Content-Based Neighborhood Recomendation System

Diego Roncancio

29 April, 2021

1. Introduction

1.1. Background

Nowadays neighborhoods are an important factor that may raise or decrease the likelihood of a person/family to move to a specific place. Thus, a recommendation system that suggest neighborhoods based on their surrounding top venues can significantly reduce the scope of search for a new place to live and bring a satisfying new home. As satisfaction has been repeatedly linked to efficiency and prosperity, helping people to find suitable neighborhoods can help improve the city. Furthermore, overtime this may give valuable information of what venues drive customer satisfaction within a neighborhood.

1.2. Problem

To build this recommendation system, location data pertaining the neighborhoods of a city and their most common surrounding venues will be needed. This project aims to recommend a neighborhood to a customer based on their levels of satisfaction with their previous neighborhoods.

1.3. Interest

Real estate agents would be interested in the information this recommendation system will provide to narrow their search. Also, there is the possibility to develop an app with this type of system that could also be economically interesting.

2. Data acquisition and cleaning

2.1. Data Sources

Now, to simplify matters this recommendation system will be constructed for the city of Toronto. The information of the neighborhoods for this city was scrapped from the list of postal codes in Wikipedia (here) and the information pertaining to the surrounding venues will be obtained using the Foursquare API for each of the neighborhoods. As for the necessary coordinates for each neighborhood a .csv file was found to assign each one of them.

2.2. Data cleaning

The information scrapped from Wikipedia had to be parsed in order to be useful, this was achieved using the BeautifulSoup package for Python. Thus, the table in the webpage was extracted and information was recorded in a pandas dataframe with the following features: Postal Code, Borough, Neighborhood, Longitude and Latitude.

2.3. Feature selection

After data cleaning, the data frame had 103 neighborhoods with 5 features. Clearly for the recommendation system, we need the types of venues surrounding the neighborhoods and assign them to the dataframe to build the recommendation matrix. Thus, a function was defined to return the venues of a neighborhood in a 500m radius, this radius was defined to not obtain null cells in faraway neighborhoods. Each venue was hot coded and grouped by neighborhood. Since this is a content based system, the user must have assigned a rating to their current (and previous if available) neighborhoods.