Python 3

The Battle of Neighgbourhood

Introduction

The goal of this analysis is to compare different big cities in the world to check their similarity. The similarity will be calsulated using the list of venues retrieved with foursquare analyzing the categories of venues contained in a city and their ditribution within the city area.

Questions that we want to ask with this analysis are: Can city of the world be clustered in groups based just on the type of venues contained in them? Does the result of this grouping make sense? Are those groups related to geography (i.e. all europeans city will belong to the same group)? Can we create a classification of city districts that span across different cities (e.g. will all city have zones with restaurants other with museums etc.. is the venue category distribution a characteristics that repeats in different cities)?

Analisys method

In order to perform this analysis we plan to take a number of cities and use foursquare to fetch the venues that are contained in the cities. Each city will be divided in a regular grid of **Search Spots**. The search spots are the points that will be used to search for venues around. Each search spot togheter with the radius used for the search will become a **City Zone**

The search spots are placed in a regular triangular grid and the radius of the search will be one third of the distrance between points.The points will be distanciated 1000 meters and the search radius of each spot will be 666 meters.

Each spot is a zone of the city for wich a profile will be contructed using the frequencies of the venues categories.

All the zones of all the cities considered will be divided in a number of groups based on the venues category frequency.

Then each city will have a profile based on the distribution of the zone categories and city will be grouped this way. City that will belong to the same group being "more similar"

The data

In this section we will create the code that will be used to fetch the data that will be required for the analysis.

Each city that will be included in the analysis will have a reference point and search grid dimension. The search grid dimension will tell how many search point will be used to difide the city in zones and to get the venues for each zone using the foursquare explpore API.

The grid will be limited and catch mostly the central part of the cities to limit the number of foursquare queries that will be necessary (this is due to the limitations of the free plan that we are using and that limits the number of queries to 950 per day)

The result of this activity will be a DataFrame for each city containing the teched venues and the cooridnates of the zone (search point) used to fetch the venue.

The search zone of one search point has an overlapping with the zone of other adiacent search points, a venue is associeted to the first search point that will intercept it in the search. If the same venue will be fetched by a search in another search point it will be ignored since the venue has already been associated to a zone and cannot belong to multiple zones.

