Matching Neighbourhoods

Coursera IBM Data Science Capstone Project

Michael Weinberger

January 2021

Introduction / Business Problem

In today's world people frequently need to relocate, be it due to work requirements or for personal reasons. Often, this involves moving to a city that you have not lived in before and may never have visited.

So how do you decide which part of a city that you do not yet know you would like to live in?

One way would be to base this decision on past experience and start from a neighbourhood that you know and like. You can then choose a new neighbourhood based on the fact that is very similar to the one you like.

This project allows you to do just that: It uses data on the presence of various kinds of local venues (https://foursquare.com/) to determine neighbourhoods that best match the area surrounding a location freely chosen by the user.

The project therefore is aimed at a broad audience, namely anyone who would like a personalised recommendation on which new neighbourhood best to move to. It thus addresses a common need. It is distinct from existing property marketing resources in that it helps to choose a neighbourhood, rather than a specific property.

To reach its audience, the project could be implemented as a web-based app. The app would furthermore be able to give additional information, such as the kinds of venues that define a chosen neighbourhood.

Data

In this project I used neighbourhood data (names and locations) of four cities: San Francisco, New York, Toronto and London. Additionally, the project relies on a freely chosen input address that is to be matched to the other neighbourhoods.
The San Francisco data has been used in the Data Visualization with Python course (https://cocl.us/sanfran_geojson). I extracted the neighbourhood names from the geojson file "features" sections and "properties" subsections. I used the averages of the "geometry" section and "coordinates" subsection values for each neighbourhood as neighbourhood locations.
I collected the New York neighbourhood name and location data from https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json (as used in week 3 of the Data Science Capstone Project course). I isolated the names from the json file "properties" sections and "name"

subsections, and the locations from the "geometry" sections and "coordinates" subsections. To reduce the number of neighbourhoods, I only kept those located in Manhattan.
The Toronto neighbourhood name data are from https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M, more specifically from the "Toronto - 103 FSAs" table on the website. I acquired the neighbourhood location data from http://cocl.us/Geospatial_data (as used in week 3 of the Data Science Capstone Project course) and merged with the name data using the postal code information. To reduce the number of neighbourhoods, I only kept those located in Downtown Toronto.
The London neighbourhood name and location data are from https://en.wikipedia.org/wiki/List_of_London_boroughs. I used the "List of boroughs and local authorities" table on the website to get names and coordinates of the neighbourhoods.
I used the Foursquare API (https://foursquare.com/) to collect venue data for the user-supplied location as well as for each neighbourhood location, applying a search radius of 600 meters and an upper limit of 100 found venues per neighbourhood. I collected the venue data using my personal Foursquare developers account credentials.

Methodology

I first summarised the collected venue data to create a table of venue occurrence counts, with each neighbourhood in a separate row and each venue type in a separate column. Only neighbourhoods featuring more than 4 total counts and venues featuring more than 20 counts were retained. For each neighbourhood, venue counts were then normalised against the total number of venues in that neighbourhood. Finally, the count data was scaled up to represent counts per 10000 total counts.

To analyse which venues characterise the input neighbourhood, I generated a bar plot of the venues with the highest count frequencies in that neighbourhood. I then used hierarchical clustering (contained in the Seaborn package, method='ward', metric='euclidean') to group the neighbourhoods according to their count data similarity and generated a heatmap showing the clustering dendrogram and the venue counts. Here, I log2 transformed the count data to allow for a clearer visualisation on the heatmap.

To rank neighbourhoods according to their venue count distance from the input neighbourhood I generated a distance matrix using data that had not been log-transformed. I then generated another heatmap depicting the count data distances between the neighbourhoods and the input address area. Finally, I created maps using the Folium package on which the neighbourhoods were marked and colour labelled according to their count distance from the input area.

Results

I first analysed the input address *175 5th Avenue NYC*. This address is located in *Flatiron*, a Manhattan neighbourhood also contained in the New York neighbourhood data. The area surrounding *175 5th Avenue NYC* featured many restaurants, with *New American Restaurant* being the most frequent venue and comprising 6% of all counts (Figure 1).

