Comparison of Neighbourhoods of Bangalore and San Francisco

By

Samriddh Lakhmani

Contents

Introduction

This project aims to compare the Neighborhoods of two cities that are Startup Capitals of their respective countries. For which we have chosen, Bangalore in India and, The Silicon Valley, San Francisco in United States of America.

The primary objective of this project is to help Startup offices in San Francisco on deciding a location in Bangalore, suitable to open a branch office. A lot of startup from The Silicon Valley look to open office's in Bangalore, and we wish to provide them guidance via this project, by helping them find Neighbourhoods similar to their current head office, in Bangalore.

Data Sources

1. The neighborhood data of the two cities is taken from *Opencity and Wikipedia,* respectively. This for which a are being shared bellow :
    - https://opencity.in/data/bbmp-wards
    - https://en.wikipedia.org/wiki/List_of_neighborhoods_in_San_Francisco

    The data from open city is in a CSV formal while the data from Wikipedia will be required to be Scrapped before being used. ***Beautiful Soup*** library will be used to scrap the data from the Wikipedia page.
    Bangalore has 199 neighbourhoods, and San Francisco has 114 neighbourhoods.
2. ***Geopy*** library to be used for geocodes.
3. ***Folium*** library will be utilized to represent the data on maps.
4. ***Scikit-learn*** will be used to utilize machine learning.
5. ***Foursquare*** *API* will be used to gather neighborhood data.

Importing Data

Importing data is divided into 3 stages.

- The first stage is getting list of neighbourhoods of the four cities from the above Wikipedia links.
- The second stage is getting location of neighbourhoods.
- The third stage is getting the venues in the neighbourhoods from Foursquare.

Neighbourhood lists of Cities

We already have the links of the Wikipedia pages from which we can get the list of neighbourhoods in each city. Beautiful Soup library is used to extract the information from the wiki tables in the pa-es. This data is stored in a pandas dataframe. Along with the neighbourhoods, the city name, the state name and the country name are stored.

Geolocation of the Neighbourhoods

The geopy library is used to get the location data of the Neighbourhoods. Now in geopy library, Nominatim service is used.

For leveraging the free service of Nominatim there is a restriction of 1 call per sec. If the calls are within a second, then a 'timeout' error is displayed. Therefore, there needs to be a one sec gap between each call. A sufficient gap of 2 seconds is provided by calling the sleep function.

Timeout Errors

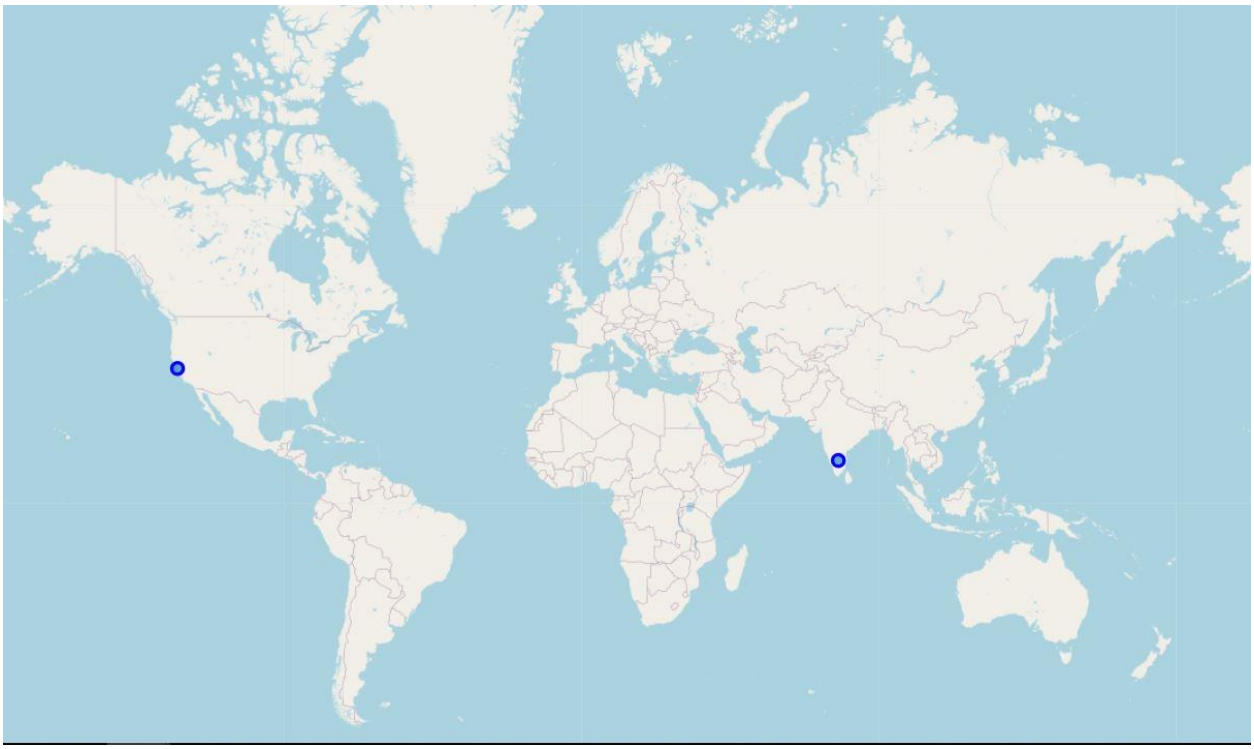
Even after providing a 2 sec gap in calls, there are timeout errors. So, to handle these errors is simple. Simply call the Nominatim service again for these locations after checking for network connectivity.