starting data

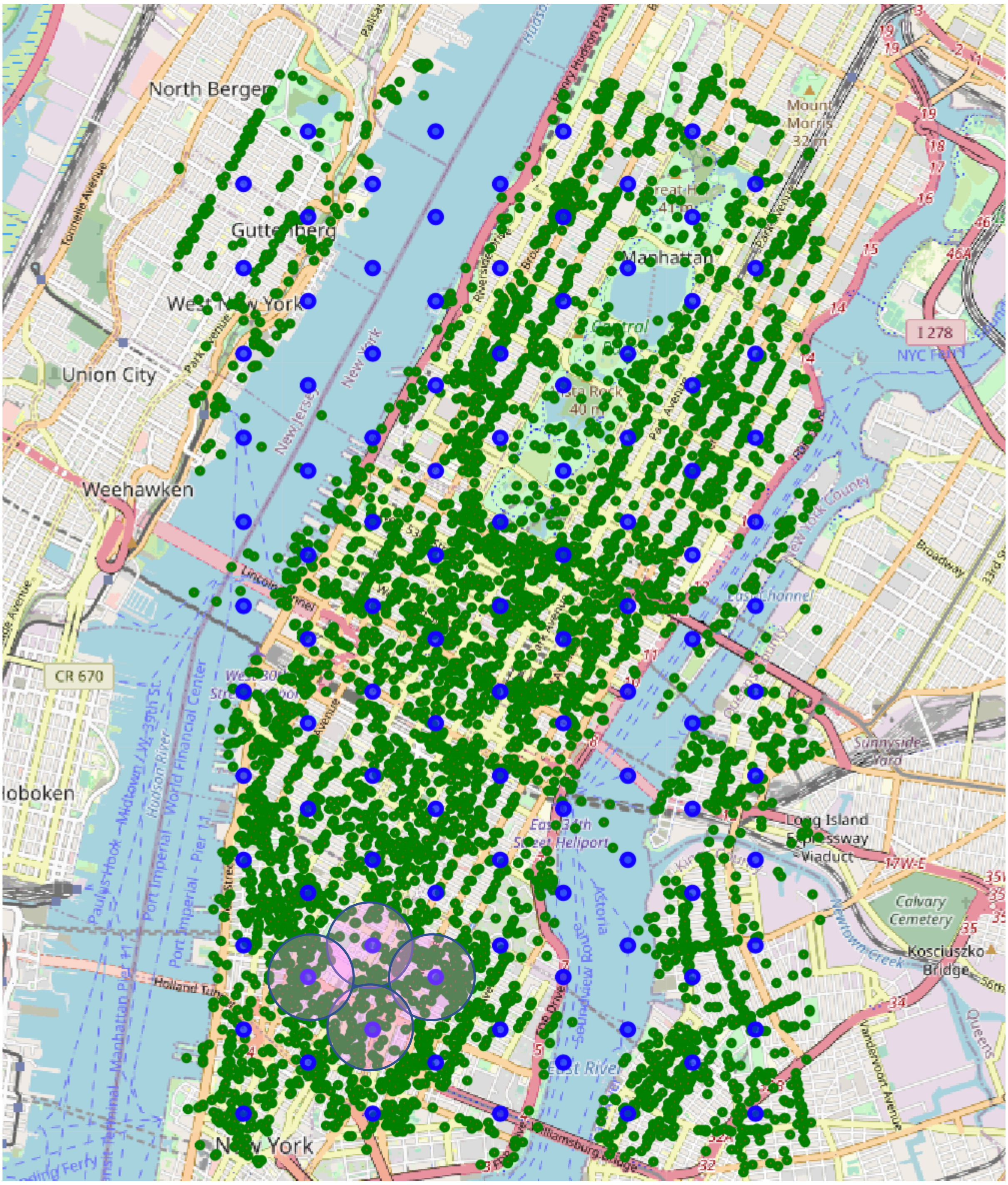
The datasets that will created through foursquare queries and that will create our start data will have the following format

[1]:

	search_spot_lat	search_spot_lon	search_radius	id	Name	Latitude	Longitude	Category
0	41.925131	12.538995	666.666667	4ca7816e76d3a093554e0c6b	Mejo De Betto E Mary	41.925766	12.539925	Roman Restaurant
1	41.925131	12.538995	666.666667	52b0a30f11d20648deaa4034	Inofficina	41.925485	12.535250	Gastropub
2	41.925131	12.538995	666.666667	4b0d105cf964a5208a4323e3	Lanificio 159	41.926047	12.539253	Performing Arts Venue

Each line represents a venue in the city . (search_spot_lat , search_spot_lon) are the coordinates of the center of the zone in wich the venue is inlcuded. The zone is a circle with radius: search_radius centered in that point. The id is the foursquare id of the venue it identifies the venue. Name is the name of the venue Latitude,Longitude are the absolute geographical coordinates of the venue Category is the Category of the venue attributed by foursquare.

For each city we will have a certain grid of searching points and of venues retrieved like in the following picture of New York:



The blue dots represent a search points, the green dots are the retrieved venues.

Each search point has a radius of search around it . The zone delimited by the search radius is a **Zone** in the city (in pink in the picture). We will use these zones to inquiry the nature of the cities and see how similar or different cities are. The zones have the same size in all the cities of our inquiry. Each zone has a small overlap to the adjacent zones. A venue that lay in that overlap is attributed to the first zone that is searched for venues. So each venue is attribute to only one zone.

The grid of search point is a triangular grid to minimize overlap of the search zones.

The Zones are a pure spatial city split rather than an administrative one. This has the advantage to make them more comparable between different cities. A zone in New York has exactly the same size of a zone in Tokyo. The expected number of venues inside a zone depends on factor that are specific of the city (like the position in the city, the density of shops etc.) rather than the way the city administration has divided its territory.

