RECOMMENDATION SYSTEM FOR ESTABLISHING NEW BUSINESSES USING GEOSPATIAL CLUSTERING FOR MULTIPLE REFERENCE POINTS

Shreayan Chaudhary *, Paras Sibal *, Dr. M.Ferni Ukrit
shreayan_sameer@srmuniv.edu.in, parassibal_viveksibal@srmuniv.edu.in, ferniukm@srmist.edu.in
Department of Software Engineering, SRM Institute of Science and Technology, Chennai

Abstract

India is an extremely densely populated country (one of the densest), with more than 1.34 billion residents. It is tough to start a business here due to high real estate costs. Moreover, entrepreneurs are scared due to high risk involved and concerns whether or not they would cover the cost with profits to thrive the business. This ML model is personalized for each user by generating its own dataset and determining the latitude and longitude of a place appropriate for setting business in any particular state involving less risk and thereby maximizes profits by attracting more customers. Hence, the objective is to find the optimal location in any given particular city or place to help set up a business for entrepreneurs, thus saving time, money and risk. Target audience consists of entrepreneurs and small-scale businessmen/women interested in setting up the business, aiming at the student and corporate demographic for maximizing profits. This model is determined to be suitable for use by entrepreneurs for setting up the business and future works.

Keywords: Clustering; Geospatial Data; Optimal; Reference Points; FourSquare API; Folium

Introduction

Establishing a new business is one of the most riskiest tasks in today's world. It is very difficult to find the "best and suitable" place for entrepreneurs to establish new business and keep them running for a long term. By using and analysing geospatial data, these techniques and technologies have made it possible to find the most optimal place. The project will be deployed as a web application running on the cloud that includes an interactive web interface as well separately offering ML solutions through a dedicated quasi-RESTful API endpoints. The Data Pipeline will consist of the following steps:
Data Collection -> Data Cleaning and Preprocessing -> Clustering -> Cluster Analysis
The Nominatim library is used to find their latitudes and longitudes. The FourSquare venues API is used to find the related categories to be looked at. We have devised a function to find categories of a given venue from our data. We have used the FourSquare API to find the venues within a particular radius specified by the user (1500m by default). For this paper, we have fetched only restaurant data for setting up a new restaurant but this method can be used for many types of business that are visible over the Google Maps such as Petrol Pumps, Entertainment Places, Hospitals, etc.

Materials and Method

The project will be deployed as a web application running on the cloud that includes an interactive web interface as well separately offering ML solutions through a dedicated quasi-RESTful API endpoints. The application will interface with a web server such as Ngnix suitable to our design and a Python WSGI-compliant HTTP server such as Gunicorn for deployment of Django-based application. The Geopy Nominatim library is used to locate the latitude and longitude for the given address, cities and countries. FourSquare API is further used to find businesses within a specified distance returning venue, postal codes, and summary. Moreover, a google map is generated using folium that further clusters the reference points and thereby minimizing the relative distance, while keeping in mind the customers that the place will attract, while maximizing the profits. We have used Haversine's formula for calculating the distance between two relative points and the cluster accuracy is calculated by using the RMS of elements in clusters and the number of elements in cluster. The cluster score is calculated using cluster accuracy and number of competitors in the cluster.

Result and Discussion

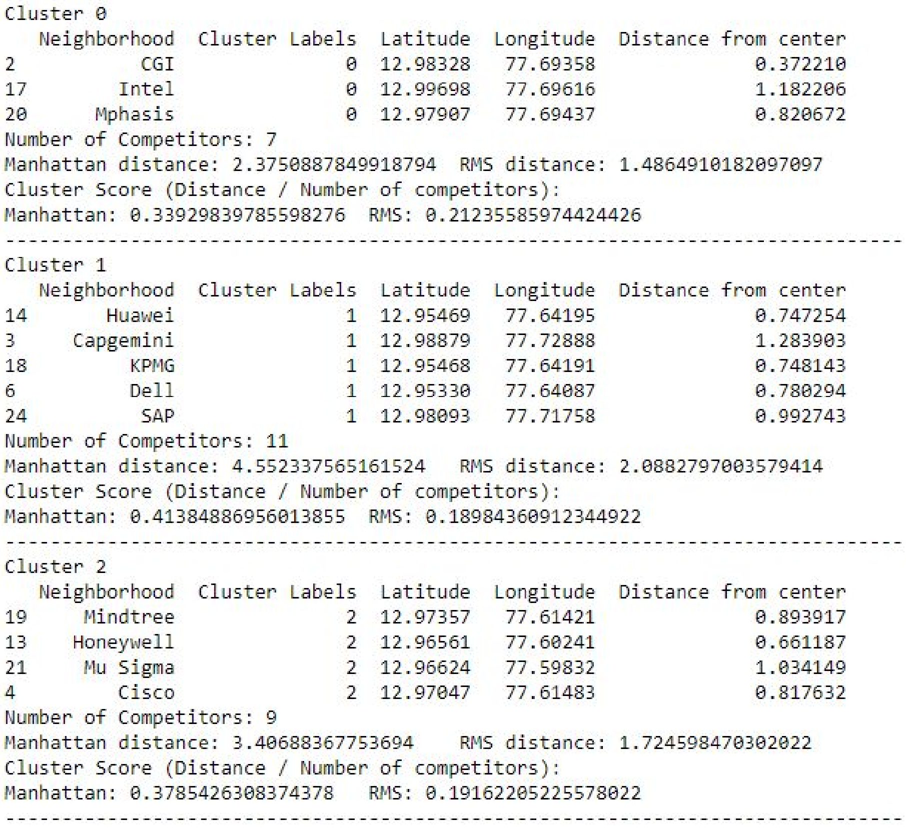
The Model returns the latitude and longitude of the place that is suitable for entrepreneurs to set up new business. This place involves less risk and capable of attracting more customers thereby increasing profits.

