Nearest Neighborhoods Project

Introduction/Business Problem

Problem

A New York property developer has decided to expand his portfolio by acquiring properties in London. The developer has experience and familiarity with neighborhoods in New York and the opportunities they represent but is unfamiliar with areas of London that could offer similar rates of return to his current properties.

The developer is looking for a way to compare neighborhoods in New York with neighborhoods in London in order to find their nearest equivalents.

Discussion

The developer has had the idea that the demographic characteristics of a neighborhood may be represented by its venues i.e. the range of theatres, restaurants, gyms, etc. For example, a working class neighborhood in Queens will have very different range of venues from the financial district of Manhattan or the hipster 'capital' at Williamsburg. The idea is that by representing the venue profile of a particular neighborhood in New York, it may be possible to identify its equivalent or 'nearest neighborhood' in London.

A potential problem is that cultural differences between the USA and Britain may account for different venue profiles. There may also be cultural differences arising from the different levels of ethnicity between the two cities. For example, New York has a smaller proportion of Asians (13.7%)[1] than London (18.5%)[2]. This may have an effect on the venues (e.g. Asian-style restaurants) in each city.

A potential solution to this problem would be to categorise restaurants generically. For example, an Italian restaurant and an Indian restaurant would both be categorised as 'restaurant'.

This method of comparing neighborhoods may be approximate, however an approximate solution is acceptable to the developer.

Notes

1. From webpage at http://worldpopulationreview.com/us-cities/new-york-city-population/ as referenced on Nov 20, 2018
2. From webpage at http://worldpopulationreview.com/world-cities/london-population/ as referenced on Nov 20, 2018

Data

Three sources of data are required for each city:

- A list of the names of neighborhoods
- The location (latitude and longitude) of each neighborhood
- A list of venues (categorised by type) in the vicinity of each neighborhood

For New York, a list of the neighborhoods and their geocordinates was provided in a lab session on this course at https://ibm.box.com/shared/static/fbpwbovar7lf8p5sgddm06cgipa2rxpe.json

Additional data was sourced as follows:

City	Data	Source
London	Names of neighborhoods (known as 'wards' in London)	Website: https://www.citypopulation.de/php/uk-wards-london.php
London	Geo-coordinates (longitude, latitude) of each neighborhood	API calls via Nominatim[1] geo-locator service used with geopy library.
London & New York	Venue details within each neighborhood.	API calls via the FourSquare venue rating webservice

The data contains:

Data	Format	Fields	Notes
NY Neighborhoods	json	properties:borough = borough (string) properties:name = neighborhood (string) geometry:coordinates:0 = longitude (float) geometry:coordinates:1 = latitude (float)	
London Neighborhoods	HTML	Table column 0 = borough or ward (string) Table column 1 = type (string)	Scraped from website. Ward is used as the neighborhood name
London Coordinates	python	geolocator.geocode.location.latitude (float) geolocator.geocode.location.longitude (float)	
London/NY Venues	python	venue.name (string) venue.location.lng (string) venue.location.lat (string) venue.categories (string)	

Notes

[1] See documentation on Nominatim at https://geopy.readthedocs.io/en/stable/#nominatim

Data Cleansing

The names of the neighborhoods were obtained by scraping a website. Initial trials with the Nominatim geocoder service showed that some of the London neighborhood names were not usable in their raw state. There were 4 problems, which were solved as follows:

Problem	Description	Solution
Neighborhood name not recognised by Nominatim geocoding service	A name that isn't recognised by Nominatim returns no longitude or latitude. For example, the neighborhood Chelsea is given on the website as 'Chelsea Riverside'	a) Convert some of the names e.g. 'Chelsea' for 'Chelsea Riverside' b) Ignore some of the less well known neighborhoods i.e. do not include in the model.
Neighborhood names combined	For example 'Knightsbridge and Belgravia'	Include the first part of the name only e.g. 'Knightsbridge'
Neighborhood name includes extraneous charactacters	For example, brackets e.g. 'Aldersgate (incl. Cheap)'	Select the substring before the offending character.
Misinterpretation by Nominatim	Nominatim misinterpreted the name and returned data for a location outside of the inner London area. For example, coordinates for 'Holland' returned instead of 'Holland Park'	Drop misinterpreted rows i.e. those returning coordinates outside of the inner London area.