Analysys methodology

We will create a city zone classification. Each zone will be associated to a typology, a label that will identify each zone with a type. We will do this using **Kmeans** clustering algorithm.

We have memorized the "zone" for all the venues that we have fetched for all our cities. We will create a DataFrame that for each zone and for each venue Category will have the number of occurrences of that Category in the zone.

city	search_spot_lat	search_spot_lon	search_radius	American Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	BBQ Joint	Bagel Shop	Bakery	Bank	...	Trail	Tram Station
barcelona	**41.380660**	**2.154853**	**666.666667**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	...	0.0	0.0
		2.169493	**666.666667**	0.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	...	0.0	0.0
		2.184133	**666.666667**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	...	0.0	0.0
		2.198773	**666.666667**	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	...	0.0	0.0
		2.213413	**666.666667**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
...	**...**	**...**	**...**	...	...	...	...	...	...	...	...	...	...	...	...	...
washington	**38.940771**	**-77.047325**	**666.666667**	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	4.0	0.0
		-77.032685	**666.666667**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
		-77.018045	**666.666667**	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0
		-77.003405	**666.666667**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
		-76.988765	**666.666667**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...	0.0	0.0

2028 rows × 127 columns

This will become the **profile** of the zone.

Then we will use these profiles to cluster all the zones of all the cities in a finite number (10) of group.

Each group will be a zone typology, i.e. Zones that lay in the same group are of the same type (at least using our criteria).

Once we have done this we will category the cities in types like we have done for the zones.

We will create a DataFrame that for each city and each zone typology contains the number of occurrence of that zone typology in the city.

Here again we can consider this as a **profile** of the city.

Wee will use agail kmean to cluster all the cities in omogeneous groups and then analyze the results to infer knowledge about cities similarities

Results

Here are the results of our analysis:

[33]:

city	0	1	2	3	4	5	6	7	8	9	city_type
tokyo	0	0	0	2	65	2	0	0	0	31	0
los_angeles	8	3	0	1	0	0	0	0	0	87	1
washington	28	0	0	0	0	2	1	0	0	72	1
boston	24	0	1	0	0	6	0	1	0	75	1
dallas	6	0	0	0	0	1	0	0	0	88	1
moscow	11	0	0	2	0	4	9	1	0	72	1
phoenix	3	0	0	0	0	2	0	0	0	94	1
philadelphia	16	0	0	1	0	1	1	0	0	81	1
barcelona	0	0	0	4	0	1	0	0	30	63	2
madrid	0	0	0	1	0	2	0	0	39	57	2
milan	0	0	42	1	0	4	0	0	0	57	3
rome	0	0	38	1	0	1	0	0	0	58	3
sao_paulo	26	0	0	2	1	17	4	0	0	54	4
london	25	0	0	15	0	6	6	0	1	51	4
new_york	48	0	0	1	0	12	0	7	0	40	4
munich	0	0	4	20	0	1	1	0	0	76	5
berlin	0	0	1	33	0	2	4	0	0	61	5
mexico_city	0	42	0	0	0	2	0	0	0	60	6
istanbul	0	0	0	0	0	3	41	13	0	47	7
paris	0	0	1	55	0	7	0	0	0	37	8

This table has a row for each city. The columns with numbers contains the number of instances of zone of the type of the name of the column.

For example Tokyo has 65 occurrences of zones of type 4.

The column city_type represent the group in which the city has been placed.

Here is the list of the zone profiles.

For each Zone type is shown the list of the top 10 categories presents in such zones (by summing all the category instances across all the zones of such type)

zone type: 0

Coffee Shop	531.0
Italian Restaurant	318.0
Pizza Place	307.0
Bar	300.0
Café	282.0
Bakery	257.0
Hotel	256.0
Gym / Fitness Center	211.0
American Restaurant	198.0
Sandwich Place	173.0

zone type: 1

Mexican Restaurant	234.0
Taco Place	177.0
Bakery	47.0
Coffee Shop	43.0
Seafood Restaurant	43.0
Restaurant	41.0
Convenience Store	38.0
Ice Cream Shop	35.0
Pizza Place	35.0
Bar	30.0

