Coursera Capstone Project Report

INTRODUCTION

Rome is one of the most visited cities across the World and our intent, as a popular Italian Data Science company, is to develop a tool able to provide some information about the main Rome municipalities, in term of their tourist attractions.

IDEA

In particular we want to develop a model that exploits the Foresquare location data to measure how the different district are similar or dissimilar in terms of touristic venues. By doing so, the tourist agency will be able to offer a customized and client-based tour across different roman neighborhoods just on the basis of their preferences (for instance, there are persons more interested in museum and cultural activities while other persons are more curious to taste local dishes or even go shopping).

In practice, the aim of the project can be reformulated in answering the above listed questions, for example:

- If I like going shopping which is the best shopping district?
- If I want to see a lot of Museum and cultural initiatives but at the same time try some local dishes, in which district I should book my Hotel?

SOLUTION

The above described purpose can be achieved by groping the different neighborhoods of Rome on the basis of the venues categories retrieved by the Foursquare API. Once all venues of a given district are collected, we can apply a clustering algorithm (ex. K-means) to group Rome municipalities into categories belonging to a specific kind of cultural tour.

DATA

One of the main steps related of our project is to accurately select and collect the data starting from which the model will be developed.

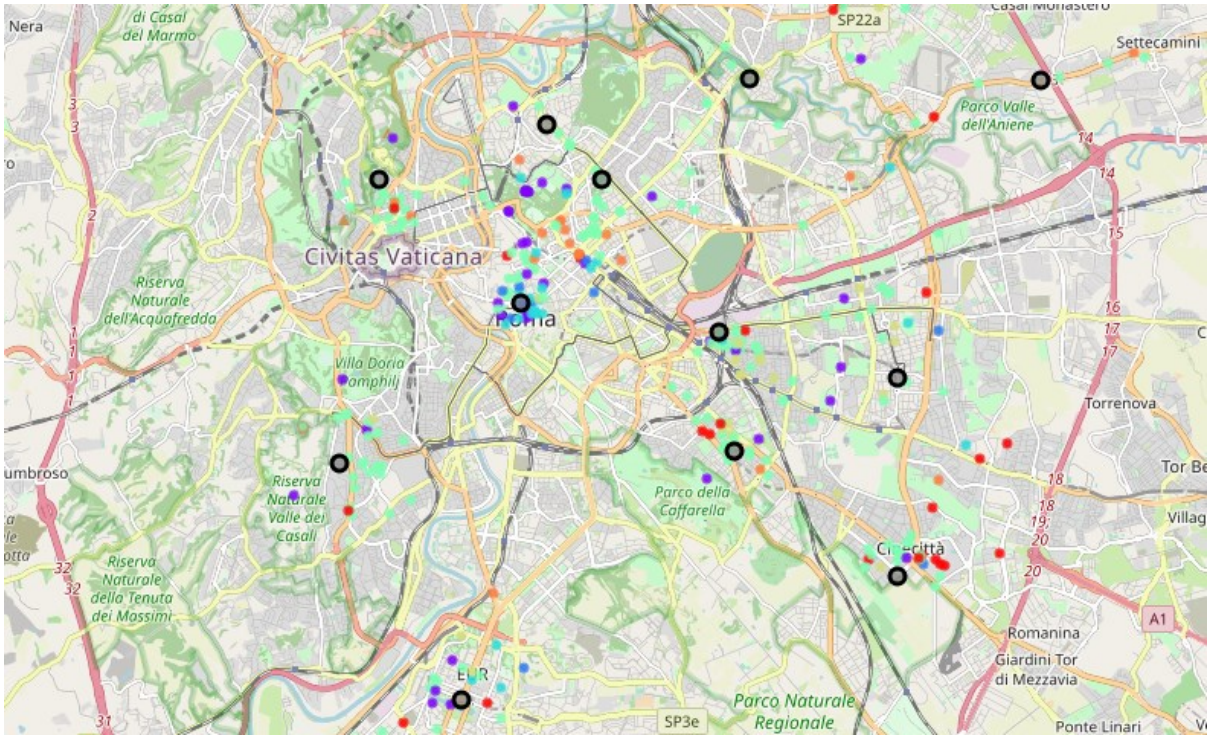
Fist, the location data of all Rome municipalities must be retrieved. Rome is composed by 19 municipalities (https://it.wikipedia.org/wiki/Municipi_di_Roma) and by knowing their name we can retrieve their geospatial coordinates with an ad hoc Python library.

Later these coordinates will be passed to the Foursquare API (https://it.foursquare.com/) and will be used to get information about venues in each neighborhood. For instance, the Foresquare API will return a bunch of information about each venue but we are interested only on the venue category. To limit the number of returned categories only some macro-categories will be investigated such as “food”, “art” and “shopping” and so on.

ORIGINAL FOURSQUARE API CATEGORY	MACRO CATEGORY CONSIDERED
BAKERY	***Food***
ART GALLERY	***Arts & Entertainment***
JAZZ CLUB	***Arts & Entertainment***
CAFÉ	***Food***
SHOPPING MALL	***Shop & Service***

EXPLORATORY DATA ANALYSIS AND METHODOLOGY

It will be useful by starting to plot the map of Rome municipalities by of the Folium python package to see how they are distributed and located geographically. Then we should use markers of different colors to over plot on the neighborhoods map all venues belonging to the macro-categories we have selected.



