IBM DATA SCIENCE CAPSTONE PROJECT REPORT

Opening a Gym/Fitness Centre in Ho Chi Minh City, Viet Nam

August 24, 2019

Cuong H. Nguyen-Dinh

Contents

Introduction

Ho Chi Minh (HCM) city (its former name is Sai Gon) is known as the most populous city in Viet Nam (with a population of over 10 million people). Moreover, HCM city is also the financial centre of Viet Nam. In this dynamic city, there are many opportunities for business activities. Therefore, it is good to think of doing business in this southern beautiful city of Viet Nam.

As the most attractive city of young people in Viet Nam, the gym/fitness service is highly required. Because this sports provides the youth not only the good health but also the well-form body. Hence, opening gym/fitness centre in HCM city promises a beneficial business.

Business problem

In this report, we try to answer the question of "Can we figure out areas in HCM city where a gym/fitness centre can be launched?" Our solution to the above business question mainly depends on data science methodology and cluster analysis of machine learning technique.

Data

We use the following data to solve the aforementioned business problem:

- Data about neighborhoods in HCM city. These are starting points to navigate companies or entrepreneurs around each neighborhood.
- Coordinate data of neighborhoods including latitudes and longitudes. This kind of data will be used to search venues and to draw map.
- Venue data is served clustering method which divides neighborhoods to groups. By analyzing these groups, we can locate areas where we can set up the gym/fitness centre.

The data used in this study comes from two sources:

- Wikipedia page[1];
- Foursquare service[2].

The data collection and processing include two steps:

1. In order to build the list of neighborhoods in HCM city, we crawl data at this URL `https://en.wikipedia.org/wiki/Category:Districts_of_Ho_Chi_Minh_City`. After having analyzed the html content, the list of 24 neighborhoods is generated. Then this data is used as the input for the next step.
2. Foursquare service is used to get coordinates of each neighborhood. Then this service is repeatedly used to get nearby venues of each neighborhood. The venue data are vectorized in order to become input for cluster analysis.

In the next section, we will describe our analytic method using these data.

[1] `https://en.wikipedia.org/wiki/Category:Districts_of_Ho_Chi_Minh_City`
[2] `https://foursquare.com/`