Hierarchical clustering showed that the venue count data of *175 5th Avenue NYC* was indeed most similar to that of *Flatiron* (Figure 2). Most neighbourhoods clustering close to the input

address were located in New York, the closest neighbourhoods from another city were located in San Francisco, such as *Northern* and *Mission*. Neighbourhoods located in Toronto and London tended to cluster further away from *175 5th Avenue NYC*.

Ranking the neighbourhoods according to their venue count similarity with *175 5th Avenue NYC*, I found again that *Flatiron* was the closest neighbourhood and that many of the neighbourhoods with very high similarity were located in New York (Figure 3). The most similar neighbourhood in San Francisco was *Northern*, the most similar in Toronto *St. James Town* and the most similar in London *Ealing*.

Marking the neighbourhood locations on folium maps of the four cities, I found that the neighbourhoods which were similar to the area surrounding *175 5th Avenue NYC* were generally located close to the city centres (Figure 4). This matched the fact that the input area was located in the centre of New York. In contrast, several neighbourhoods located further away from the city centres could not even be analysed as their total venue counts were too low.

Next, I analysed a different input address (*45 Rue Greneta, Paris*), which I chose randomly. Here, *French Restaurant* and *Bakery* were the most frequent venues in the surrounding area (Figure 5).

The most similar neighbourhoods were *Little Italy* in New York, *Northern* in San Francisco, *Southwark* in London and *St. James Town* in Toronto (Figure 6).

Discussion

The analysis described here indicates that the algorithm is able to faithfully analyse and cluster neighbourhoods based on Foursquare venue data. This can be used to address the problem described in the introduction: For example, someone who likes the area around the Rue Greneta in Paris might be interested in the Little Italy neighbourhood in Manhattan as it features similar venues.

A caveat is that the quality of the clustering depends on the number of venues that can be retrieved for a neighbourhood. If neighbourhoods show a high dissimilarity to the input neighbourhood, this might be due to a low venue count in those neighbourhoods. If the input neighbourhood is very dissimilar to all other neighbourhoods, it might be good to check that the total venue count for the input neighbourhood is sufficiently high.

Here, analysis of neighbourhoods is confined to San Francisco, New York, Toronto and London. To make the algorithm more general, it would be best to have access to a standardized database of city neighbourhood names and locations. This might not be available as freeware though.

Conclusions

The neighbourhood matching algorithm allows for comparing the area surrounding an input address to other neighbourhoods by using Foursquare venue data.

A user can choose an input address of interest and determine which neighbourhood is most similar (or dissimilar) to that input area.

At the moment, the analysis is confined to San Francisco, New York, Toronto and London. However, the algorithm can theoretically be generalised to include a wider range of cities.

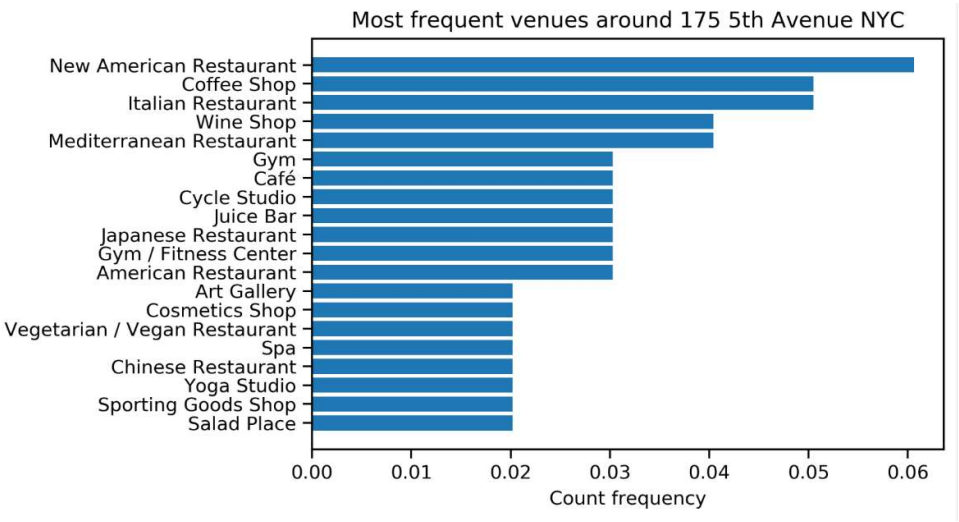

Figure 1: The most frequent Foursquare venues in the area surrounding *175 5th Avenue NYC*.

