Facility Location Planning in India

Abhay Sobhanan

May 21, 2020

Introduction

India is considered as one of the growing economies with high potential for development and success. The huge number of English-speaking youth and their educational skills attract many multi-national companies. Such companies also benefit tax reductions from the Indian government in addition to low infrastructural and running costs compared to the western countries. The given problem is commonly known as facility location planning. It is a computationally hard problem with increasing constraints present in the model. In this report, we will only look at a small data and categorize it using an unsupervised machine learning algorithm. Location selection problems are not only limited to commercial field applications but even a person looking to move to a new country or a place can utilize this problem to find out the right city or neighborhood for his stay.

Interest to stakeholders:

Selection of the right location for your business or new house is a crucial decision which profoundly influences your professional and/or personal growth over the coming years. The search for a perfect location requires many additional details, such as the company profile, location-specific requirements, presence of competitors, etc. However, this analysis can be looked at as a way of preliminary shortlisting of preferable cities for your team. Instead of wasting a considerable amount of time in contacting different dealers for help across the country, shortlisting through data analysis will result in the benefit of low-cost and time.

Now, are you a person looking for a notable difference in the living environment? This study will also help you to sort out major cities across multiple Indian states to find your future home. Personal preferences for facilities highly vary, apart from social and cultural interests. With the advancement in development, your favorite hangout place can vary among a garden, a beach or dine-out and shopping facilities. You can obtain an initial idea from this analysis which groups different cities based on their popular venues.

Data

Data Requirement:

For the described problem, we require a list of Indian cities and their corresponding geo-coordinates for visualization and to acquire other related information. If latitude and longitudes of a city are unavailable in the dataset, one may use *pgeocode* Python package to obtain the details. Using Foursquare API, we extracted the list of most popular venues at these locations with a limit of 20 and radius = 5 km.

Data Acquisition:

We extract the data of Indian cities and their corresponding geographical coordinates. The data is scrapped from the website: "https://simplemaps.com/data/in-cities". The data is readily available for download as CSV file. Else, we may use *BeautifulSoup* python package to scrap the web data. For each selected city for the analysis, we obtain the venues list via "https://api.foursquare.com/v2/venues" using a Client ID and Secret provided by the platform.

Data Cleaning:

Initially, the dimensions of the Pandas dataframe containing location and their geo-coordinates had the dimensions: 212 rows and 9 columns. We rename and select the columns of our requirements such as: Location (City), Latitude, Longitude, State and Population. Later, the rows with NA entries for Population were dropped to filter out the main cities with census data. Please note that the Foursquare venues data for the Indian cities are not very accurate and contains fewer entries. In addition, due to the Covid-19 lockdown rules laid out by the central and state governments, many public places and transport services have stopped functioning. These changes will be reflected in the obtained data. Since these changes cannot be altered at present, we will go ahead with the analysis after the preliminary modifications.

