The Similarity between the neighbourhoods between Mumbai and Chennai

Malavika Rajesh Vikraman

July 21st,2019

1.Introduction

1.1 Background

Mumbai (/mʊmˈbaɪ/, also known as Bombay /bɒmˈbeɪ/, the official name until 1995) is the capital city of the Indian state of Maharashtra. As of 2011 it is the most populous city in India with an estimated city proper population of 12.4 million. The larger Mumbai Metropolitan Region is the second-most-populous metropolitan area in India, with a population of 21.3 million as of 2016.Mumbai lies on the Konkan coast on the west coast of India and has a deep natural harbour. In 2008, Mumbai was named an alpha world city. Chennai (/ˈtʃɛnnaɪ/ (About this soundlisten); also known as Madras /məˈdrɑːs/ (About this soundlisten) or /-ˈdræs/, the official name until 1996) is the capital of the Indian state of Tamil Nadu. Located on the Coromandel Coast off the Bay of Bengal, it is the biggest cultural, economic and educational centre of south India. According to the 2011 Indian census, it is the sixth-most populous city and fourth-most populous urban agglomeration in India. The city together with the adjoining regions constitute the Chennai Metropolitan Area, which is the 36th-largest urban area by population in the world.Chennai is among the most-visited Indian cities by foreign tourists. It was ranked the 43rd-most visited city in the world for the year 2015.Chennai and Mumbai are the most modern cities in India and people tend to migrate to Chennai and Mumbai more.

1.2 Business Problem

People from Mumbai migrating to Chennai and People from Chennai find it difficult to to find a neighbourhood similar to the one they were living in.Similarly , when you shift from one neighbourhood to another neighbourhood in Chennai or Mumbai itself you find it difficult to find a similar neighbourhood with somewhat similar amenities and facilities.

1.3 Interest

This will help people to migrate easily and find themselves in a similar surrounding which will make their shift easy.

1.4 Data

To solve this problem, we will need the following data:

- List of Neighbourhoods in Chennai
- List of Neighbourhoods in Mumbai
- Latitude and Longitudes coordinates of these neighbourhoods. This is required in order to plot the map and also to get venue details
- Venue data, especially regarding all type of venues found in a particular neighbourhood, so that it is easy for clustering

1.5 Sources of Data and Methods to extract them

https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai

This link gives us the major neighbourhoods in Mumbai city.

https://en.wikipedia.org/wiki/List_of_neighbourhoods_of_Chennai

This link gives us the major Neighbourhoods in Chennai city.

We will use web scraping techniques to extract the data from the Wikipedia pages , with help of Python requests and beautifulsoup packages. Store it as CSV separately for Chennai and Mumbai. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder Packages which will give us latitude and longitude coordinates of the neighbourhoods.

After that using Foursquare API, we will collect the venue data surrounding any particular neighbourhood.Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers.Foursquare API will provide many categories of the venue data. This is a project that will make use of many of data science skils,from web scraping(Wikipedia),working with API(Foursquare),data cleaning,data wrangling,to machine learning(K-means clustering) and map visvualization(Folium).In the next section, we will present the Methodology section where we will discuss the steps taken in this project , the data analysis that we did and the machine learning techniques that was used.

1.6 Methodology

Firstly,we need to get the list of neighbourhoods in the city of Chennai and Mumbai.Luckily, for us the details are available in Wikipedia.We will do scraping using Python Requests and using Beautifulsoup packages to extract the list of neighbourhoods data. However, that is just a list of names .We need geographical data in form of latitudes and longitudes in order to use Foursquare data.To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical data in form of latitudes and longitudes .After gathering the data , we will populate the data into a pandas Dataframe and visualise the neighbourhoods in a map using Folium package. This allows Ud to perform a sanity check to make sure the geographical coordinates data returned by Geocoder are correctly plotted in the city of Mumbai and Chennai.

Next, we will use the Foursquare API to get the top 200 venues that are within 1000 meters radius.We need to register a Foursquare Developer's Account in order to obtain the Foursquare ID and Foursquare secret key.We need to make API calls to Foursquare in the geographical coordinates of the neighbourhoods by a python function. Foursquare will return the venue data in json form and we will extract the venue name , venue category,venue latitudes and longitudes.With the data ,we can check how many venues were returned from each neighbourhood and examine how many unique categories can be curated from all the neighbourhood and taking the mean frequency of each venue category occuring .By doing so, we are preparing data that can be used for clustering of the neighbourhoods.