$$Haversine's\ Formula : Distance\ d\ between\ 2\ coordinates\ (x_1, y_1)\ and\ (x_2, y_2)$$

$$d\ =\ 2r\ arcsin\ (\sqrt{sin^2(\tfrac{x_2-x_1}{2}) + cos(x_1)\ cos(x_2)\ sin^2(\tfrac{y_2-y_1}{2})}\ )$$

$$Proximity\ Measure\ =\ \frac{RMS\ of\ elements\ in\ cluster}{Number\ of\ elements\ in\ cluster}$$

$$Cluster\ Score\ =\ \frac{Proximity\ Measure}{Number\ of\ Competitors\ in\ that\ cluster}$$

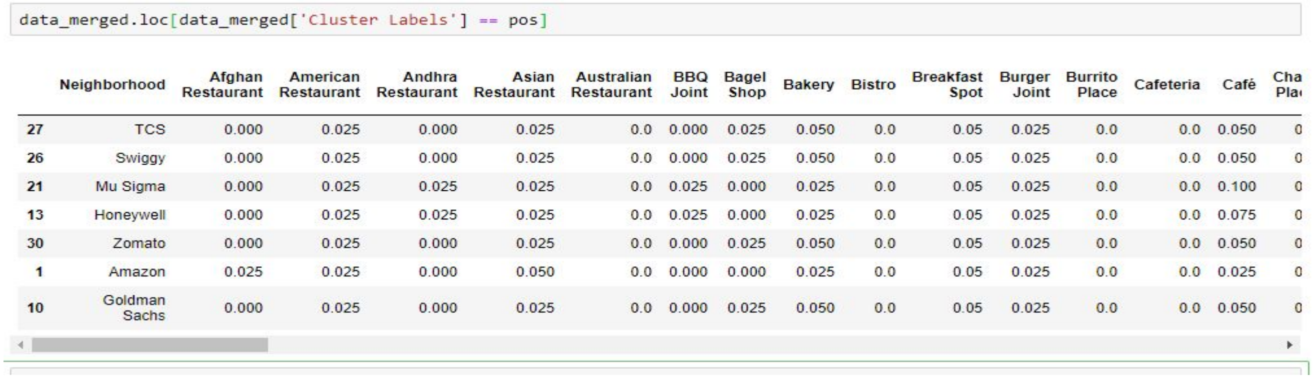


```
Cluster 0
   Neighborhood  Cluster Labels  Latitude  Longitude  Distance from center
2           CGI               0  12.98328   77.69358              0.372210
17        Intel               0  12.99698   77.69616              1.182206
20      Mphasis               0  12.97907   77.69437              0.820672
Number of Competitors: 7
Manhattan distance: 2.3750887849918794  RMS distance: 1.4864910182097097
Cluster Score (Distance / Number of competitors):
Manhattan: 0.33929839785598276  RMS: 0.21235585974424426
-----------------------------------------------------------------------------
Cluster 1
   Neighborhood  Cluster Labels  Latitude  Longitude  Distance from center
14       Huawei               1  12.95469   77.64195              0.747254
3     Capgemini               1  12.98879   77.72888              1.283903
18         KPMG               1  12.95468   77.64191              0.748143
6          Dell               1  12.95330   77.64087              0.780294
24          SAP               1  12.98093   77.71758              0.992743
Number of Competitors: 11
Manhattan distance: 4.552337565161524   RMS distance: 2.0882797003579414
Cluster Score (Distance / Number of competitors):
Manhattan: 0.41384886956013855  RMS: 0.18984360912344922
-----------------------------------------------------------------------------
Cluster 2
   Neighborhood  Cluster Labels  Latitude  Longitude  Distance from center
19     Mindtree               2  12.97357   77.61421              0.893917
13    Honeywell               2  12.96561   77.60241              0.661187
21     Mu Sigma               2  12.96624   77.59832              1.034149
4         Cisco               2  12.97047   77.61483              0.817632
Number of Competitors: 9
Manhattan distance: 3.40688367753694    RMS distance: 1.724598470302022
Cluster Score (Distance / Number of competitors):
Manhattan: 0.3785426308374378   RMS: 0.19162205225578022
-----------------------------------------------------------------------------
```

