Capstone Project

The Battle of Neighborhood

Segmenting and Clustering Neighborhoods of San Francisco and

Los Angeles

By,

Charles C Gomes

Contents

Introduction

San Francisco and Los Angeles are two major cities in California.

Brief information about both cities:

- San Francisco: officially the City and County of San Francisco, is a city in, and the cultural, commercial, and financial center of, Northern California. San Francisco is the 13th-most populous city in the United States, and the fourth-most populous in California, with 883,305 residents as of 2018.

- Los Angeles: officially the City of Los Angeles and often known by its initials L.A., is the most populous city in California, the second most populous city in the United States, after New York City, and the third most populous city in North America. With an estimated population of nearly four million,[11] Los Angeles is the cultural, financial, and commercial center of Southern California.

Objective

In this project, we will study in details the area classification using Foursquare data and machine learning segmentation and clustering. The aim of this project is to segment areas or neighborhood of San Francisco and Los Angeles based on the most common places captured from Foursquare.

Using segmentation and clustering, we hope we can determine:

1. The similarity or dissimilarity of both cities
2. Classification of area located inside the city whether it is residential, tourism places, or others

Data

The data for neighborhoods is acquired from following sources

Los Angeles - https://data.lacity.org/api/views/nwj3-ufba/rows.csv?accessType=DOWNLOAD

San Francisco - https://data.sfgov.org/api/views/xfcw-9evu/rows.csv?accessType=DOWNLOAD

Additionally Foursquare data api was used for getting different kind of venues for segmentation and clustering.

Results

Cluster 1: San Francisco: Tourism

Cluster 2: San Francisco: Residental and Tourism

Cluster 3: San Francisco: Tourism

Cluster 1: Los Angeles: Tourism

Cluster 2: Los Angeles: Residential based on the Park, Convenience Store and Yoga Studio.

Cluster 3: Los Angeles: Mixed.

Discussion

Based on cluster for each cities above, we believe that classification for each cl uster can be done better with calculation of venues categories (most common) i n each cities. Referring to each clsuter, we can not deterimine clearly what repr esent in each cluster by using Foursquare - Most Common Venue data. What is lacking at this point is a systematic, quantitative way to identify and distinguish different district and to describe the correlation most common venues as record ed in Foursquare. The reality is however more complex: similar cities might ha ve or might not have similar common venues. A further step in this classificatio n would be to find a method to extract these common venues and integrate the s patial correlations between different of areas or district. We believe that the cla ssification we propose is an encouraging step towards a quantitative and system atic comparison of the different cities. Further studies are indeed needed in orde r to relate the data acquired, then observe it to more meaningful and objective r esults.

Conclusion

With the help of Foursquare API, we were able to capture the venue informatio n and using venue information, we can figure out the similarities or dissimilariti es of San Francisco and Los Angeles. We did classification of Neighbourhoods as Residential, tourism or Mixed. In conclusion, both cities San Francisco and Los Angeles have tourism as similarity as well as there are some residential are as. It is somewhat clear that in San Francisco, the residential and tourism neigh borhoods are mixed compare to Los Angeles.