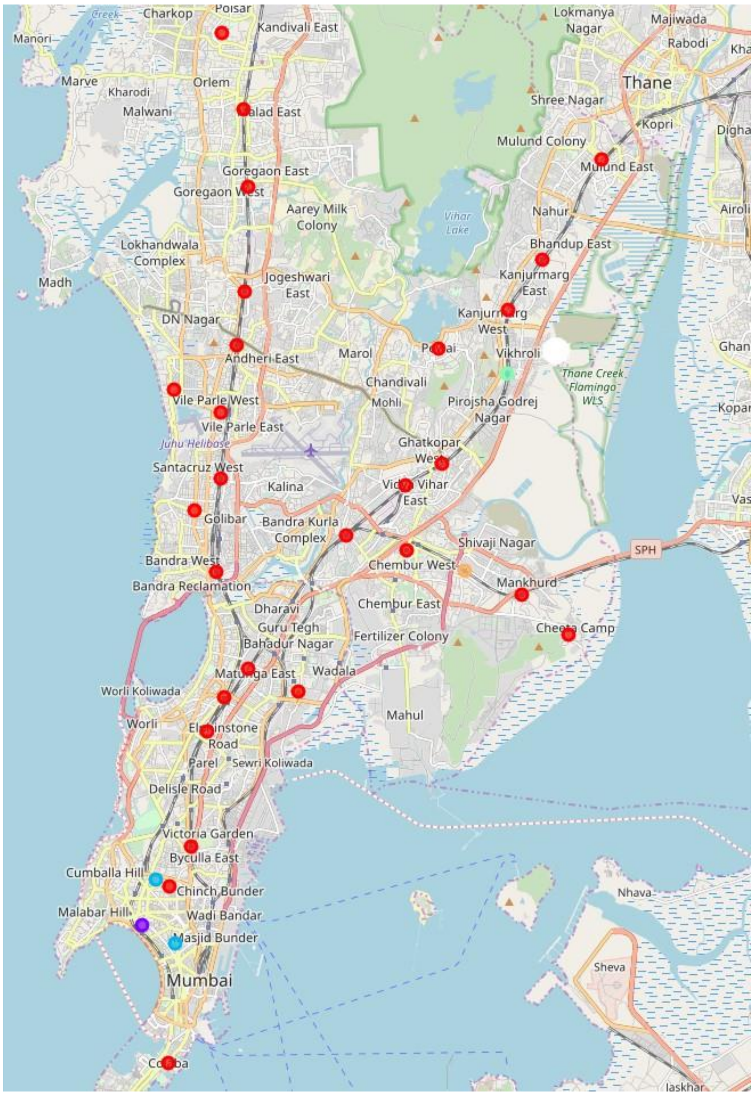
Then individually clustering the data of Mumbai and Chennai gives similar Neighbourhoods within themselves. Combining them will give neighbourhoods among Chennai and Mumbai.

We do clustering by using K-means clustering ,it identifies k number of centroids and then allocates every data point to the nearest cluster , while keeping the centroids as small as possible.it is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project .We will cluster Mumbai and Chennai's neighbourhoods separately into 5 clusters based on on their frequency of occurrence of each venue category each, While the clustering of the neighbourhoods of Mumbai and Chennai into clusters of 7 .

The results will allow to have clear picture of all neighbourhoods are similar to each other with same venues around them making it simple for us to shift to a neighbourhood with almost same facilities.

1.7 Results

- Clustering Of Mumbai

  Cluster labels

  - 0

    Basically consists of neighbourhoods having more Indian Restaurants,Resrtaurants, Ice Cream Parlors,Fast Food Restaurants

  - 1
  - Basically the neighbourhoods having more Juice Bars and Electronics Stores
  - 2

    Basically the neighbourhoods having more Bakery and Department stores.