Fig.3. Sample cluster scores calculation for Bangalore with radius of 2.5km, n_clusters=10.

Conclusion

On sorting the clusters by decreasing order of cluster score, we will get the best places to set up the new business. These places balance the tradeoff of having the most target audience while minimizing the number of competitors for the same. Hence the cluster score will be the most suitable metric to determine the most optimal place for setting up a new business.



```
data_merged.loc[data_merged['Cluster Labels'] == pos]
```

	Neighborhood	Afghan Restaurant	American Restaurant	Andhra Restaurant	Asian Restaurant	Australian Restaurant	BBQ Joint	Bagel Shop	Bakery	Bistro	Breakfast Spot	Burger Joint	Burrito Place	Cafeteria	Café	Cha Pla
27	TCS	0.000	0.025	0.000	0.025	0.0	0.000	0.025	0.050	0.0	0.05	0.025	0.0	0.0	0.050	0
26	Swiggy	0.000	0.025	0.000	0.025	0.0	0.000	0.025	0.050	0.0	0.05	0.025	0.0	0.0	0.050	0
21	Mu Sigma	0.000	0.025	0.025	0.025	0.0	0.025	0.000	0.025	0.0	0.05	0.025	0.0	0.0	0.100	0
13	Honeywell	0.000	0.025	0.025	0.025	0.0	0.025	0.000	0.025	0.0	0.05	0.025	0.0	0.0	0.075	0
30	Zomato	0.000	0.025	0.000	0.025	0.0	0.000	0.025	0.050	0.0	0.05	0.025	0.0	0.0	0.050	0
1	Amazon	0.025	0.025	0.000	0.050	0.0	0.000	0.000	0.025	0.0	0.05	0.025	0.0	0.0	0.025	0
10	Goldman Sachs	0.000	0.025	0.000	0.025	0.0	0.000	0.025	0.050	0.0	0.05	0.025	0.0	0.0	0.050	0

Fig.1. The optimal clusters along with competitor details with radius=1km and 2km respectively

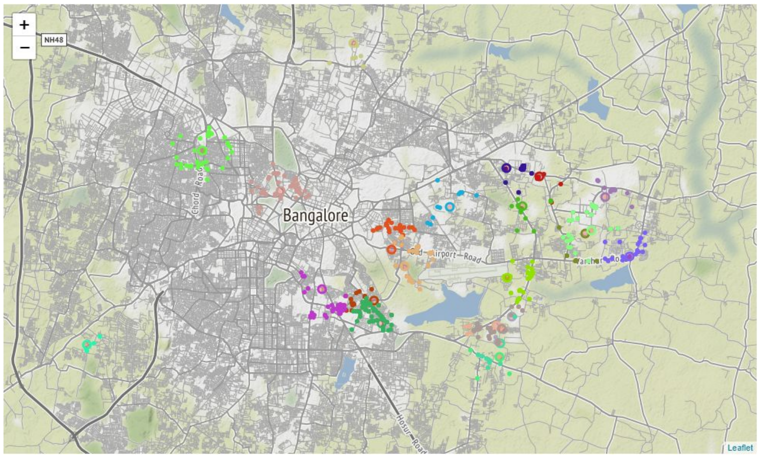


Fig.2. Example of formation of clusters along with reference points.

References

1. Y. Zhong, J. Li and S. Zhu, "Clustering Geospatial Data for Multiple Reference Points," in IEEE Access, vol. 7, pp. 132423-132429, 2019.
2. Rabindra Barik, Ankita Tripathi, Suraj Sharma, Vinay Kumar, Himansu Das. MistGIS: Optimizing Geospatial Data Analysis Using Mist Computing. Progress in Computing, Analytics and Networking Journal, Springer (2018) https://doi.org/10.1016/j.compenvurbsys.2011.11.003