Using the data

Once cleansed, the neighborhood and venue data for London and New York will be combined in a single DataFrame containing both sets of data. Effectively the neighborhoods for London and New York will be treated as a 'super'-city. This ensures that the data is conformed i.e. that the same attribute/columns are considered for each city. A 'City' column will be included so that we can keep track of which neighborhood belongs to which actual city.

The data will be clustered according to the frequency of various types of venue.

The desired outcome is that each cluster will contain a mix of New York and London neighborhoods. In this way it will be possible to say that New York neighborhood x is similar to London neighborhood y because they are in the same cluster. For example:

Cluster 1 may contain:

City	Neighborhood
London	Kennington
New York	Gramercy

Cluster 2 may contain:

City	Neighborhood
London	Kentish Town
New York	Lower East Side

In these results we can say that Kennington in London is similar to Gramercy in New York. Both of those neighborhoods are sufficiently different from Kentish Town in London and Lower East Side in New York as they are in a different cluster.

Methodology

The requirement is to group 'like with like' across the mix of New York and London neighborhoods. K-means Clustering was selected as an unsupervised machine-learning algorithm for its ability to group elements on the basis of similarity to each other.

The method involved 4 stages:

1. Obtaining the names of neighborhoods in New York and London
2. Using the Nominatim geocoding service to obtain the longitude and latitude of each neighborhood
3. Using the FourSquare API to obtain the venues within a 500m radius of the neighborhood
4. Clustering the venue data based on category of venue using the K-means algorithm. This assigns a cluster number to each New York and London neighborhood.

Venue Category to Numeric conversion.

The K-means clustering method operates on numeric data, seeing as the Euclidean distance between examples must be calculated. However, after stage 3 is completed all we have is a list of venues and venue categories held as text. A method is required to convert the venues and venue categories into a set of numerical values for each neighbourhood. This is achieved as follows:

> 4a. Perform 'one-hot' encoding of the category data. For example, if a venue has a category of 'Theatre' then 'Theatre' is created as a column in the data set with a 1 set for that venue. Venues that aren't a theatre get a 0 in that column. All venue categories therefore become columns populated with a 1 or 0 for each venue as appropriate.
>
> 4b. The venue one-hot encodings are summarised for each neighborhood by using the 'groupby' function on each neighborhood. The mean is taken for each 'one-hot' column, so for example, if 10% of the venues in a neighborhood are restaurants then the neighborhood column gets a value of 0.1 in the 'Restaurant' column. This also has the effect of normalising the data, so a large neighborhood with 100 venues gets a proportional score for each venue category, in the same way as a small neighborhood with only 50 venues.

Choosing the number of clusters

When using K-means or any clustering algorithm, it is an arbitrary choice as to the number of clusters to create. Bearing in mind the number of neighborhoods being analysed, a small number like 2 or 3 would not have differentiated the neighborhoods sufficiently from each other. On the other hand, a very large number like 20 may have created clusters that weren't sufficiently different from each other. There was also the danger that many clusters may have been created clusters populated with just New York or London neighborhoods. In the end 10 was selected as this seems to be a reasonable number in the middle.

Method for interpreting results

To find the 'nearest' London neighborhoods to a New York neighborhood, the method is to note the New York neighborhood's cluster number, then to look at the London neighborhoods in the same cluster. The results of the cluster analysis were exported to a spreadsheet to make the lookups easier to perform.

