Searching where live in Lima

Diana Quintanilla
Marzo 2019

1. Introduction

a. Background

For the last few years I have been living in the city of Lima. Specifically in a remote district, for causes of independency I want find a apartment according to my preferences about the neighborhood. For example, one might want to look at all areas which has proximity to pubs, cafes, public transport etc.

b. Problem

¿Where in Lima city will I buy or alquile a apartment?
To figure this out I want to compare the various neighbourhoods in Lima to see how similar they are, and finally figure out if there is a neighbourhood with my preferences. I will do this based on the different venues and ammenities available in the direct vicinity of the various neighbourhoods.

c. Interest

People who are looking to find apartment in Lima, especially the people who are starting their apartment search and has certain preferences.

2. Data

I used two datasets for the above problem:

A. A list of neighbourhoods in Lima at [http://www.movistar.com.pe/documents/10182/24285/3G.pdf/4a97babb-9b45-4c39-bcc4-d4acf6155ca8]. This includes
   i. Name of neighbourhood
   ii. Name of district
   iii. their geographic coordinates

B. The second dataset I used is Foursquare venues data with categories, this will be used in collaboration with the above dataset.

3. Methodology

This section is divided into three main parts, data wrangling/Cleansing, Exploratory analysis and machine learning method used to solve the problem.

Data Wrangling/ Cleansing: Data about neighbourhoods obtained from Movistar website in format pdf, that I have converted to csv format, after I only select center neighbourhoods to Lima city and I drop columns not used and rename columns, the result is shown in Fig 1.

Fig 1. Part of dataset about neighbourhoods

	Código UBIGEO	Centro poblado	Distrito	Provincia	Región	LATITUD	LONGITUD	CLASIFICACIÓN INEI
717	1501010001	LIMA	LIMA	LIMA	LIMA	-12.046679	-77.03230	URBANO
719	1501040001	BARRANCO	BARRANCO	LIMA	LIMA	-12.149599	-77.02474	URBANO
720	1501050001	BREÑA	BREÑA	LIMA	LIMA	-12.056910	-77.05366	URBANO
737	1501080001	CHORRILLOS	CHORRILLOS	LIMA	LIMA	-12.174429	-77.02482	URBANO
741	1501130001	JESUS MARIA	JESUS MARIA	LIMA	LIMA	-12.069999	-77.04524	URBANO

Exploratory Analysis: The dataset has 29 districts and 53 neighborhoods, for see this neighborhoods shown in a map (Fig 2).

Fig 2. Neighbourhoods in a map

Also the data about places was obtained from Foursquare API, then I combined all result about venues near to neighbourhoods from above dataset (Fig 1). You can

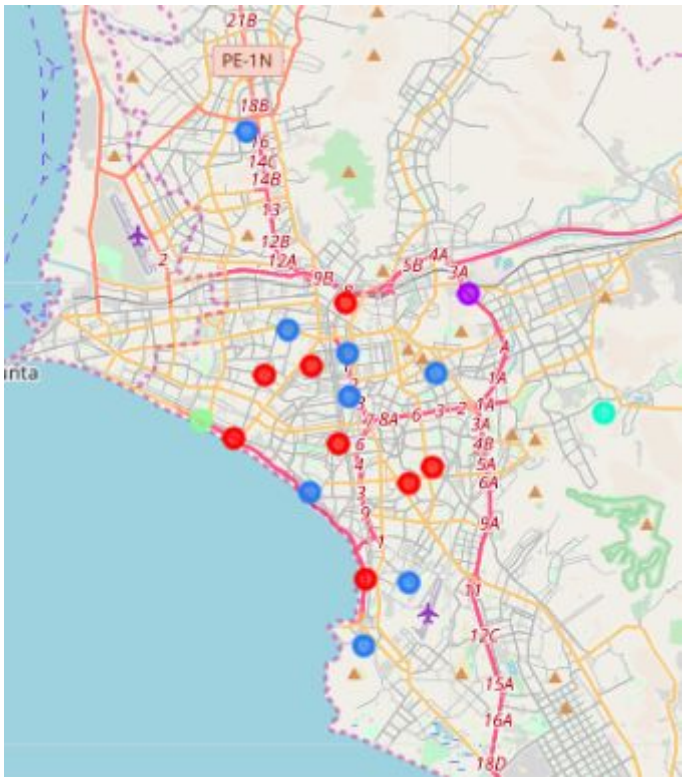
see this result in the Fig 3. How venue and zone are categories, I take this categories information and create a dataframe with a one hot encoding of these data.

Fig 3. Results from Foursquare

	Neighborhood	Zone	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	LIMA	URBANO	-12.046679	-77.0323	Plaza Mayor de Lima	-12.045983	-77.030565	Plaza
1	LIMA	URBANO	-12.046679	-77.0323	Casa Bernardo O'Higgins	-12.047588	-77.032498	Art Gallery
2	LIMA	URBANO	-12.046679	-77.0323	Palacio Municipal de Lima	-12.045283	-77.030917	City Hall
3	LIMA	URBANO	-12.046679	-77.0323	Teatro Municipal de Lima	-12.045770	-77.034839	Theater
4	LIMA	URBANO	-12.046679	-77.0323	Galería Municipal Pancho Fierro	-12.045495	-77.031341	Art Gallery

I created five different clusters based on the venue categories provided by data and I plotted all districts the on the map with your cluster label. (Fig 4)

Fig 4. Clusters



Machine Learning: I decide to use K means Clustering for the problem because after the exploratory analysis it was obvious that we can categorise neighbourhoods into different homogenous clusters based on the venue categories. Further on, we

need to create our dataset in a format where rows will represent the sectors and all columns will be venue categories and we will calculate the score for every category in each sector as per their presence or absence in the sector. We sort the categories for each sector in order of their occurence, for example if a sector has more cafes and pubs, the sector will get higher score for the cafes and pubs and less for other categories. The previous step will be sufficient to execute data mining technique I chose to accomplish this task(K mean clustering). K means clustering will segment the data into groups where groups are similar among themselves but different from other groups in terms of occurence of venue categories. For our dataset, K mean clustering will append another column to the dataset which will depict the cluster number, similar sectors will be grouped together.

4. Results

I obtained 5 clusters from K means clustering executed in the previous step.

Cluster 1. Contains many restaurants and diners, as well as some other services, like art gallery

Cluster 2. Seems to be dominated by cofee shop and mallS

Cluster 3. Is mostly linked to open areas for bar,parks and music

Cluster 4. Is the most homogeneous one (it only contains one neighborhood)

Cluster 5. Is characterized by shops, gym and restaurants

The above clusters segment the data into 5 homogeneous groups which exhibit similarities among themselves in terms of venue categories are being heterogenous from the other groups. For example, If an individual is interested in staying near to places with proximity to pubs. bar and cafes, Cluster3 is a great choice or if restaurants and art are of interest to someone, sectors under cluster 1 are the most suitable.

5. Discussion

For a major differentiation between clusters can search features as population, buildings, bus stop, but this data in this moment not was founded.

6. Conclusión

As demonstrated in the Result section, we can effectively help in the apartment search process for our customers by clustering the area based on the venue categories and filtering the cluster to narrow down their search area. Customers can then prefer to focus their apartment search by analysing various trade offs. Those who are confused can efficiently compare various areas and choose the one which is most suitable for them.