THE BATTLE OF THE NEIGHBORHOODS!

Coursera Capstone Project

Rodrigo Link Federizzi

April 02, 2021

1. Introduction

After the completion of a certain Data Science course, an young and brave Brazilian data scientist is flooded with job offers from all over the world. After thoughtful consideration, he realizes he really likes where he currently lives and would like to have the same kind of venues around him wherever the job is.

From the job offers, he selects a few cities that seem interesting and sets out to compare his current city's attractions to the neighborhoods of each new possible location, trying to identify similarities. Two approaches will be used:

a. Based on the similarity, we will present a list of the three neighborhoods from each city that better match his current location, display a metric of the similarity between them and present three new venue types what he should try out at the new location.

b. A clustering technique will be used to group all the neighborhoods into sets and then classify the current location. This way he can see if the ones that better match his current neighborhood, even if they are all in the same city. This would mean that particular city is a better fit over all.

2. Dataset

The first input needed from the user is his current location, so that we can create the standard we will use to compare the cities. *The format is 'Neighborhood, City, Country'*. The second

input we need is the list of possible cities. *The format is a list with each 'City, Country'*. From these, we have to manually find and webscrape the list of neighborhoods or districts and their coordinates.

With all the coordinates, we will use the Foursquare API to explore 100 venues around a 1km radius from the center of each neighborhood. We'll construct a Dataframe where each row is a neighborhood from a different city and each column is the percentage of each type of venue.

For the first approach, we will compute the correlation between each row and the data from the current location. The end result is a list with the 3 top neighborhoods that better match, the correlation, and the top 3 venue types that are present at the new location, but not at the current one.

The second approach is to use clustering to identify a group of neighborhoods that is closer to the original location. We'll try a different number of clusters to optimize the separation. We'll fit the current location data into the model and present the list of neighborhoods that are in the same cluster.

3. Methodology

Our initial input is the list of cities where the Data Scientist is considering job offers and his current location. We want to compare how similar are the attractions from the many neighborhoods at different cities from different countries. We will get the list of attractions from Foursquare, but we need to get geographic coordinates (latitude and longitude) from each neighborhood. But our initial data is just the list of cities.

We were able to find the list of districts that are part of each city on the internet (Wikipedia and local government pages). To extract the information, we employed the *BeautifulSoup* library to read, transform and structure the data into a table we could use in the following step.

The following procedure was to obtain the coordinates of central geographical point at the center of each neighborhood from the list we retrieved from the webpages. To do that, we employed a library called *geopy*, where we can get latitude and longitude of any address we query. The format we used was ‘Neighborhood, City, Country’. We found comments saying that this library was unreliable and we can assure that it really is. The main bottleneck of the project was this step, and we were able to run it completely only once, but we were smart enough to save it to a file.

With the coordinates, we queried the Foursquare API for the top 100 activities from each neighborhood. The main goal was to get a list of the types of venues in the area and count how many of each kind were close by. There were some neighborhoods with less than 100 results, but this was expected, because some of the areas are in the suburbs or rural areas of the cities. From the

list and count of attractions, we were able to identify the composition of each neighborhood. This is the information used to study the similarities.

For our first approach, the idea was to calculate the Pearson Correlation between the composition of each neighborhood with that from the current location of the Data Scientist. This returns a number between one (the composition is exactly the same) and minus one (they behave exactly opposite to each other), with zero meaning no correlation at all.

We studied each city, looking at the three neighborhoods with the highest correlation to the current location. For each one, we looked up three types of venues that are unavailable at the original place, to see what new activities are possible if he decides to move.

For the second approach, we used Unsupervised Machine Learning to see if the algorithm was able to group all the neighborhoods in clusters with similar compositions. We decided for k-means clustering. There was not a clear indication for the optimum number of clusters, so we tried the best with what we got. At the end, we ended up with 53 neighborhoods in the same group as the current location.

4. Discussion

For the city of Porto Alegre, Brazil, the three neighborhoods with the highest correlation where Cidade Baixa (0.47), Praia de Belas (0.41) and Floresta (0.40). For the city of Wellington, New Zealand, the three neighborhoods with the highest correlation where Horokiwi (0.63), Oriental Bay (0.46) and Pipitea (0.43). For the city of Stockholm, Sweden, the three neighborhoods with the highest correlation where Riddarholmen (0.52), Gamla Stan (0.51) and Älvsjö (0.40). For the city of Vancouver, Canada, the three neighborhoods with the highest correlation where Downtown (0.50), West End (0.44) and Fairview (0.23).

We don’t believe the three types of venues were really relevant. From our own experience at the current location, many of the venues that were absent don’t show because there are more than 100 activities in the area. ‘Hotel’, ‘Bar’ and ‘Buffet’ are available, but they probably are not high ranked in Foursquare.

The clustering technique brought some interesting insights. 20 of the 53 members of the cluster were from Stockholm, 18 from Vancouver, 14 from Wellington and only 1 from Porto Alegre. If we combine with the information from the first approach, 5 of the 10 first results are from Stockholm.

5. Conclusion

We developed a notebook to try to help to identify similarities between the current location of a Data Scientist and many possibilities from the job offers he is considering. We used webscraping, data transformation and APIs to be able to have a considerable dataset to analyze the problem.

Using correlation and clustering techniques, we end our report with the following suggestions:

All things equal, the order of choice for the Data Scientist should be:

1. Stockholm (either Riddarholmen or Gamla Stan);

2. Vancouver (Downtown)

3. Wellington (Oriental Bay)