Exploratory Analysis

An early exploration of the London neighborhood data showed that many of the neighborhood names were not recognised by the Nominatim geocoding service and required cleansing. Among the problem were that some neighborhoods had been combined, for example 'Knightsbridge and Belgravia', some had extraneous information after the name e.g. ‘Aldersgate (incl. Cheap)’. Code was added to the notebook to strip out the extraneous information and convert the names into a usable form.

The first time clustering was performed resulted in several clusters containing New York-only or London-only clusters. This may have been due to the fact that there were columns for each type of restaurant, for example ‘German Restaurant’ or ‘Japanese Restaurant’. This removed 15 out of 480 different venue categories.

It is quite likely that cultural differences between New York and London mean that some speciality restaurants e.g. German would be more prevalent in one city than the other. This was resolved by converting any restaurant category into a single category for ‘Restaurant’. After making this change only 2 clusters contained New York-only neighborhoods, one containing two neighborhoods and the other containing one neighborhood. This was considered to be an acceptable result.

Results

There were 2 notebooks for the solution:

Notebook	Url
Obtain London neighborhoods through web-scraping and geocoding via Nominatim geocoding web service	https://github.com/tommymato/Coursera_Capstone/blob/master/GetLondonData.ipynb
The neighborhood venue clustering method	https://github.com/tommymato/Coursera_Capstone/blob/master/Nearest%20Neighborhood.ipynb

For convenience, in addition to the notebook, the cluster data was downloaded to a spreadsheet here: https://github.com/tommymato/Coursera_Capstone/blob/master/NYLonClusters.xlsx

The clusters were as follows:

Cluster #	Number of New York neighborhoods	Number of London neighborhoods
0	109	4
1	2	0
2	87	19
3	1	4
4	0	20
5	57	10
6	31	85
7	8	3
8	5	3
9	1	0

Some specific examples

New York Neighborhood	Cluster #	London 'Nearest Neighbors'
Battery Park City	5	Gospel Oak, Graveney, Plumstead, Regent's Park… and others
Lower East Side	6	Bishopsgate, Camden, Chelsea… and others
Somerville	8	Latchmere, Streatham Wells, Thamesmead Moorings

Anomalies

It can be seen from the above table that cluster 1 has 2 NY neighborhoods and no equivalent London neighborhoods in that cluster. Likewise for cluster 9 which has 1 NY neighborhood and no London equivalents. That makes a total of 3 NY neighborhoods (out of 301) without a London equivalent. The 1% of anomalies is deemed an acceptable rate. It could be said that the 3 neighborhoods in question, Neponsit, Oakwood and Port Ivory are in a class of their own.

Discussion

Further analysis would be required to determine if this clustering technique could be useful in identifying neighborhoods with equivalent investment potential. Other data may be required to make a more complete model, such as demographics, proportion of property types (residential, commercial, industrial, etc.) and data on property prices.

In this clustering exercise we changed all restaurant types to a single 'Restaurant' category in order to remove the effect of cultural differences between the US and UK. Additional cleansing may be performed, for example grouping specific sport venues into a single ‘Sport’ category. There may be many ‘Baseball Parks’ in NY but very few in London. Likewise, Bagel and Donut shops could be categorised as the more London-friendly ‘Bakery’.

Regarding the anomalies listed above, it is still be possible to compare neighborhoods by calculating the Euclidean distance of one neighborhood from another with respect to venue data. This was not performed due to the small number of anomalies present.

Conclusion

The attempt to compare New York with London neighborhoods was successful given that most of the clusters had a mix of New York and London neighborhoods. Whether the venue category cluster analysis is useful to a property developer is open to further analysis. Perhaps the model could be improved by incorporating demographic data (distribution of age, income bracket, etc.) and additional data sets for property types and prices in each neighborhood. The venue analysis is a good first step and encourages further research.

There may be other potential applications, providing guidance to tourists perhaps, and the technique could be applied to comparisons between other cities.