Missing/No Coordinates

Due to spelling errors in the name of some locations, it will not resolve into coordinates. These can be rectified by using different spellings. Some locations will not resolve despite that. Then that data is procured manually searching on Google Maps.

World Map

A world map is generated for each city shown as markers. San Francisco is in the United States of America, While Bangalore is on the opposite side of the world, in India.



Venues

Getting Venues

Venues are places located in the neighbourhoods like restaurants, hotels, cafes, parks etc. Foursquare API was used get the list of venues for a neighborhood. Since free version of the Foursquare API is used, a maximum of 100 venues can be retrieved.

Now the size of each neighbourhoods might not be equal. Especially between cities. To search venues about position of the coordinates, radius needs to be given. The Table below illustrates the Radius Considered

City	Area (km^2)	No. of Neighbourhoods	Avg Neighbourhood Radius Considered (km)
Bengaluru	709	195	1000
San Francisco	121.4	114	500

Studying Venues

To study the venues, the dataframes containing the venues are grouped by neighbourhoods and summed up.

For Bengaluru, there are some neighbourhoods with a lot of venues while many have very few venues. Realistically this is not true, and this can be considered as Foursquare not having detailed venues for all neighbourhoods. The data provided by Foursquare is crowd sourced and hence, it can be concluded that Foursquare is not popularly used in Bangalore.

For individual clustering, one hot encoding is done for neighbourhoods of each location. While for complete clustering, all venues are combined into the same dataframe and then one hot encoding is done. This makes sure that all the types of venues are considered.

Individual Clustering Results

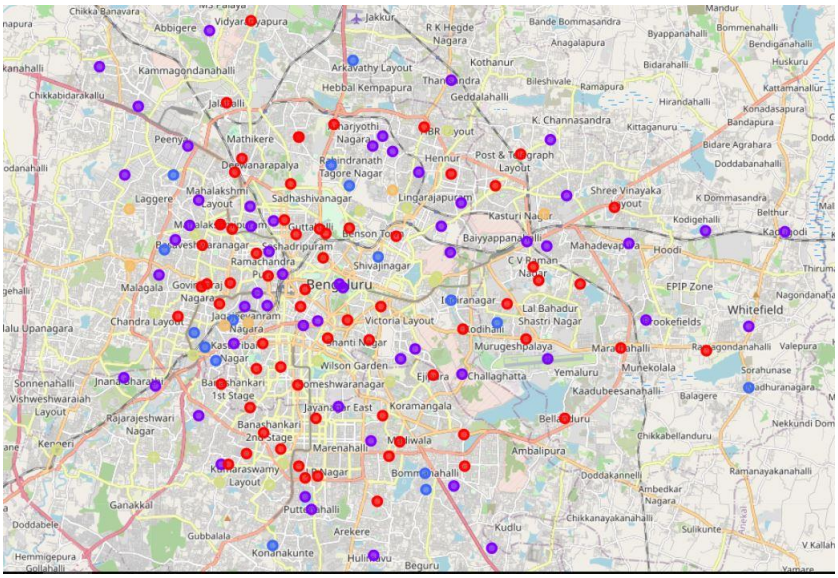
Individual Clustering will help understand how the individual locations can be clustered. To be consistent with all the individual location clustering and the complete clustering, there are going to be 5 clusters.

It must be noted that cluster labels are not the same across different locations. So, Cluster 0 in Bengaluru is not the same as Cluster 0 in San Francisco.

