The Battle of the Neighborhoods

Clustering of Cities

Maxime Menu

April 18, 2020

1 Introduction

1.1 Background

Cheap air travel and globalization has made overseas traveling easier. Another consequence of globalization has been the cross-influencing of different cultures. This means that a lot of countries and their capital-level cities share many features in the kind of venues they manifest. This information can be used by travel agencies and alike to easily recommend and advert new travel destinations to their new customers.

1.2 Problem

Given the preference a customer might have for New York City, which city should a travel agency recommend to their customer with the assumption that travel distance is irrelevant. This project aims to answer this question.

1.3 Interest

Travel agencies, airline companies and other parties involved in the travel industry would be interested to know how to recommend next travel destinations to (potential) customers.

2 Materials & Methods

2.1 Data Gathering

Three types of data were gathered to carry out this analysis.

- Boroughs and neighborhoods for 6 selected capitals (New York, Toronto, Tokyo, Seoul, Berlin and London)
- Coordinates for neighborhoods
- Venue categories of FourSquare
- Venues in a set radius for each of the coordinates

Boroughs and neighborhoods

The purpose of boroughs and neighborhoods was to gather subdivisions of the cities which would later be clustered based on venues. In order to bring in a diversity of neighborhoods 6 capital-like cities were selected from 3 parts of the world (North America, Europe, Far East Asia).

This data was gathered from different sources, depending on the city.

New York: The data was downloaded from a public dataset provided by the Applied Data Science Capstone course as a JSON file. (https://cocl.us/new_york_dataset)

Toronto: The data was scraped from the Wikipedia page for Canadian postal codes. (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)

Tokyo: The data was scraped from the Wikipedia page for Tokyo special wards. (https://en.wikipedia.org/wiki/Special_wards_of_Tokyo)

Seoul: The data was scraped from the Wikipedia page for the districts of Seoul. (https://en.wikipedia.org/wiki/List_of_districts_of_Seoul)

Berlin: The data was scraped from the Wikipedia page for boroughs and neighborhoods of Berlin.
(https://en.wikipedia.org/wiki/Boroughs_and_neighborhoods_of_Berlin
)

London: The data was scraped from the Wikipedia page for London postal districts.
(https://en.wikipedia.org/wiki/London_postal_district)

Coordinates

The purpose of coordinates data was to be able to gather venue data with the Foursquare API in the next step, as the API requires coordinates.

This data was gathered from different sources, depending on the city.

New York: Coordinates were contained in the same file gathered for the boroughs and neighborhoods data or New York.

Toronto: The data was downloaded from a public dataset provided by the Applied Data Science Capstone course as a csv file. (http://cocl.us/Geospatial_data) This dataset was merged with the neighborhoods data for Toronto based on postal codes.

Tokyo: The coordinates were gathered by using the geopy package for python with the Nominatim geocoder for OpenStreetMap data.

Seoul: The coordinates were gathered by using the geopy package for python with the Nominatim geocoder for OpenStreetMap data. In case no result could be returned for the neighborhood names, coordinates were gathered by manually inputting the neighborhood names in google maps. (https://www.google.com/maps)

Berlin: The coordinates were gathered by using the geopy package for python with the Nominatim geocoder for OpenStreetMap data. In case no result could be returned for the neighborhood names, coordinates

were gathered by manually inputting the neighborhood names in google maps. (https://www.google.com/maps)

London: The coordinates were downloaded from a public dataset containing coordinates for every postal code in the United Kingdom. (url: https://www.doogal.co.uk/files/postcodes.zip) This dataset was merged with the neighborhoods data for London based on postal codes.

Coordinates were visually confirmed by mapping them with the Folium package for python.

Venue Categories

The foursquare API returns the most specific category for a given venue. Spatial bias can however be included in the categories if they are too specific (e.g. Japanese restaurant, Korean restaurants etc.) which lead to the need to gather all Foursquare categories with their hierarchies as to be able to infer super categories of venues.

This data was gathered by using the categories Foursquare API.

Venues

The purpose of venues data was to assign features to neighborhoods based on the categories of venues found in these neighborhoods. These features were then used in the clustering steps in the analysis part.

This data was gathered by using the explore API of Foursquare. A limit of 100 venues in a radius of 500 meters for every set of coordinates was gathered.