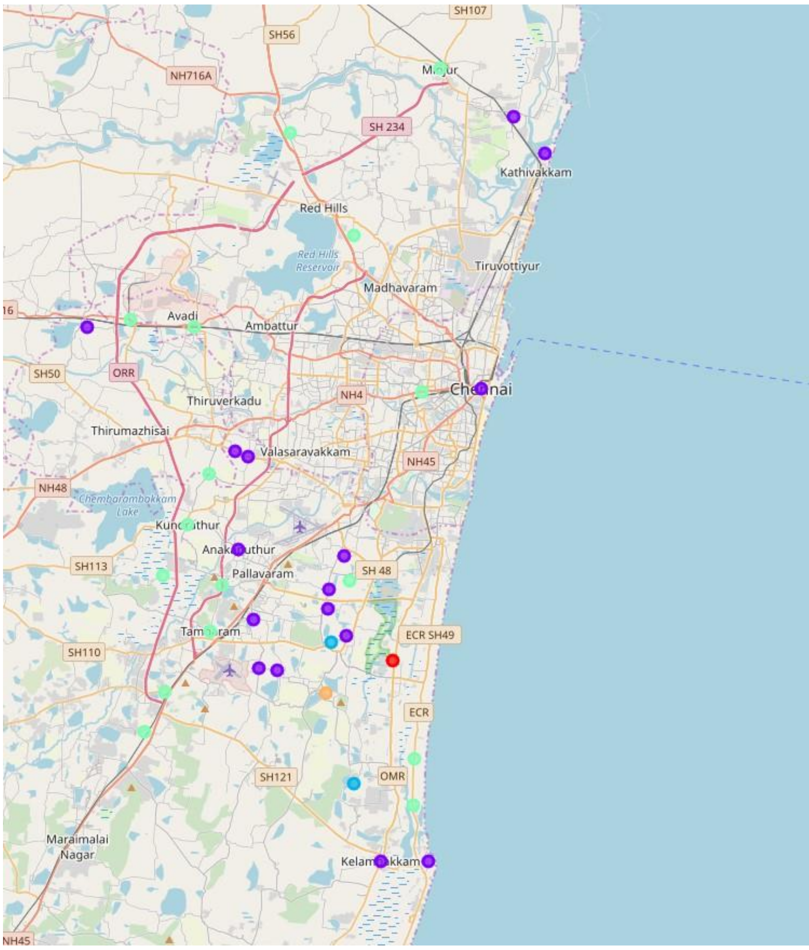
  - 3

    Neighbourhoods having more Café’s , Trials and Coworking Spaces.

  - 4

    Has Hot Dog Joints, Food trucks and pub more in these neighbourhoods

- Clustering of Chennai



- 0

  The neighbourhoods that have more resorts and Golf Courses are grouped in 0 cluster.

- 1

  The neighbourhoods that have more Indian restaurants , Convenience Stores,Cosmetics Stores, Daycares with Bus Station/Train Station .

- 2

  Neighbourhoods having more Tea Stations and Indian Restaurants

- 3

  Neighbourhoods having Cosmetics stores, Convenience Stores , Daycare , Department Stores and Electronis Store more are grouped in 3.

- 4

  Neighbourhoods having Bakery and Men’s Store more are grouped together in this cluster.

- Clustering of Mumbai and Chennai

- 0

  Neighbourhoods that contain Pizza Places, Food and Fast Facilities More

- 1

  Neighbourhoods that contain Restaurants (different kinds) More.

- 2

  Neighbourhoods that conatin Tea Rooms more

- 3

  Neighbourhoods that contain Cafes and Multiplexes More.

- 4

  Neighbourhoods that contain Tea Rooma and Men' Stores More.

- 5

  Neighbourhoods that contain Resort and Golf Course more.

- 6

  Neighbourhoods that contain Convenience stores , Cosmetic shops , Day care ,Department stores and Electronics Stores more.

For detailed understanding See Notebook:

https://github.com/malavikarajeshvikraman/CapstoneDatabase/blob/master/Capstone%20Project.ipynb

1.7 **Discussions**

So since we have clusters formed now its easy for us to identify , where to shift towards.

Like a person who’s living in neighbourhood which in cluster 0 can shift to a similar neightbourhood by going for another neighbouhood of cluster 0.Sometimes we can see the clusters having only one neighbourhood which means that there will be no other neighbourhood that you can shift to with most facilities same.

1.8 **Conclusion**

In this project,we have gone through the process of identifying a problem , specifying the data required , extracting and preparing the data, performing machine learning by clustering the data into 5 to 7 clusters based on similarities and lastly providing information to people regarding how to shift from one neighbourhood to another without feeling much of a change.