Chris O'Sullivan
Data Science Professional Certification
10 June 2020

Employee Secondment: Neighbourhood Comparisons and Recommendation

Background Introduction

A large multi national firm offers secondment opportunities to different locations around the world for employees; this can enhance both the employee's and the businesses performance by offering a variety of experience and exposure to alternative markets.

The HR team have asked a data scientist to produce an analysis that can make pertinent suggestions of where to base living arrangements for employees who are about to be seconded.

This is important for the firm to provide a high standard of job satisfaction, looking after their employees needs in what can be a difficult transition abroad for many people. Ideally this analysis would recommend regions of cities that are most similar to the employee's home city in terms of nearby amenities.

Problem

- Given the current living location of an employee, can the analyst find the most similar neighbourhood in alternative cities given a list of common amenities?
- In our example we will find the most suitable neighbourhood within Montreal and New York for an employee who is currently living in Camden, London. We will compare the two cities to London and determine which is the most suitable location for our employee's secondment, considering only local amenities in this case.
- This will be done by performing a similarity analysis in a K Means Classifier on the list of common venues to determine the locations with the most similarity to Camden Town.
- A general visual comparison of the two cities can also be performed from the results of the analysis.

Data Sources and Usage

The following data sources will be used to demonstrate our solution to this location clustering problem:

- City Neighbourhoods for London, Montreal and New York:
    - To obtain the list of neighbourhoods, we will perform web scraping. For example shown below are the results of the London web scraping from https://en.wikipedia.org/wiki/List_of_London_boroughs. This data already contains latitude and longitudinal coordinates but must be cleaned to split them.

	Borough	Inner	Status	Local authority	Political control	Headquarters	Area (sq mi)	Population (2013 est)[1]	Co-ordinates	Nr. in map
0	Barking and Dagenham [note 1]	NaN	NaN	Barking and Dagenham London Borough Council	Labour	Town Hall, 1 Town Square	13.93	194352	51°33′39″N 0°09′21″E / 51.5607°N 0.1557°E	25
1	Barnet	NaN	NaN	Barnet London Borough Council	Conservative	Barnet House, 2 Bristol Avenue, Colindale	33.49	369088	51°37′31″N 0°09′06″W / 51.6252°N 0.1517°W	31
2	Bexley	NaN	NaN	Bexley London Borough Council	Conservative	Civic Offices, 2 Watling Street	23.38	236687	51°27′18″N 0°09′02″E / 51.4549°N 0.1505°E	23
3	Brent	NaN	NaN	Brent London Borough Council	Labour	Brent Civic Centre, Engineers Way	16.70	317264	51°33′32″N 0°16′54″W / 51.5588°N 0.2817°W	12
4	Bromley	NaN	NaN	Bromley London Borough Council	Conservative	Civic Centre, Stockwell Close	57.97	317899	51°24′14″N 0°01′11″E / 51.4039°N 0.0198°E	20

- To obtain the latitude and longitude for Montreal and New York, we will use a the geopy geocode python module to extract the data from the borough and city combination, amending the results to our data frame. An example of the final data for New York is shown below.

	Neighborhood	City	Latitude	Longitude
0	Central Bronx	New York	40.846651	-73.878594
1	Bronx Park	New York	40.858847	-73.875904
2	High Bridge	New York	40.842233	-73.929305
3	Hunts Point	New York	40.812601	-73.884025
4	Kingsbridge	New York	40.878705	-73.905141

- In total there are 32 London neighbourhoods, 19 Montreal neighbourhoods and 54 New York neighbourhoods.

- Foursquare venue data for all neighbourhoods to produce a data frame of amenities.
  - We will select the 100 nearest amenities in a 1000m radius to the coordinates we have been given. An example of this call is shown below for Montreal:

	Postcode	Postcode Latitude	Postcode Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Ahuntsic-Cartierville	45.541892	-73.680319	Sapori Di Napoli	45.540799	-73.685707	Italian Restaurant
1	Ahuntsic-Cartierville	45.541892	-73.680319	Parc Marcelin-Wilson	45.540585	-73.685730	Park
2	Ahuntsic-Cartierville	45.541892	-73.680319	Subway	45.537962	-73.679149	Sandwich Place
3	Ahuntsic-Cartierville	45.541892	-73.680319	AMT Gare de Bois-de-Boulogne	45.540292	-73.676976	Train Station
4	Ahuntsic-Cartierville	45.541892	-73.680319	Aréna Marcellin-Wilson	45.540662	-73.685701	Hockey Arena

  - In total we found 1257 venues in London and 215 venue types, 284 venues and 109 types in Montreal and 2138 venues and 277 Types in New York.
  - To perform representative analysis, we will reduce the data frame to contain only common venue types across the three cities. This is a field of 77 venue types, a reasonable number for classifying similarity across neighbourhoods, including items such as train stations, restaurants, gyms and bars. We will train our classifier on this data. This data will be categorical so to train a K means classifier, this data will be converted to numeric data via one hot encoding for all venue types.
  - A similarity analysis will allow us to characterise the locations in Boston and New York that are most similar to the borough of Camden Town and to perform a general visual comparison of the three cities.