# Use machine learning to predict business start locations intelligently

22 Dec, 2018

# Introduction:

Small businesses are grouping a critical section in a developed or developing economy. Traditionally, owners or entrepreneurs of small businesses start their business according to experience or very limited market investigation due to investment limitation or personal knowledge shortage. Big companies normally conduct this by its special department where there are professional analyst or data scientist, or out sourcing this to professional vendors. To facilitate small business owners make an informal decision, especially for the business location, a machine learning powered solution introduced in this report.

# **Method:**

In this report, a creative method is deployed. Firstly, generate a vector of category and then attend the DBSCAN clustering, after that, the prediction - a range of neighbourhood areas are suggested due to the clustering result.

## Data Collection

So the data in the report combines the Foursquare location-based business data and Toronto city borough and neighbourhood information.

### Four Square Location data

There are many famous data providers who can provide business data based on location like Google, Facebook and Foursquare. Foursquare location data is deployed through its comprehensive and convenient web-based restful API because it is free and growing every day. The data is formated as json , an example is like:

{'meta': {'code': 200, 'requestId': '5c1c447bdb04f53ab6735573'},

'response': {'suggestedFilters': {'header': 'Tap to show:',

'filters': [{'name': 'Open now', 'key': 'openNow'}]},

'headerLocation': 'The Beaches',

'headerFullLocation': 'The Beaches, Toronto',

'headerLocationGranularity': 'neighborhood',

'totalResults': 5,

'suggestedBounds': {'ne': {'lat': 43.680857404499996,

'lng': -79.28682091449052},

'sw': {'lat': 43.67185739549999, 'lng': -79.29924148550948}},

'groups': [{'type': 'Recommended P

…

The report mainly focus on fields 'venue.name', 'venue.categories', 'venue.location.lat', 'venue.location.lng'.

### Toronto Borough and Neighbourhood data

All the analysis is based on the Toronto, CA, so public borough and neighbourhood data is also extracted from public wikipedia page [link](\"https:/en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M\"). The data has the format:

| Postcode | Borough | Neighbourhood |
| --- | --- | --- |
| M1A | Not assigned | Not assigned |
| M2A | Not assigned | Not assigned |
| M3A | [North York](https://en.wikipedia.org/wiki/North_York" \o "North York) | [Parkwoods](https://en.wikipedia.org/wiki/Parkwoods" \o "Parkwoods) |

So cleansing and pre-processing need to be conducted before using it.

## Data pre-processing

### Foursquare location data

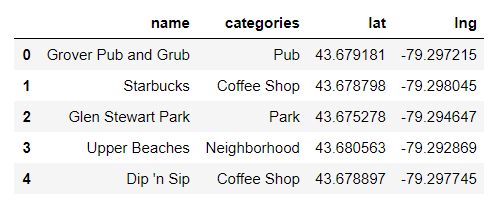
Through Foursquare *Explore* interface, an amount of recommendation venues are retrieved with json format. The data has much more items than the report needed. We focus on the following attributes/features and need get them into DataFrame:

Venue : The venue name for data exploration analysis as index

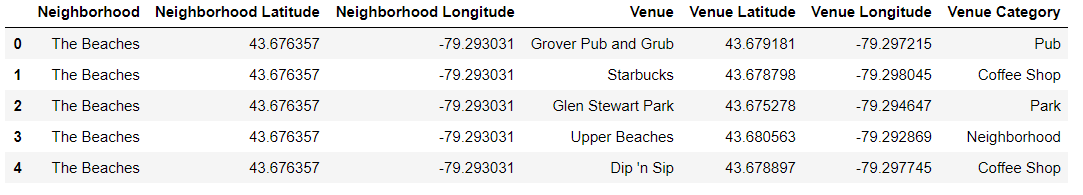
Venue Latitude :The location of venue for data exploration analysis

Venue Longitude : The location of venue for data exploration analysis

Venue Category : The main attribute to define the features of a neighbourhood



After combining with Borough and Neighbourhood :



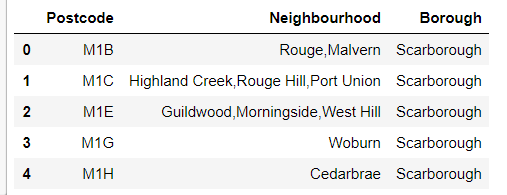
### Borough and Neighborhood

**Data Formatting:**

As mentioned in [Section 2.1.2](#_Toronto Borough and Neighbourhood data), borough and neighborhood data is from wikipedia, it have to be extracted and change to table data (here, Pandas DataFrame)

**Missing Values wrangling:**

Both borough and neighborhood data contains ‘Not Assigned’, if borough is ‘Not Assigned’, delete it as it is the main critical feature required. If neighborhood is ‘Not Assigned’, replace the name with borough name to keep borough as possible as we could. After cleansing, these data will be like:

