good decision when they consider moving to another cities.

2. Problem

   Finding identical neighborhoods indifferent cities or countries so that can help provide a perception of similar neighborhoods which provide a large number of insights in order to make a decision of choosing a neighborhood that far away, but somehow feels like home.

II. DATA ACQUISITION AND CLEANING

1. Data sources

   In this capstone project, I use two sets of data. The first dataset consists of Toronto's different boroughs and their respective postcode. The second dataset consists of New York's different neighborhoods and their respective geometric coordinates. The two datasets are attached in Data folder, which is submitted together with this report.

   Or you can find it here:

- Toronto dataset: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- New York dataset: https://cocl.us/new_york_dataset

2. Data cleaning

   The first data source is a Wikipedia page that contains postal code of Toronto city in table. To scrape the data from URL, BeautifulSoup has been used to extract the table data. I would go through some more step, the dataframe obtained consists of PostalCode, Borough, Neighborhood.

   The second data source in the link above is in json format, which consists of many different features. I would format the json file so that the data finally resulted in a dataframe that consists of Borough, Neighborhood, Latitude, Longitude.

   There are some problems with the Toronto dataframe, it has some values under "Borough" column that were not assigned in the first place. So, these rows would be dropped. There are some rows under "Neighborhood" column that had no value assigned to. My solution is copying the value from the "Borough" column of the respective row into the "Neighborhood" column.

3. Features selection

   Now we have the datasets of different neighborhoods and their respective geometric coordinates for the city of Toronto and New York. So, we will come up with different venues that these different venues have to offer.

   Foursquare API provides an access to an enormous database consisting of venues from all around the world including rich variety of information such as addresses, tips, photos, etc. Having signed up for a Foursquare Developer account, using the Client ID and Client Secret, it is possible to make an API request in order to retrieve venue's information

   By feeding a function with Neighborhood name and its coordinates, using Foursquare API, different venues were extracted. After performing one-hot-encoding and grouping rows by Neighborhoods, the Toronto and New York datasets seemed to have 250 features. Both the dataframes would be combined into one single dataframe in order to perform clustering operation.

4. Dimensionality reduction

III. METHODOLOGY

IV. RESULTS
V. DISCUSSION