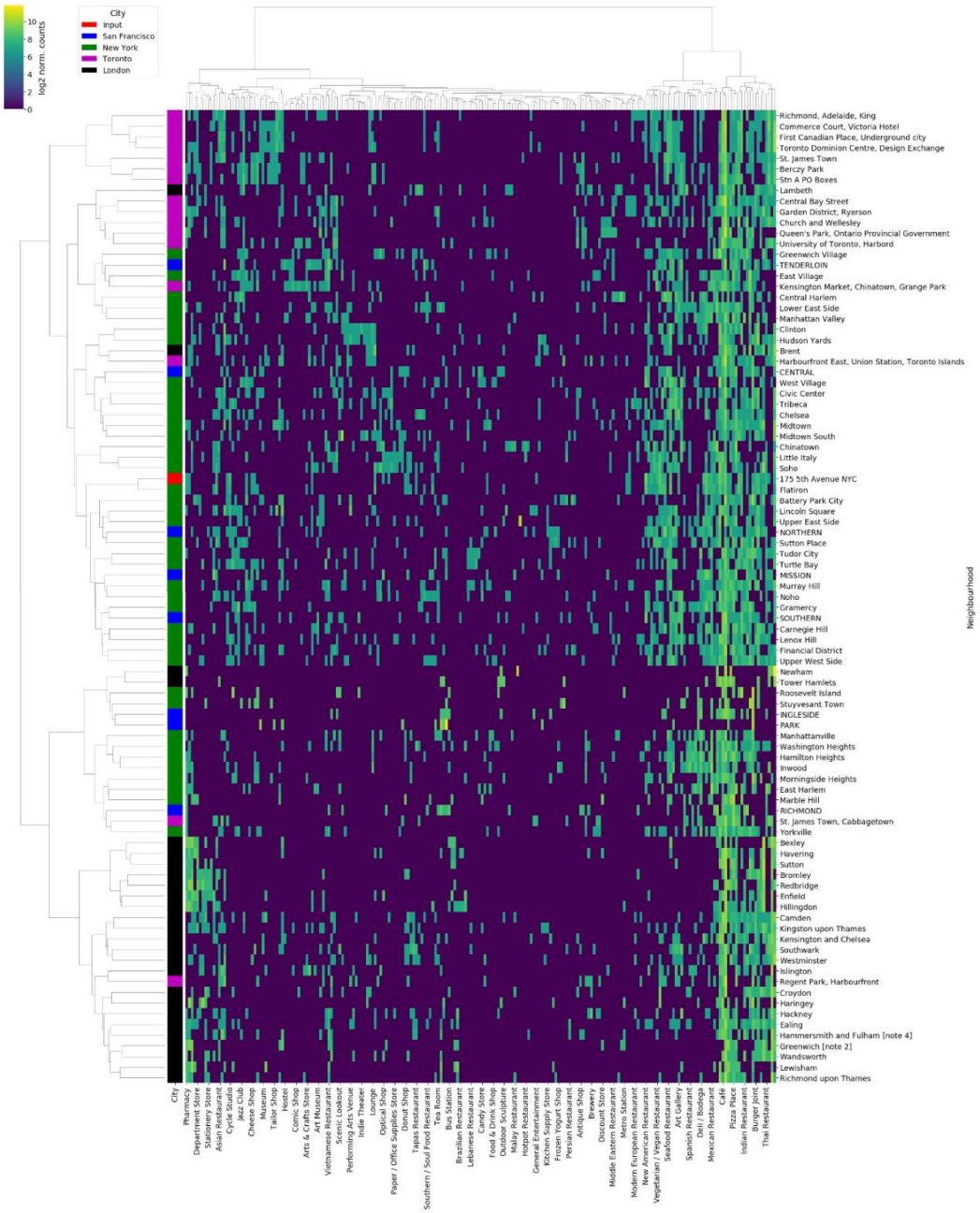

Figure 2: Heatmap showing venue counts across the area surrounding *175 5th Avenue NYC* and all other neighbourhoods. Shown are log2-transformed counts. Both neighbourhoods and venues have been clustered hierarchically. The colour bar to the left indicates which city each neighbourhood is located in.

