8/26/2019

Riya Manish Bhandekar

Milvik, Enhancio

Venue location selection

An anomaly detection approach

# **Introduction**

In this capstone project we will explore the idea of where to build new sports facilities.

## Background

To set a background for the discussion, the local government wants to build more sports facilities in the selected neighbourhood. The sports facilities are of 2 main categories – 1. Athletics and Sports 2. Gymnasium / Fitness Centre. They want to ensure that they achieve a balance between the types of facilities they build. So they want to ensure they increase the number of gyms in areas where there are fewer gyms etc.

This is our problem domain.

## Problem

The problem we will solve is deciding where we should build sports facility in a neighbourhood and of what type it should be. This approach is similar to anomaly detection where we find outliers i.e. areas having less sports facilities.

## Stakeholders

The stakeholders of this project would be local sports departments OR people interested in opening some type of sports facility business in the neighbourhood for example a chain of Gyms like Life Time Athletic. They want to decide what type of facility and/or also where it is best to build it. In order to benefit the new business they will want to select areas with no facility or over-crowded facilities to profit from the existing crowd.

1. Data Sourcing, Understanding and Wrangling

The primary data source for this project is the location data provided by the provider FourSquare. We will access the FourSquare REST APIs to get venue and other details. Following sections look at the APIs we will use and understand the data they return.

## Data API

The important FourSquare API for this project are listed below :

* Search For Venues: GET: XXX/v2/venues/search
* Get Venue Categories: GET: XXX/v2/venues/categories
* Get Details of a Venue: GET: XXX/v2/venues/VENUE\_ID

## Problem & Data Understanding

The problem requires that we apply the model that we come up with on a certain area or neighbourhood and we should be able to run it again on other neighbourhoods. So it should except new parameters in terms of area/neighbourhood and also venue categories we want to differentiate between.

We will work with Venue Dataretrieved from FourSquare. Each Venue belongs to a list of categories wherein there is a parent category. We will be interested in venues of following parent categories for our fitness project :

* **Athletics & Sports (4f4528bc4b90abdf24c9de85) (exclusively)**
* **Gym / Fitness Centre (4bf58dd8d48988d175941735)**

We will assume the user has a location in mind around which they want to check availability of fitness facilities. Hence we will assume a target geographical location given by user as input data also.

## Methodology

On analysis the problem appears to be of a location based, spatial nature. The data sets we can gather are un-labelled and un-supervised. We are interested in finding how venues of particular type (Gym or Non-gym) are scattered around a given location - where the collections are denser and where they are sparse. Thus we think that **CLUSTERING** machine learning technique is the best way to develop our model for analysis and prediction. We say this because the data points (venues) which belong to a segment or cluster are similar both spatially and by type. Thus putting them in the same cluster. We can say that our final conclusion will be made around the **outlier detection ability of our model** telling us the sparse points in our data set where are clusters are formed based on following 4 features –

1. Geographical location (latitude, longitude)

2. Venue Category

3. Likes received for the venue

## Exploratory Data Analysis

We will develop the model using a neighbourhood in Toronto, Canada – Ryerson,Garden District.

Data collection and preparation

The data set we have gathered has 50 samples and 3 important features – latitude, longitude and Category. The other features we will use for final conclusion are – venue name, Distance from provided location, user likes. We will convert the category names using the pandas get\_dummies method to float values and then use standard scaler to fit transform the values to standard forms before building the model.

A screenshot of a cell phone

Description automatically generated

Data visualization

We plotted the venue locations on a BaseMap to study their distribution in the area.

A picture containing text

Description automatically generated

Looking at the data points plotted above it appears that **DBSCAN** is the best clustering algorithm for developing the model. The reason is that it has the ability to create clusters which can be **nested (cluster within a cluster)** and also show the **correct outliers** considering a given range. As it is not required to specify the expected number of clusters hence we will get the right idea of the outliers. The algorithm will not “try” to include all the points into some cluster thus it will not skew the cluster formation showing us correct locations with lesser sports facilities nearby.

1. Predictive model

# Out original data was collected from a radius of 50 Km around the location of interest. We gathered the maximum allowable venues which was 50. We built the DBSCAN model based on location and category data with epsilon as 2 (km) and number of data points as 5 (sports facilities). This created in total 4 clusters (including outlier set) and **14 outliers for our model.**

# A screenshot of a cell phone Description automatically generated

# **We can see the outlier distribution in the map below** which is of our interest as we are using the anomaly detection method. The outliers (grey coloured data points) show us the sports facilities which do not belong to any cluster based on location and category. Thus, we can build new facilities near them using the user likes data to decide.

# A close up of a device Description automatically generated

# We filtered the data frame to include only the outlier data points as shown below and came to the solution of our problem using our predictive model.

# A screenshot of a cell phone Description automatically generated

A screenshot of text

Description automatically generated

Thus few options we can suggest for possible locations and types of sports facility is as below based on the table and map above,

* **Life Time Athletic Gym** near Eglinton Park (43.707430 -79.405359) or Markland Wood Golf Club (43.629057 -79.581131)
* **Baseball or Tennis facility** near YMCA Mississauga (43.586857 -79.644184) or The Country Club (43.793174 -79.589762)

1. Discussion

The model that we have built is completely parametrized for sports facilities. Thus we can easily provide a different neighbourhood latitude and longitude and use the same model to suggest locations for new sports business and its type. Also with a few minor changes we can even change the model to predict other types of venues like cinema halls and hotels which can be modelled for outlier detection. Thus we can answer questions like where can be open a hotel or new cinema house to get best business results.

We are focussing on a small area thus we are able to focus on small distances which will help us to provide more services to same users where existing services are “less liked” or over-crowded. We can also check if our new facility is at a permissible distance from our neighbourhood of interest.

1. Conclusion

# To conclude, we have built a model which looks at venues of a particular category from a “helicopter view” giving us an idea in a 10 by 8 cm area how they are geographically distributed. This helps us to cluster them using additional information like their category. Using a spatial model helped us to use the fact that we do not want to open a business where there are already a lot of other similar businesses leading to a loss for us. The user ‘likes’ data helped is to decide which areas has users visiting other sports facilities in large numbers and what types are they. Thus we feel we have used location coordinates and venue types in a very useful way for our predictive modelling.