IBM

IT company site determination in HYDERABAD using clustering

By Deepak Kumar Gunturu

April 2020

BACKGROUND:

The city of Hyderabad, in the state of Telangana, is growing to be one of the largest IT locations in India with technology and innovation taking a rise in the city. Lots of multinational software companies such as Google, Microsoft, Verizon, Deloitte, Oracle, Wipro, and Tech Mahindra have set up their offices in the city in technology parks and IT. One such as popular location is Hitech city, a hub for technology where all the companies are located. In the recent years, a lot of new incubators and other locations in Hyderabad have begin to come into the picture bringing about the potential for expansion of existing companies and increase in startup venture opportunities. Thus, identification of the ideal locations for IT/software companies to set up their offices based on the venues is a requirement of paramount importance for any venture.

INTRODUCTION/BUSINESS PLAN:

The main objective of this project is to search for ideal locations for existing IT/software companies or startups by analyzing the location data from the Hyderabad and the different venues and spots surrounding those different locations. Data science methodology will be used to collect the data regarding the locations and this data is going to be subjected to a clustering algorithm to answer the following question: what is the ideal location to set up a company headquarters or branch office in the city of Hyderabad.

TARGET AUDIENCE/POTENTIAL CLIENTELE:

The initial target audience this is targeted for are the owners and CEOs of IT/software companies from any part of the world who want to set up their company's branch offices in a neighborhood in Hyderabad. Startups which require new office spaces to set up their operations would also be able to make use of this application. From 2018, the requirement for the market space has gone up from 8% to 27% for the IT sector alone. It can gradually be later scaled up for organizations wanting to scale up their operations in other cities as well.

DATA:

The following data is will be collected and used for this project:

- Neighborhood data for Hyderabad collected from the Wikipedia page https://en.wikipedia.org/wiki/Category:Neighbourhoods_in_Hyderabad,_India. Scope of the project is determined by this.
- The latitude and longitude positions of the neighborhoods that are presented in Hyderabad. Required for venue data as well as visualizing the map of the possible locations and neighborhoods.
- Venue data and other data related to the IT companies. This will be required to perform the clustering algorithm for the aim of the project.

SOURCES AND DATA EXTRACTION:

The Wikipedia page for the neighborhoods in Hyderabad contains the data for 200 neighborhoods. The names of these neighborhoods are extracted using web scraping with the BeautifulSoup package in Jupyter notebooks. Geoencoder library in Python is used to gain the latitude and longitude values of each of their neighborhoods and their venues.

The Foursquare API is used to get the venue data for the each of the neighborhoods. The details that we get from the Foursquare API are categorical values that we can use for clustering together the neighborhood data and use for solving the business problem which is determining ideal locations for setting up the headquarters of a location for an IT/software company or startup. In the next sections, the data science methodology will be used to explain the low level processes for collecting (BeautifulSoup and Foursquare) and curating (Pandas and NumPy) the data and then applying the k-means clustering(SciKit-Learn) algorithm on it to determine the ideal locations segmented into different clusters of neighborhoods.