Figure 3: Heatmap of neighbourhood venue count distance to *175 5th Avenue NYC*. Distance values were scaled to lie between 0 and 1. A low distance indicates high similarity with the input area.

Figure 4: Folium maps of San Francisco, New York, Toronto and London. Neighbourhoods have been marked by circles. Circle fill colour indicates the venue count distance to *175 5th Avenue NYC*. Neighbourhoods that were excluded from the analysis due to low venue counts are coloured grey.

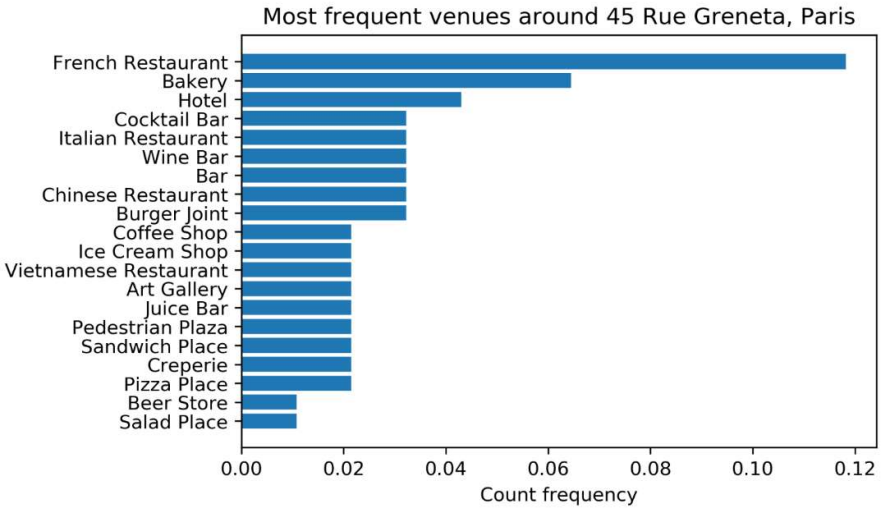

Figure 5: The most frequent Foursquare venues in the area surrounding *45 Rue Greneta, Paris*.

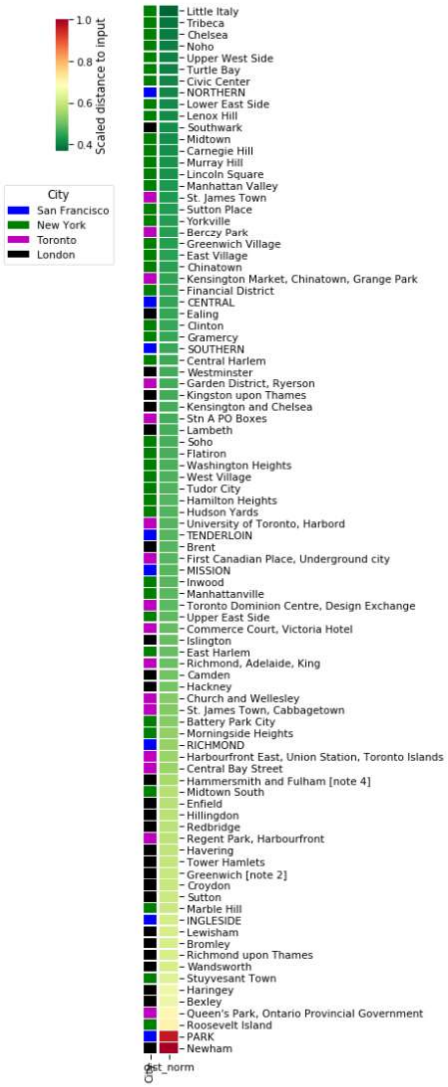

Figure 6: Heatmap of neighbourhood venue count distance to *45 Rue Greneta, Paris*. Distance values were scaled to lie between 0 and 1. A low distance indicates high similarity with the input area.

