How Similar are Neighbourhoods in London, Sydney and New York?

Aaron Armour

Introduction

Deciding to shift to a new city is a big decision to make, even more so when that city is in another country. Imagine that you are in the situation of moving overseas and deciding where to live. An important consideration for you when making this decision might be to find an area which is similar, in terms of the amenities on offer, to where you live presently. In addition to the neighbourhood you choose to live in, you are likely to place some importance on the surrounding neighbourhoods within in the city as well. That is, you would also appreciate a degree of similarity of your new city with your current city.

The audience for this project is people who are planning a shift between two of these cities, and are considering where to live in the new city. This project will also be of interest to people living in one of these three cities and considering which of the other two cities they might like to shift to.

London, Sydney and New York are all the financial capitals of their respective countries. There will be many workers in the banking, insurance and financial services sector (amongst others) who shift between two of these cities. So I expect the findings of this project to be of use to a large group of people.

There are other factors to take into account in deciding where to shift, but this project will provide relevant information for one important aspect of this decision.

I intend to answer the following three specific questions in this project:

1. How similar are neighbourhoods across the different cities; are there similar neighbourhoods in other cities? Or do neighbourhoods tend to be most like other neighbourhoods within that same city?
2. For each neighbourhood under consideration which are the most similar neighbourhoods, one from each of the other two cities?
3. How similar are each pair of cities? (This would allow us to say Sydney is more similar to London than New York, for example.)

Data

In order to limit the number of neighbourhoods under consideration, I have chosen to focus on the central city areas, the inner boroughs of London, the City of Sydney and Manhattan. I will use the following pages on Wikipedia to obtain lists of neighbourhoods in each of the three cities:

- https://en.wikipedia.org/wiki/List_of_areas_of_London
- https://en.wikipedia.org/wiki/City_of_Sydney
- https://en.wikipedia.org/wiki/List_of_Manhattan_neighborhoods

As an example, the Wikipedia page List of areas of London has the following table:

Location	London borough	Post town	Postcode district	Dial code	OS grid ref
Abbey Wood	Bexley, Greenwich [7]	LONDON	SE2	020	TQ465785
Acton	Ealing, Hammersmith and Fulham[8]	LONDON	W3, W4	020	TQ205805
Addington	Croydon[8]	CROYDON	CR0	020	TQ375645
Addiscombe	Croydon[8]	CROYDON	CR0	020	TQ345665
Albany Park	Bexley	BEXLEY, SIDCUP	DA5, DA14	020	TQ478728
Aldborough Hatch	Redbridge[9]	ILFORD	IG2	020	TQ455895
Aldgate	City[10]	LONDON	EC3	020	TQ334813
Aldwych	Westminster[10]	LONDON	WC2	020	TQ307810
Alperton	Brent[11]	WEMBLEY	HA0	020	TQ185835
Anerley	Bromley[11]	LONDON	SE20	020	TQ345695
Angel	Islington[8]	LONDON	EC1, N1	020	TQ345665
Aperfield	Bromley[11]	WESTERHAM	TN16	01959	TQ425585
Archway	Islington[12]	LONDON	N19	020	TQ285875
Ardleigh Green	Havering[12]	HORNCHURCH	RM11	01708	TQ535895
Arkley	Barnet[12]	BARNET, LONDON	EN5, NW7	020	TQ225955
Arnos Grove	Enfield[12]	LONDON	N11, N14	020	TQ295925
Balham	Wandsworth[13]	LONDON	SW12	020	TQ285735
Bankside	Southwark[14]	LONDON	SE1	020	TQ325795

The Wikipedia page https://en.wikipedia.org/wiki/List_of_places_in_London has a table which classifies each London borough as inner or outer. This information can be used to filter the London neighbourhoods to just those belonging to the inner boroughs.

I will use Python's geopy module to obtain the latitude and longitude of each neighbourhood. When this fails I will manually obtain the information through some other method, such as a google search. An example showing how to use geopy is as follows:

In [2]:
```
from geopy.geocoders import Nominatim
locator = Nominatim(user_agent='My_Coursera_Applied_Data_Science_Capstone')

place = locator.geocode('Canary Wharf, London')
place
```

Out[2]:
```
Location(Canary Wharf, Isle of Dogs, London Borough of Tower Hamlets, London, Greater London, England, E14 4HE, United Kingdom, (51.50562, -0.0257
169, 0.0))
```

In [3]:
```
lat = place.latitude
long = place.longitude
print(lat, long)
```

```
51.50562 -0.0257169
```

I will use location data from Foursquare, making API calls to their 'venue explore' endpoint to obtain a list of venues from a given neighbourhood. The results obtained from Foursquare are a list of recommended places, below is an example of such a recommendation. The key piece of information I will extract from this data the name of the category that the venue belongs to, which is 'Plaza' in the example below.


```
{'reasons': {'count': 0,
  'items': [{'summary': 'This spot is popular',
    'type': 'general',
    'reasonName': 'globalInteractionReason'}]},
 'venue': {'id': '4c4991b19f2ad13a0fd77653',
  'name': 'Canary Riverside',
  'location': {'address': 'Canary Riverside',
   'lat': 51.50644504299903,
   'lng': -0.028794868546109256,
   'labeledLatLngs': [{'label': 'display',
     'lat': 51.50644504299903,
     'lng': -0.028794868546109256}],
```

```
   'distance': 232,
   'postalCode': 'R M14',
   'cc': 'GB',
   'city': 'Canary Wharf',
   'state': 'Greater London',
   'country': 'United Kingdom',
   'formattedAddress': ['Canary Riverside',
    'Canary Wharf',
    'Greater London',
    'R M14',
    'United Kingdom']},
  'categories': [{'id': '4bf58dd8d48988d164941735',
    'name': 'Plaza',
    'pluralName': 'Plazas',
    'shortName': 'Plaza',
    'icon': {'prefix':
'https://ss3.4sqi.net/img/categories_v2/parks_outdoors/plaza_',
     'suffix': '.png'},
    'primary': True}],
  'photos': {'count': 0, 'groups': []}},
 'referralId': 'e-0-4c4991b19f2ad13a0fd77653-0'}
```


These results will be used to populate a DataFrame where each row contains a neighbourhood and the category of a venue in that neighbourhood returned by the Foursquare query. I will use one-hot encoding and convert the categories into dummy variables. I will then group by neighbourhood and average to obtain a vector of observed proportions of venues in each category for a given neighbourhood. In this way, I will obtain a feature vector for each neighbourhood. (This technique was taught in the week 3 lab of the Applied Data Science Capstone on Coursera.org.)

I will aim to get data on 50-70 neighbourhoods in each of the three cities. I would consider dropping neighbourhoods where there are too few venues returned by Foursquare. If there is an excess of neighbourhoods, I will favour those with more venues returned from Foursquare.