Battle of Neighbourhoods

Capstone Final Report

Introduction & "Business" Problem

When visiting a major city in another country for a short period, it can be difficult to establish the orientation and location of interesting "neighbourhoods". Unless aided by friends or family who know the destination city well, it is common to get stuck in the tourist quarters, and miss the vibrancy of each city's more interesting parts. In London, where I live, tourists are rarely seen outside of relatively few and relatively central locations: Westminster, the West End, the South Bank, and Shoreditch. This could be partly due to the contrivances of guide books and partly due to the overwhelming size of modern cities.

To get a sense of what each neighbourhood in a destination city is like, it can be useful to "translate" their characteristics to a more familiar place. For example, Peckham and Brooklyn are more similar than, say, Peckham and Midtown, which would be a helpful orientation device for a native Londoner (particularly if they are in search of craft beer and hipsters).

This problem is contemplated by the site Hoodmaps (https://hoodmaps.com), which relies on a crowdsourcing technique to characterise the neighbourhoods of cities around the world. This is a really interesting approach, however, I propose a different approach, leveraging existing location data, such as venue-related data offered via the Foursquare API, which would be interesting to tourists. By building a venue-based profile for each neighbourhood in a number of cities, it would be possible to employ a clustering approach across cities to identify similarities in neighbourhoods based on the venue types located there.

To test this concept, we will use three major cities, about which there will be sufficiently large datasets available: London, New York, and Paris. I have selected these cities, as I have first-hand experience of spending significant periods of time in each, and can compare the results with the "ground truth", given that clustering will be unsupervised without a formal evaluation technique.

Data

The data about the neighbourhoods will be sourced through two means:

1. First, we need to partition each city into identifiable neighbourhoods, and obtain geographic coordinates for the centre of each. We will explore the use of postal / zip codes to do this initial partitioning. We will present the results of this on a map for each city.

- New York: https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm
- London: https://www.milesfaster.co.uk/london-postcodes-list.htm
- Paris: https://www.worldpostalcodes.org/en/france/arrondissement/map-of-postal-codes-in-paris

2. For each neighbourhood in each of the three cities, and using the coordinates obtained for each, we will identify the principal venue types based on their frequency of occurrence using the Foursquare API.

We will assess various clustering techniques, using the Scikit-learn library, to then segment neighbourhoods and identify similar neighbourhoods across the three cities.

Data Exploration

On extracting the postcode / zip code data from the sources outlined in the 'Data' section, it was identified that London and New York contained a significantly higher number of locales than Paris. This is a limitation of using postcodes to segment each city into identifiable locales.

- To make the comparison useful, London and New York were restricted to the most central 45 locations. This reduced the number of locales in London and New York to a more relevant segment representing the respective city centres.

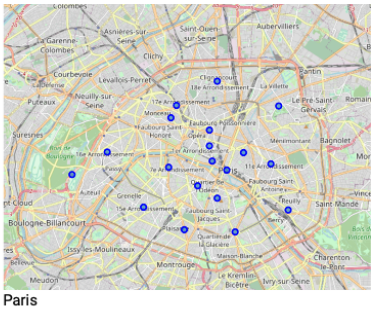
City	Locales
London	121
New York	178
Paris	19

City	Locales
London	45
New York	45
Paris	19

- After extracting the postcode / zip code data from the sources outlined in the 'Data' section, Geocoder was used to geocode the locales to a set of latitude and longitude coordinates, for use in the Foursquare API. Folium was used to create a map of each city and its respective locales.

New York

London



Paris

- Next, venue data was obtained for each locale identified. A radius of 500 was supplied, and results were limited to 100 venues.
- 375 individual venue categories were identified across the three cities. Only 217 of these venues were present in more than one of the cities, however, and the 158 present in only one city were removed, since these would provide no useful comparison.