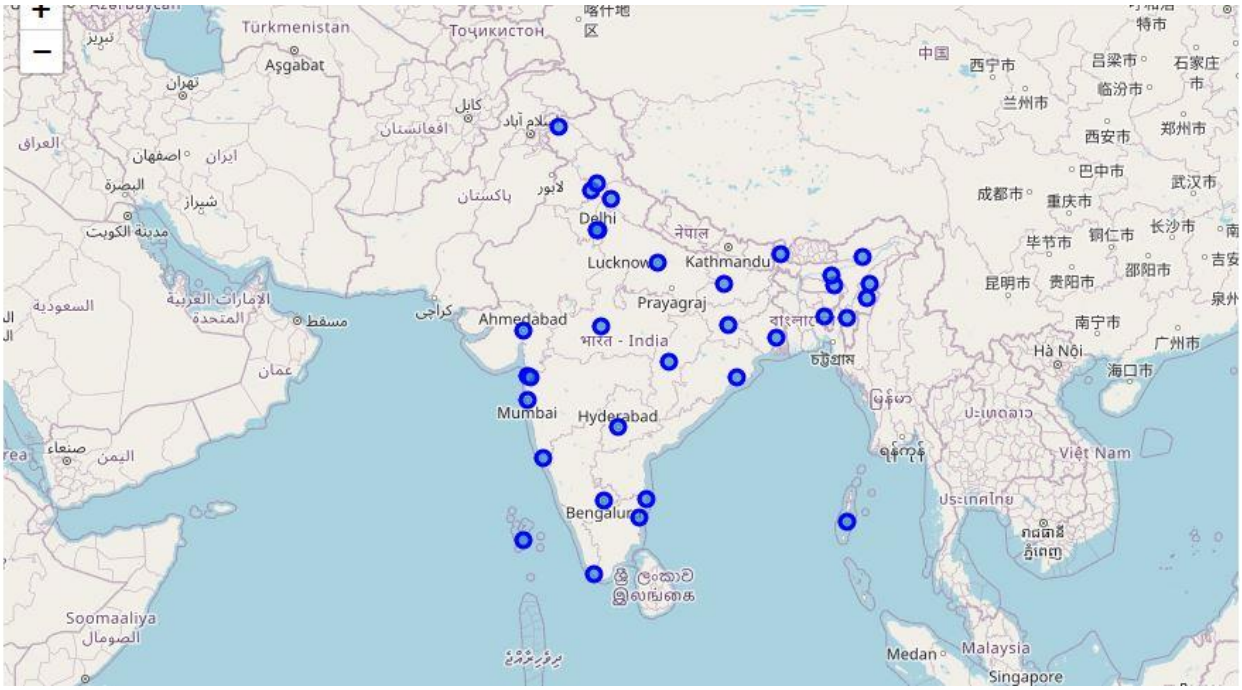
Feature Selection:

The geographical coordinates of the given list of cities will be looped into a function which calls the nearby popular venues within a radius of 5 km. We will make a limit of 20 for each call since the data diversity is less for the selected country. Now, the venues will be grouped together by their locations and can be easily categorized like Restaurants, Sports complex, Shopping malls,

etc. These parameters and their ranking will give the readers an idea about the places of regional attractions.

Methodology

After the preliminary data cleaning, the cities were sorted based on the decreasing order of their population size. The industrial and commercial cities such as Mumbai, Delhi and Kolkata ranked top in the list. Using the geographical coordinates of these cities, locations were plotted in an interactive (zoom in/out) map of India.



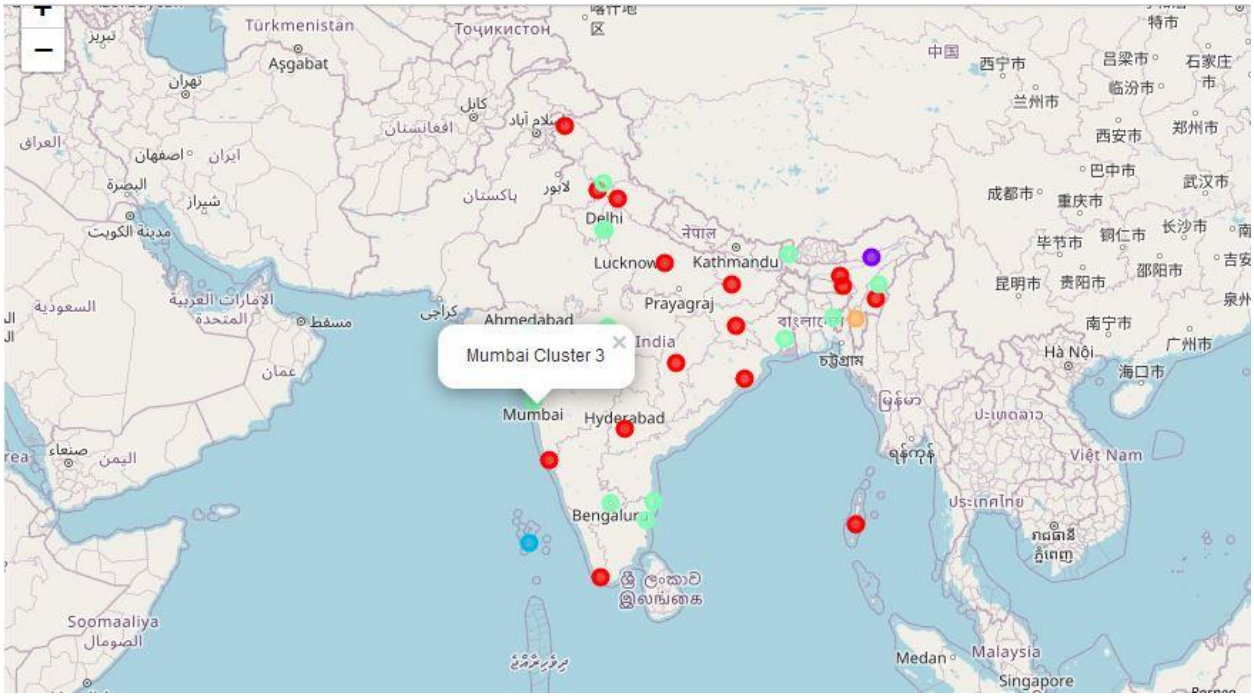
We performed the clustering analysis using the data of these 33 cities. Foursquare API provided some of the popular venues around these locations. They were categorized and merged for each of these locations. Since any prior learning information to categorize the places were unavailable, we have to rely on the unsupervised learning techniques to find the results.