The methodology adopted relays on a clustering algorithm since unsupervised learning seem the best approach suited for this problem. K-means was chosen for clustering because it's a reliable and affordable machine learning technique often used a first guess to look how the solutions space looks like.

The choice of the k-parameters is done by constructing the classic elbow curve and looking at the number of clusters that minimize the metric which is chosen as a measure of error.

At this point the data should be prepared and pre-processed. This can be done by the so call "one-hot encoding" procedure to encode categorical data into numeric. The results will be a large dataframe containing 1 if a given category is present in the neighborhood, 0 otherwise.

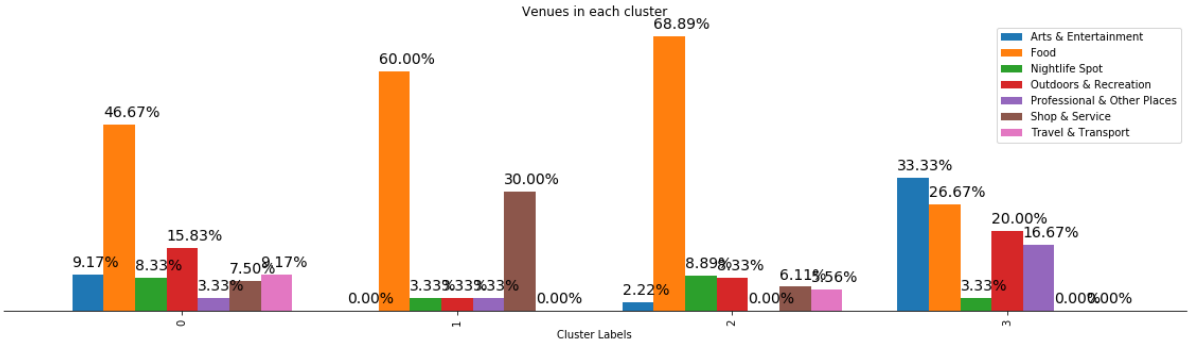
	Municipe	Arts & Entertainment	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Shop & Service	Travel & Transport
327	Portuense	0	1	0	0	0	0	0
55	Parioli	0	0	0	0	0	1	0
313	Portuense	0	1	0	0	0	0	0
271	Eur	0	0	0	1	0	0	0
105	Monte Sacro	0	1	0	0	0	0	0

Once the clustering is performed, we should be able to classify each municipe according to its touristic attractions. For instance, since both neighborhoods "Appio-Latino" and "Eur" have the category Food as the most common and the category Shop & Service as the secondo, they will be put more likely in the same cluster.

	Municipe	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
0	Appio-Latino	Food	Nightlife Spot	Shop & Service	Outdoors & Recreation	Travel & Transport	Professional & Other Places	Arts & Entertainment
1	Centocelle	Food	Nightlife Spot	Shop & Service	Outdoors & Recreation	Arts & Entertainment	Travel & Transport	Professional & Other Places
2	Centro Storico	Arts & Entertainment	Food	Outdoors & Recreation	Professional & Other Places	Nightlife Spot	Travel & Transport	Shop & Service
3	Cinecittà	Food	Shop & Service	Professional & Other Places	Outdoors & Recreation	Nightlife Spot	Travel & Transport	Arts & Entertainment
4	Eur	Food	Shop & Service	Outdoors & Recreation	Arts & Entertainment	Nightlife Spot	Travel & Transport	Professional & Other Places

RESULT VISUALIZATION

Finally, we end up with four clusters, as shown in the picture above. According to these results we can define three "type" of perfect tourist for Rome.



1. Cluster 0: All-round lover
2. Cluster 1: Food and shopping lover
3. Cluster 2: Just food lover
4. Cluster 3: Art and Outdoors activities lover

As we can see also from the map, most of the Rome Municipalities fall in the cluster 0 and 2. Of course, Rome is a must for people who loves food but that's not a very original discovery. Anyway, it offers also a lot of cultural spots especially in the town center which is full of museums and historical buildings.

Food is the most recurrent categories, and this is due in part to the Foursquare API categories classification but also to the type of city we are looking at. Rome compared to other cities has a bunch of restaurant and café. For this reason when looking at the results is important also to look at the second most common category to establish the tourist "type.

DISCUSSION

Of course, here we should point out the relevance of the choice of the k parameters on the obtained results. Obviously, a higher value of the number of clusters would have improved the accuracy of the separation between them but I would be affected the likeliness and flexibility of the model. In fact, the scope of this work was to find a city tour that corresponds to a general profile of a tourist and not to a specific and too much customized one.

A big model limitation of course is constituted by its simplicity since we don't have created ad hoc categories, but we have used the built-in categories in the Foursquare API to make the classification. Given the large amount of these categories by doing so we end up with a result that is less accurate. A suggest for future work is to search directly for specific categories venues (the ones of the tourist sector for example) that suit better your purpose where and do not take so wide range of categories instead.

It's important also to notice that algorithm other than K-means were not tested.

CONCLUSIONS

In this work a clustering of the main Rome municipalities was performed to find what venues categories they have in common. The output of the model developed could be use by local tourist agency to propose a customize tour based on the people preference.