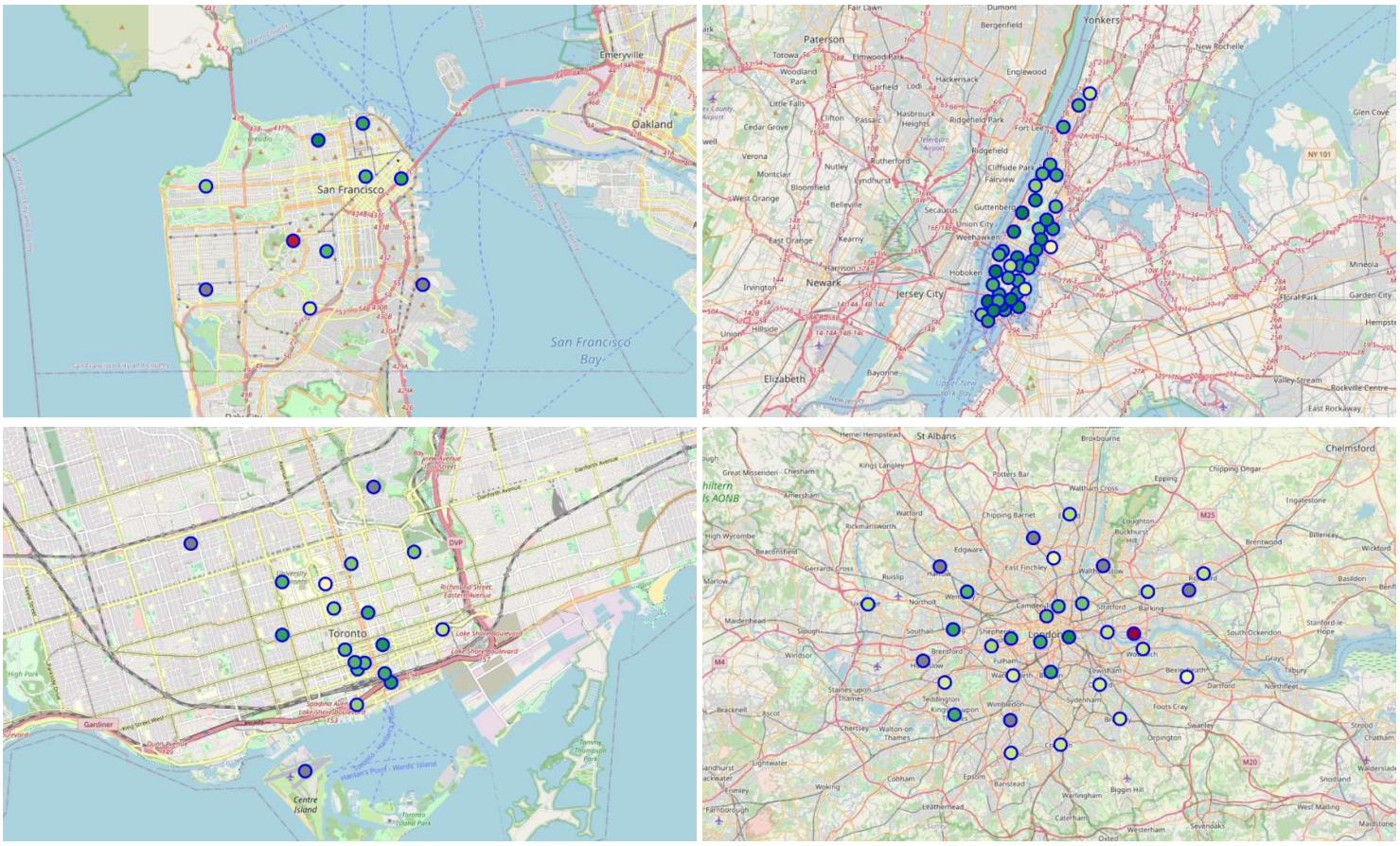

Figure 7: Folium maps of San Francisco, New York, Toronto and London. Neighbourhoods have been marked by circles. Circle fill colour indicates the venue count distance to *45 Rue Greneta, Paris*. Neighbourhoods that were excluded from the analysis due to low venue counts are coloured grey.