SEGMENTING AND CLUSTERING LISBON'S NEIGHBORHOODS

RICARDO ENES

Discussion and Background

When you hear of Europe's hot destinations, Portugal is surely at the top of the charts. At the center of this beautiful country you find its capital, Lisbon. The world is full of wonderful and exciting cities to visit, and Lisbon is one that should be on everyone's bucket list. It has everything within easy reach, from galleries and museums to the seaside and charming nearby villages. Lisbon is also extremely budget-friendly, making it easy to navigate and explore in depth.

Soon I will be leaving my hometown, Braga, in the north of Portugal to start my adult life in Lisbon, working on 9 to 6 job. I am that kind of person that needs to go out and socialize on a daily basis, in order to maintain its sanity. This is where we have a problem... I know nothing about Lisbon's hot spots, due to only have visited the city three times. As such, I am "combining business with pleasure": IBM's data science capstone project with the urge to know more about Lisbon's most cultural and coolest neighborhoods.

This project aims to nullify these urges by segmenting and clustering Lisbon's major metropolitan area, in terms of social and cultural venues. With this in mind, I will gather data by the means of web scraping and Foursquare's API.

This endeavour will mostly be useful to myself, but it would be very rewarding if by any chance (close to none), common people like myself could make use of this data and get a bigger picture about what Lisbon has to offer and where to find it.

Data Gathering and Methodology

The first step is acquiring Lisbon's metropolitan area neighborhoods names and its geographical coordinates. To solve this need I have done web scraping and used Foursquare's API.

The website from where the neighborhoods names were collected using BeaufifulSoup library,

- `https://pt.wikipedia.org/wiki/Lista_de_freguesias_de_Lisboa`

The next step is fetching the coordinates for the given neighborhoods, using Geolocator from the GeoPY library.

Finnaly, we create a dataframe that includes all of the information gathered so far. Obviously, the dataframe is then properly cleaned, removing any value that does not suite our needs, such as invalid/wrong neighborhoods names and coordinates.

Having finished the basic data gathering procedures, we utilize the Foursquare's API to query about the venues of each one of Lisbon's neighborhoods. This includes, the type of venue (or category) and it's name.

All of the data gathered will then be put together into a single dataframe including the following columns:

(1) 'Neighborhood'

(2) 'Venue Name'
(3) 'Venue Latitude'
(4) 'Venue Longitude'
(5) 'Venue Category'

We will then analyze the data gathered to find out the most common venues per neighborhood using one hot encoding and utilize the proper visualization techniques to take some conclusions has we move further is this project.

Lastly, we will try to cluster the neighborhoods based on the venue categories and use K-means clustering. To find the best value for the number of clusters, we will use the elbow method, therefore guaranteeing the best possible clustering outcome. In sum, our expectations is that based on the existing or non-existing similarities between neighborhoods, we will be able to divide them into several different clusters.