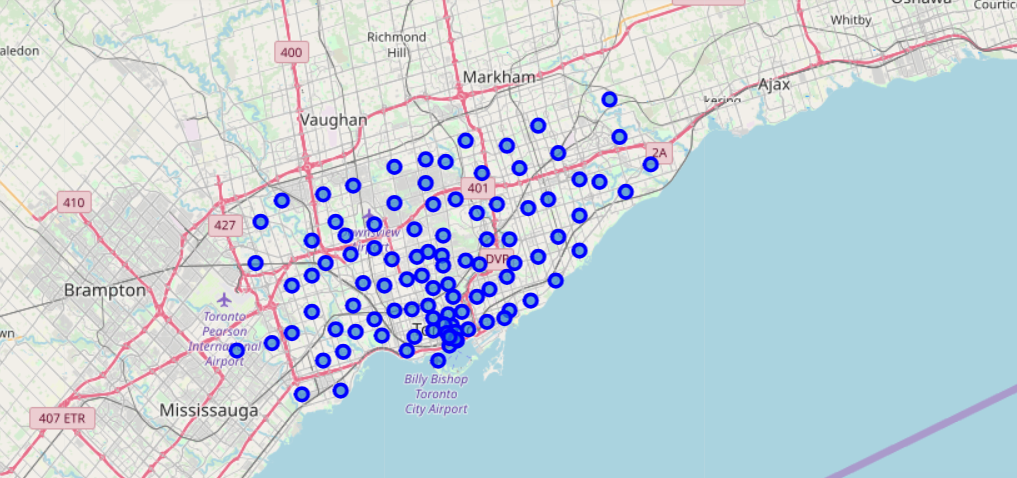


## Exploratory Analysis

This section show the way how data is visualization and understood and then generate the characteristics of the neighborhood as the basis to allow business owner match their business to specific neighbourhood areas.

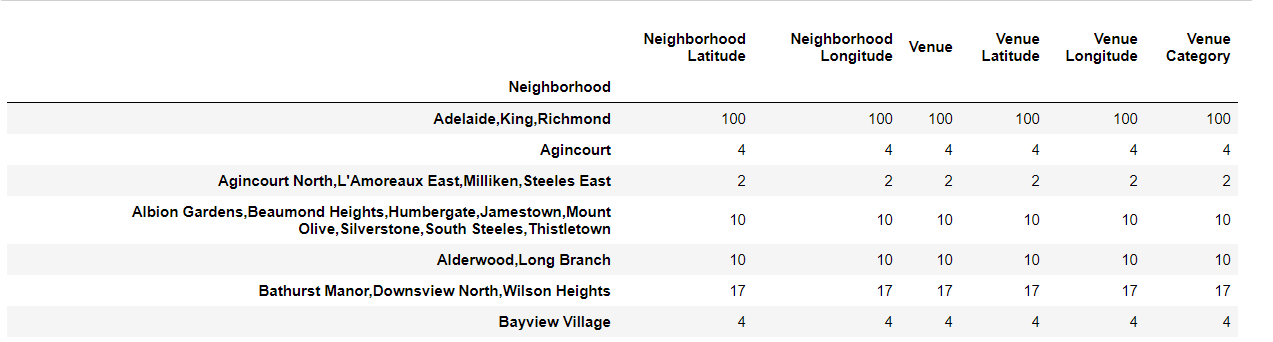
### Describe an overview of neighbourhood data.

The neighbourhood data is labeled on the map. The business requirement is to find a location for their business in specific neighbourhood areas.



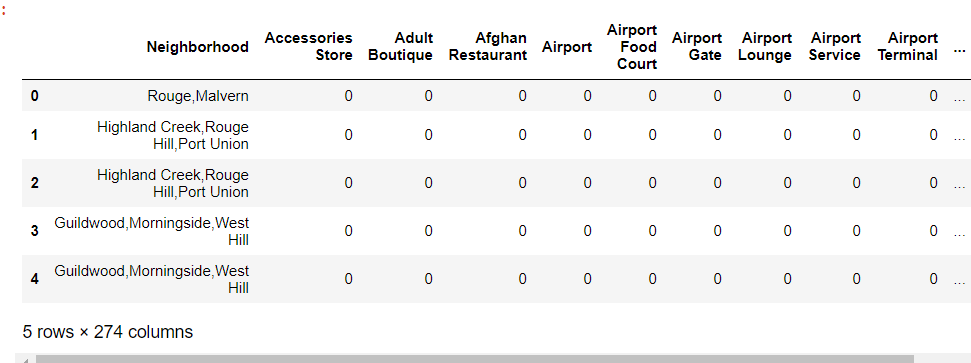
### Grouping Analysis

After venues data - Foursquare location data is acquired, grouping technique is deployed to see the characteristic of each neighbourhood. This is because the goal is find a correlation on neighbourhood instead of venue itself. From the following table, “Venue Category” is a key attribute to our case, more information should be extracted from it besides the number.