zone type: 2

Italian Restaurant	541.0
Pizza Place	282.0
Café	257.0
Hotel	231.0
Ice Cream Shop	174.0
Plaza	139.0
Restaurant	114.0
Cocktail Bar	101.0
Japanese Restaurant	82.0
Bakery	75.0

zone type: 3

French Restaurant	419.0
Café	336.0
Hotel	328.0
Bar	298.0
Italian Restaurant	261.0
Coffee Shop	220.0
Bakery	203.0
Pizza Place	147.0
Vietnamese Restaurant	131.0
Restaurant	131.0

zone type: 4

Japanese Restaurant	252.0
Convenience Store	225.0
Ramen Restaurant	207.0
Café	188.0
Sake Bar	157.0
Coffee Shop	147.0
Chinese Restaurant	142.0
BBQ Joint	137.0
Italian Restaurant	127.0
Soba Restaurant	122.0

zone type: 5

Hotel	201.0
Clothing Store	184.0
Coffee Shop	174.0
Italian Restaurant	138.0
Café	137.0
Boutique	96.0
Cosmetics Shop	96.0
Plaza	82.0
French Restaurant	80.0
Bakery	78.0

zone type: 6

Café	345.0
Hotel	175.0
Coffee Shop	161.0
Restaurant	123.0
Turkish Restaurant	114.0
Park	88.0
Dessert Shop	82.0
Bakery	82.0
Gym / Fitness Center	64.0
Bar	62.0

zone type: 7

Boat or Ferry	114.0
Café	49.0
Art Gallery	35.0
Park	26.0
Coffee Shop	24.0
Restaurant	19.0
Nightclub	17.0
Seafood Restaurant	14.0
Gym	11.0
Bar	10.0

one type: 8

Spanish Restaurant	337.0
Tapas Restaurant	213.0
Restaurant	201.0
Hotel	161.0
Bar	140.0
Café	116.0
Mediterranean Restaurant	115.0
Bakery	115.0
Coffee Shop	91.0
Plaza	75.0

zone type: 9

Café	754.0
Coffee Shop	592.0
Italian Restaurant	553.0
Hotel	538.0
Park	510.0
Pizza Place	486.0
Restaurant	446.0
Bar	409.0
Bakery	385.0
Mexican Restaurant	300.0

As we can see the zone of type 4 seems quite typical of a Japanese city and is logical to see that kind of zone present only in Tokyo . Zone 2 is an Italian Zone because of the disproportion of Italian restaurants and Pizza places .

If we look at our result table we see that Zone 2 is massively present in the two Italian cities included in the analysis and scarce in the remaining cities.

Zone 1 is a Mexican one, the only place we found it is Mexico City and Los Angeles.
Zone 3 is French (but. Is quite present in German cities too), Zone 8 is a “Spanish” one.

There are other zone types that are less obvious.

Take type 5 which caused three city with apparently nothing in common to be placed together (London, New York and Sao Paulo).
Type 5 seems to have a big presence of clothing stores and hotels.
The other cities in which zone of type 5 is present are Paris, Milan and Boston and Moscow. Maybe we have to reconsider our perception of Paris and Milan as Fashion Capitals in favor on New York , London and Sao Paulo (even if we should distinguish between quantity and quality).

Conclusions

The conclusions of our analysis are a little deceiving. We were in search of some obscure connections between different cities around the world. The result of our analysis is that city that share the same country/culture are more similar in term of venue distribution than exotic ones. Which seems to be quite an obvious result.
One problem could also be the kind of data we use that has too much localization. For example, it would have been interesting to compare city zones by the distribution of more generic venue categories, one in which an Italian Restaurant and a French Restaurant are included in a more generic category Restaurant. This would move the comparation of the zones more toward the distribution of generic categories of venues (restaurants, museums, shops etc.). This, maybe, would have cause more similarities in distant cities that reside in different continents and lead to less obvious insights on City similarities. Unfortunately, the effort necessary to review all the almost 700 different categories found by foursquare in the cities under analysis was beyond the budget of this small project.