Bengaluru

- Looking at the clustering of the neighbourhoods in Bengaluru, there are 5 clusters with 1 cluster as possible outliers.
- Majority of the neighbourhoods are in clusters 0,2,3 and 4, with cluster 4 having the maximum.
- Whereas, cluster 3 & 4 account for 75% of the total Neighbourhoods.

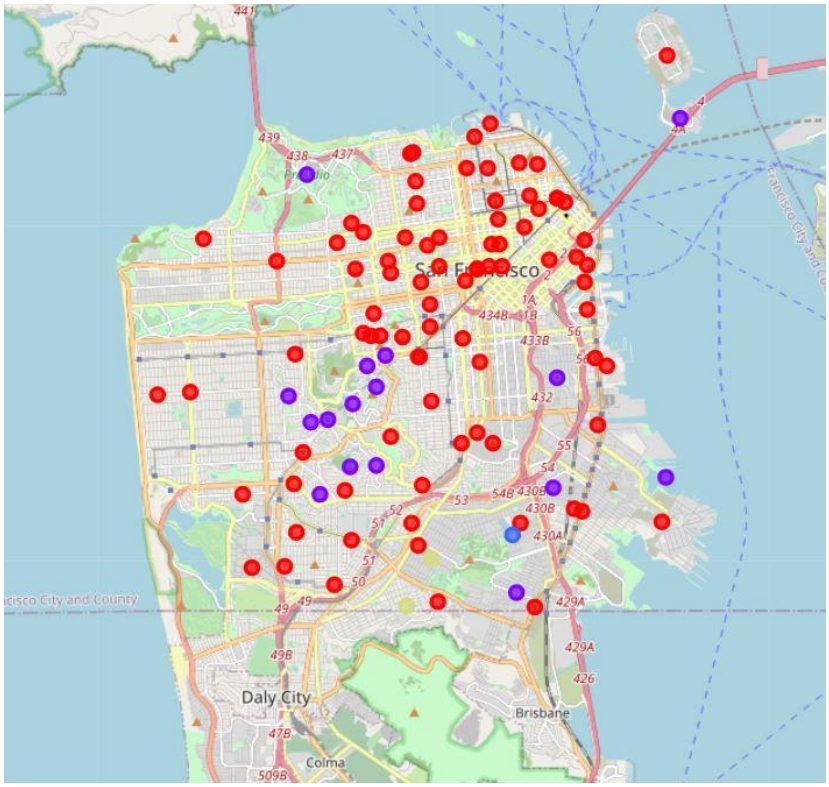
Cluster Labels	Neighbourhood
0	16
1	7
2	17
3	58
4	69



San Francisco

- The neighbourhoods can be divided into 5 clusters and possible 3 possible outliers.
- Major Neighbourhoods are in clusters 0 and 4, with cluster 4 having the maximum.
- Whereas, cluster 0 & 4 account for 93% of the total Neighbourhoods.

Cluster Labels	Neighbourhood
0	18
1	5
2	2
3	1
4	90



Complete Clustering Results

The complete clustering gives some interesting results. Bengaluru and San Francisco have a lot of common Neighbourhoods.

- Cluster 0: Shows 71 (Out of 198) Locations from Bangalore and 114 (Out of 116) from San Francisco.
- Cluster 1: Shows 16 (Out of 198) Locations from Bangalore and 1 (Out of 116) from San Francisco.
- Cluster 2: Shows 7 (Out of 198) Locations from Bangalore.
- Cluster 3: Shows 71 (Out of 198) Locations from Bangalore.
- Cluster 4: Shows 2 (Out of 198) Locations from Bangalore and 1 (Out of 116) from San Francisco.

An important observation is that Most of the Neighbourhoods in San Francisco (98%) show similarity to the 71 Neighbourhoods in Bangalore.

		Country
Cluster Labels	City	
0.0	Bangalore	71
	San Francisco	114
1.0	Bangalore	16
	San Francisco	1
2.0	Bangalore	7
3.0	Bangalore	71
4.0	Bangalore	2
	San Francisco	1

Discussion

The objective of this analysis was that if there is a company, planning to open a Branch office in Bangalore already with its headquarters in San Francisco, help them find a suitable Neighbourhood

Based on Complete Clustering neighbourhoods cluster 0 has Neighbourhood that are similar neighbourhoods. So, if the office is in the neighbourhoods of these clusters in San Francisco then a new franchise can be opened in the neighbourhoods of the same clusters in Bangalore.

Conclusion

The result showed that the branch franchise can be opened in Bengaluru though more data and analysis is needed. More data from neighbourhoods in Bangalore is important.