Present In	Number
One city	158
Two cities	99
All (three) cities	118

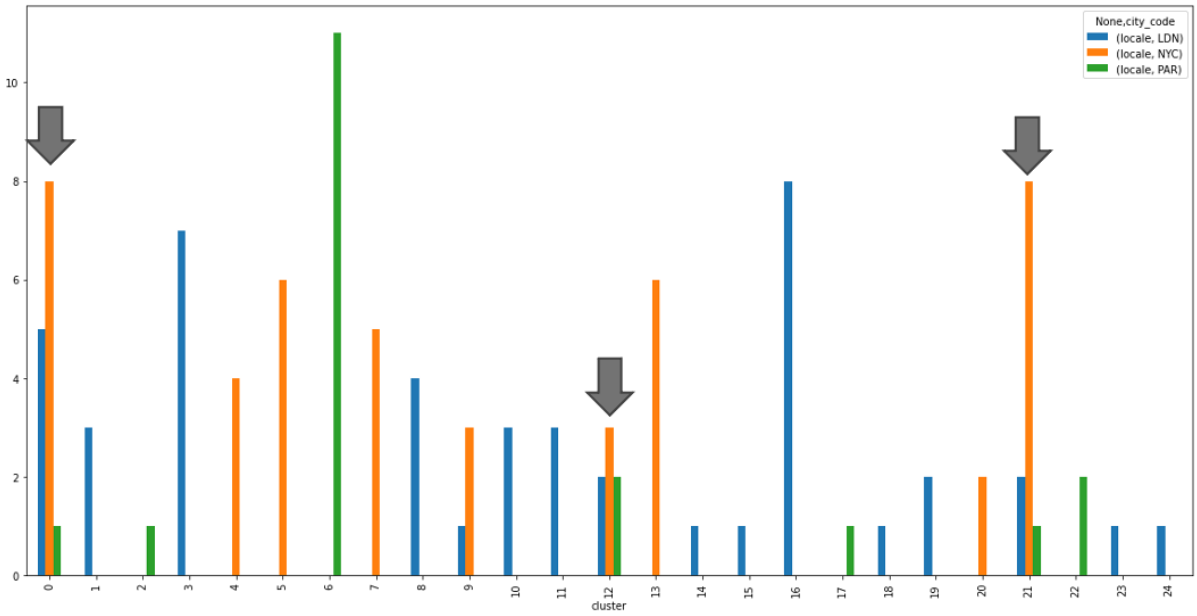
- The 10 most popular venues across all three cities were as follows:

French Restaurant
Hotel
Grocery Store
Café
Pub
Lake
Bike Rental / Bike Share
Art Gallery
Bar
Bakery

Methodology

Clustering

- To identify similar locales across the three cities, K-Means clustering was used. Various values of K were used to identify an appropriate number of clusters.
- Given a large number of venue categories and locales, a large value of K was chosen - values from 15 to 30 were explored.
- A graph showing the clusters across cities, for K = 25, with the three most interesting clusters highlighted:



Results & Discussion

- The locales in each city, with a few exceptions, were most similar with other locales in the same city. This makes sense: although London, Paris, and New York are vibrant, multicultural, global cities, each has a distinctive culture.
- Two particularly interesting clusters were identified, highlighted on the chart.
  - **Cluster 0. This appears to relate to wealthy residential neighbourhoods.**
    - In NYC, six of the eight zip-codes in this cluster were located in and around the Upper East Side.
    - In London, three of the five zip-codes in this cluster were located in Kensington & Chelsea.
  - **Cluster 12. This appears to relate to the touristic centre.**
    - In London, this cluster contains both Soho and Covent Garden, collectively the West End.
    - In Paris, this cluster contains the 1ème arrondissement, which contains many of the most well-known tourist attractions.

Conclusion

- Overall, the results from clustering were interesting, however, to generalise and improve recommendations on neighbourhood similarity across the cities and make these more useful to visitors, there are three key recommendations resulting from the analysis carried out:
  - Separate clustering of each neighbourhood from other neighbourhoods in the same city, to control / prevent the intra-city similarity from dominating the

clustering.

- Explore a richer set of data on venues, either from Foursquare, or from another API, such as venue price ranges (such as the indicators shown on Google and Yelp), to further differentiate venues, or using transport data to gauge how busy areas are at different times in the week.
- Explore alternatives to using postal / zip codes to segment neighborhoods. These often merged neighbourhoods, which in fact have distinctive characters. For example, W1 in London covers “Mayfair, Marylebone, Soho", which are three quite distinct areas of central London.