Nearest Neighborhood 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.