Creating a neighbourhood recommender tool

As part of the IBM Data Science Capstone Project

Marco Weber

22. January of 2021

1 Introduction

1.1 Background

International mobility is becoming a more and more common occurrence among young professionals. Widespread adoption of English as lingua franca, facilitated international mobility and increasing interest in a diverse workforce lead many aspiring talents to move outside their home country and settle in a variety of new places. While companies have already introduced data-driven business models to address the specific needs of these people on a global scale, such as Airbnb, Craigslist and others, people interested to move still have to rely on word-to-mouth and cumbersome internet research. While they are interested to move they still want to make sure their standards of daily life and the offer of leisure activities stay on a somewhat similar level than their current neighborhoods.

1.2 Problem

The following paper will tackle the issue young professionals face to secure a similar life standard, while moving to other places. For that, an exemplary case of a person currently living in Berlin's neighborhood of "Friedrichshain-Kreuzberg" that wants to move to Manhattan is introduced. By creating a clustering model, the current neighborhood will be compared to the neighborhoods in Manhattan to find the ones with the highest similarity. This model could then be further developed to address any start and target destination, if the exemplary model yields satisfactory results.

2 Data

2.1 Data Sources

Two data sources are required to build the desired model:

1. Neighborhood names and geodata: To create an overview over the start destination and the possible destination targets, the information will be extracted from Wikipedia with the web scraping tool "BeautifulSoup". This will also facilitate the extraction of geodata for each neighborhood. The geodata of latitude and longitude will be extracted with the geolocation tool "Nominatim" to extract precise locations for each neighborhood.

2. Venue information: Based on the data acquired in point 1 the most common venues will be extracted from the Foursquare API for all defined destinations. Based on this data the comparison algorithm will be constructed to determine the similarity between the start and the possible target locations.

3 Methodology

3.1 Data extraction and curation

The address and geodata for the current address were procured manually, while the Manhattan neighborhoods were web scraped from Wikipedia and the geodata were seized with a geolocator. For the current address and the possible target neighborhoods the most relevant venues were collected from the Foursquare API and stored in two separate tables. To secure data integrity, an intersection of the venues in both tables was performed to eliminate venues types that appeared in only one table and thus would not add to the similarity value. Also, all empty or erroneous columns were deleted from the tables.

3.2 Model construction

To find the target neighborhoods with the highest similarity to the current one, a neighborhood profile of the current address was created by using the count of each venue type as multiplicator value. Then a matrix multiplication of both current and target neighborhoods was conducted to receive the similarity value per value type per target neighborhood. To receive the overall similarity score the sum of all venue type values was computed. Finally, all Manhattan neighborhoods were sorted descending according to their similarity value, which range from 1 being the lowest similarity to 153 being the highest similarity value.

3.3 Model visualization

To make the result more readable, a “folium” map was created with each of the neighborhoods as location marker added. To illustrate the top similar neighborhoods, the top 3 were marked in green, the top 10 were marked in yellow and the rest were marked in blue.

Figure 1: Target locations, color-indicated

4 Discussion

The top 3 neighborhoods in similarity that the young professional could investigate further were:

Neighborhood	Similarity-Score
Bowery	153
Lower East Side	152
Rose Hill	150

Furthermore, all neighborhoods in the top 10 of similarity seem to be located quite closely to the downtown area. There were also 2 locations outside of Manhattan (not visible in the photo), which was probably due to an erroneous lookup by the geolocator. Technically, a process to delete neighborhoods outside the relevant boundaries could be implemented additionally. Other than that the tool provided satisfying results in determining attractive locations that would not have been identified as easily by research or other measures.

5 Conclusion

In this capstone project a quick and effective recommender tool to compare a specific address's similarity to possible target destinations based on the existing venues in the current neighborhood was created. The tool can also provide some flexibility to adjust current and target destinations rather easily. This could therefor help to facilitate or at least support the process of finding a suitable neighborhood to move into.