Capstone Project
Sao Paulo x New York

Introduction

This work is to provide the details of my final peer reviewed assignment for the IBM Data Science Professional Certificate program – Coursera Capstone.

1. A description of the problem and a discussion of the background - In the Americas there are many important cities that perform a leading role in the world economy. In particular we have the cities of New York, in the USA and the city of Sao Paulo in Brazil. Both cities have many things in common, like the fact that both are a melting pot of different cultures and both where build by many immigrants from various countries in Europe, Asia, Africa. In modern days, it would be helpful to have a way to compare the neighborhood of those cities to help the expatriates to chose places to go, to go for short of long assignment knowing in advance what they can encounter. Research Question: What are similar neighborhoods I can find in Sao Paulo and New York City?

Brief information about both cities:

- São Paulo is a municipality in the Southeast Region of Brazil. The metropolis is an alpha global city (as listed by the GaWC) and the most populous city in Brazil, the Western Hemisphere and the Southern Hemisphere, besides being the largest Portuguese-speaking city in the world, and the largest city south of the Tropic of Capricorn. The municipality is also the world's 12th largest city proper by population. The city is the capital of the surrounding state of São Paulo, the most populous and wealthiest state in Brazil. It exerts strong international influences in commerce, finance, arts and entertainment. (source: https://en.wikipedia.org/wiki/São_Paulo)
- New York (NY), is the most populous city in the United States. With an estimated 2018 population of 8,398,748 distributed over a land area of about 302.6 square miles (784 km2), New York is also the most densely populated major city in the United States.[10] Located at the southern tip of the state of New York, the city is the center of the New York metropolitan area, the largest metropolitan area in the world by urban landmass[11] and one of the world's most populous megacities,[12][13] with an estimated 19,979,477 people in its 2018 Metropolitan Statistical Area and 22,679,948 residents in its Combined Statistical Area.[3][4] A global power city,[14] New York City has been described as the cultural,[15][16][17][18][19] financial,[20][21] and media capital of the world,[22][23] and exerts a significant impact upon commerce,[21] entertainment, research, technology, education, politics, tourism, art, fashion, and sports. (source: https://en.wikipedia.org/wiki/New_York_City)

In Module 3, we explored New York City. In the Week 3, I have explored the city of Sao Paulo. For both we segmented and clustered their neighborhoods. Both cities are very diverse and are the financial capitals of their respective countries. I decided to adopt the interesting idea to compare the neighborhoods of the two cities and determine how similar or dissimilar they are. Is New York City more like Sao Paulo? I think this is something that will be interesting for many travellers, business workers that need to go to those two Cities. In this project, we will study in details the area classification using Foursquare data and machine learning segmentation and clustering.

Using segmentation and clustering, the goal is to determine:

1. Similarity or dissimilarity of both cities.
2. Classification of area located inside each city and find similar clusters.
3. Generate data that can be used in the future for apps for the travelers

Data

We will need geo-location information about that specific borough and the neighborhoods for the two cities. In Sao Paulo I found that in a WiKi and built up a data base. For New York, the data was acquired based on Module 3.

We will need data about different venues in different neighborhoods of each specific borough. In order to obtain that information, we will use "Foursquare" locational information. Foursquare data will be used for segmentation and clustering.

Using segmentation and clustering, the goal is to determine:

1. Similarity or dissimilarity of both cities.

2. Classification of area located inside each city and find similar clusters.

3. Generate data that can be used in the future for apps for the travelers.

Methodology

I am going to use the same methodology as in Week 3 lab. Below are some considerations:

We already have a list of the principal boroughs/neighborhoods of the two cities (shown in the maps above).

We will use the Foursquare API to explore neighborhoods in both cities, New York and Sao Paulo

Explore function to get the most common venue categories in each neighborhood,

Run a Clustering Algorithm to group the neighborhoods into clusters

K-means clustering algorithm will be used. Folium library to visualize the neighborhoods in New York and Sao Paulo and their emerging clusters.

Using Foursquare API to get venues at surounding area of both Sao Paulo and New York areas

Results

Using the methodology, I was able to determine for Sao Paulo and for New York (Manhattan):

- 10 Most common venues per neighborhood (Sao Paulo and Manhattan)
- Sao Paulo and Manhattan Clustering using K-means
- Sao Paulo and Manhattan Clustering visualization
- Sao Paulo and Manhattan Clusters selection for comparison

Discussion

With the data on the clusters for each cities above, we can have an idea of each one and the type of venues categories (most common) would you find in each cities. Our original idea is to use this data to help the expatriates to chose places to go, to go for short of long assignment, based on the similarities of the type of Neighborhood they will find.

We can certainly say that Sao Paulo has many "Bairros" (Neighborhoods) that are very similar to what you would find in Manhattan and the data helps to identify them

In general, with the initial data obtained for this project we can make the following assumptions for each cluster: If you are looking for places to stay and have many alternatives nearby:

- Cluster 2 and 5 in Manhattan
- Cluster 2 and 4 in Sao Paulo

One idea to enhance this initial project would be to go deeper in the correlation of the Neighborhood by finding a quantitative way to identify and distinguish their difference based on most common venues founded via Foursquare. Another idea, is to design an App where the user could give her/his most important venue that she or he would like to find in the target Neighbourhood at the city and run this analysis and come back with possible suggestion. I think this would be very helpful and can be done in a future project.

I believe that this project is a first step towards a quantitative and systematic comparison of the different cities, and specially that is bring information about the city of Sao Paulo, which I could not find any similar work done before.

Conclusion

Although I think the goals of this project were met (Applied Data Science Capstone Project) there is definitely room for further improvement and development as discussed above.

My initial experience as a Data Scientist trying to do this project and others, is that the time you spent in getting the right data to answer your initial "Research Question" represents the majority of the time spent. The right definition of the problem to be solved is key to make sure you go after the right data needed. In this aspects I may say from my experience that the Data Science work and methodology is very similar to scientific methodology, and the risks associated with finding out that your objective is too aggressive and you will need more analysis / data to accomplish are present.

Paulo Cesar Calabria

Sao Paulo, SP - Brazil pcalabria@fei.edu.br

Addendum

This capstone project will be graded by your peers. This capstone project is worth **70%** of your total grade. The project will be completed over the course of **2 weeks**. Week 1 submissions will be worth **30%** whereas week 2 submissions will be worth **40% of your total grade**.

For this week, you will required to submit the following (23/Dec):

1. A description of the problem and a discussion of the background. (**15 marks**)
2. A description of the data and how it will be used to solve the problem. **(15 marks)**

For the second week, the final deliverables of the project will be (27/Dec):

1. A link to your Notebook on your Github repository, showing your code. (**15 marks**)
2. A full report consisting of all of the following components (**15 marks**):

- Introduction where you discuss the business problem and who would be interested in this project.
- Data where you describe the data that will be used to solve the problem and the source of the data.
- Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any

inferential statistical testing that you performed, if any, and what machine learnings were used and why.

- Results section where you discuss the results.
- Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.
- Conclusion section where you conclude the report.

3. Your choice of a presentation or blogpost. (**10 marks**)