K-means clustering is an unsupervised learning algorithm which partitions the given data points into different clusters, where each of the clusters is of a specific category. We fix a value of k=5 and segment the given data points based on the category ranking of top visited venues in a radius

of 5 km. The *KMeans* function algorithm from the *sklearn* package in Python is called for the execution of our problem.

Results

The program output successfully segmented the list of 28 Indian cities (5 were excluded due to fewer venue data points from Foursquare) into 5 clusters. The following map of different cluster colors visualize the result:



Cities belonging to each Cluster:

Cluster-0	Hyderabad, Lucknow, Patna, Srinagar, Ranchi, Chandigarh, Thiruvananthapuram, Raipur, Bhubaneshwar, Dehra Dun, Shillong, Imphal, Port Blair, Panaji, Dispur
Cluster-1	Itanagar
Cluster-2	Kavaratti
Cluster-3	Mumbai, Delhi, Kolkata, Chennai, Bengaluru, Ahmadabad, Bhopal, New Delhi, Puducherry, Agartala, Shimla, Kohima, Gangtok, Daman, Silvassa
Cluster-4	Aizwal

Discussion

From the clustering of Indian cities, we obtain the following observations and recommendations:

1. Cluster-0 is a group of Tier-1 cities across the country. If a person prefers a steady job (government or private) and hassle-free living environment, these cities should be a good choice to settle.
2. Cluster-1 contains only the city Itanagar which is in the eastern end of India. There is less structural development, and we can quickly notice this from the top-visited places across the city. It is also important to note that this state, Arunachal Pradesh, has the lowest population density across India.
3. Cluster-2 contains one city named Karavatti. On observing the most visited places, it can be inferred that this is a tourist destination of natural beauty. Karavatti is actually the capital of Lakshadweep island. For tourism, especially for international visitors, business planning in the city can be beneficial as a business.
4. Cluster-3 grouped together most of the metro/top-developed cities in the country. Multi-national companies and other Indian start-ups should prefer setting up their main offices in one of these cities for the purpose of networking and business development.
5. Cluster-4 contains Aizawl, which is also a city in a less developed area. However, compared to Cluster-1, this city is more developed, and more facilities are available.

Conclusion

This analysis of multiple cities for their classification looks to be a promising preliminary analysis. The described methods can be used directly or improved according to the situations to get a shortlist of cities of one's interest. Both time and cost will be saved by businesses and people who utilize data analysis methods. While the results may seem intuitionistic for an Indian citizen who has travelled across multiple Indian cities, the same method and model can be applied to find a suitable neighborhood based on several factors of importance, say, housing prices, industrial facilities or even crime rate.

References

1. IBM Data Science Course: https://www.coursera.org/learn/applied-data-science-capstone

2. Foursquare: https://foursquare.com/developers/apps
3. Location data: https://simplemaps.com/data/in-cities

Disclaimer:

The given analysis is part of the user's learning exercise and should be verified before its usage. The outcomes are heavily relying on the input data, and thus any discrepancy in the data can result in considerable variations in the model.

Thank you for taking your time to read my work!