

2. A description of the data and how it will be used to solve the problem. **(15 marks)**

The required data for the project are taken from foursquare.com. The analytics need to explore two datasets. The first one is the IT industries location data in the city and the status of the company which includes categories like

1. Completed
2. Started
3. Planning
4. Under construction

The data is available in https://en.wikipedia.org/wiki/List_of_tech_parks_in_Chennai.

The another data set is the geographical location data of the city to plot the clusters in the city map. The dataset is available in https://en.wikipedia.org/wiki/List_of_neighbourhoods_of_Chennai.

As the project focuses on the location of the IT industry and not on the IT companies working in the tech park those details are cleaned during the data cleaning and preparation step. In the same way the location also takes the postal address and latitude and longitude to get the geographical location.

Later the Status are converted to numeric data as status code to cluster it based on the status of the techno park. And two clusters are formed based on location and status.