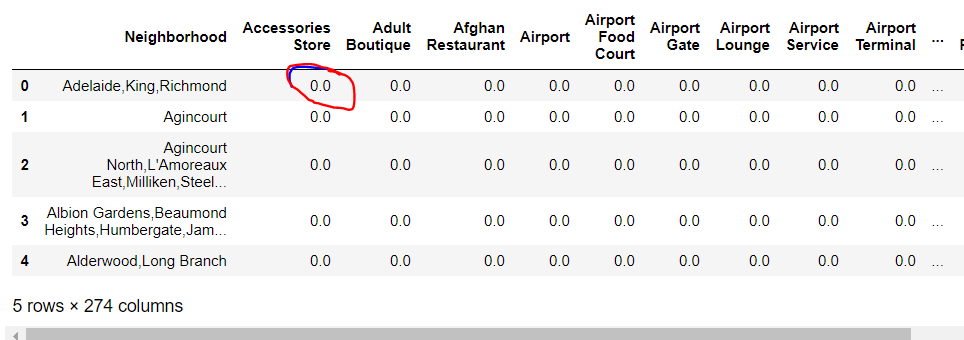


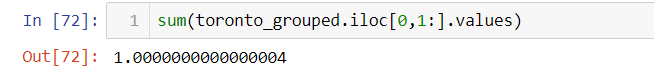
### Dummy Features creation.

To retrieve more information from “Venue Category”, we create dummy features of each category, expand one feature into 274 features/attributes.



Same as before, we focus on neighbourhood , group and aggregation need deployed. Basic mean is deployed to get a proportion of each category in the all categories. So from each number in the table, it give the category significance for a specific category in a neighbourhood. The sum of numbers of different categories in a neighbourhood should be one.





Through this category score vector, the speciality of each neighbourhood is define, it is ready to create a model to solve our problem.

## Modeling - Machine Learning DBSCAN

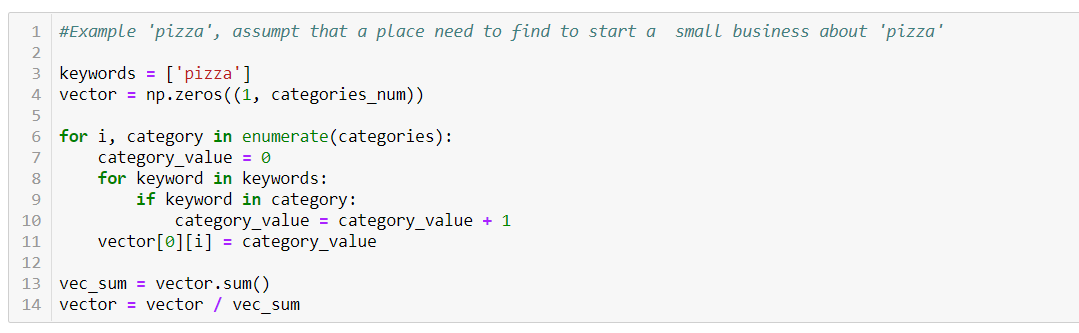
To get the prediction result, there two main steps.

### Business Vector generation

Generate a fake vector from key words described by owner of small business.

Actually, category name for venues are the combination of typical key words like ‘coffe’, ‘cake’. So we match the key words from business owner with the category name and spread them into the categories vector in a uniform distribution way. (This could be improved with more information and give each words specific weights)

The ‘pizza’ example:

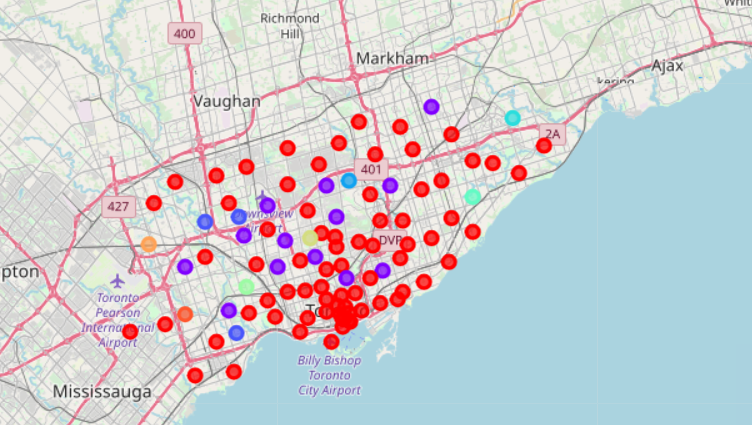


### Give suggestion due to clustering result.

Add the vector in the neighbourhood and clustering to see which cluster it is assigned.

The report choose DBSCAN(Density based spatial clustering of Application with Noise) own to its main advantage: 1) DBSCAN can identity outliners or more noise tolerance. As the fake vector is generated from key words, its distribution is not assured , in other word, maybe too noisy. 2) In some cases, the business is not proper to the whole Toronto - all neighbourhood area could not match , that is, it is a kind of outliner. If using other clustering method like K-Means, it can not deal with this situation.

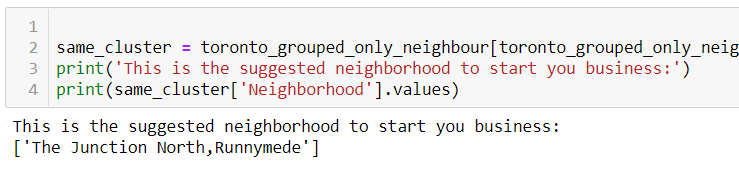
The following is the clustering result.



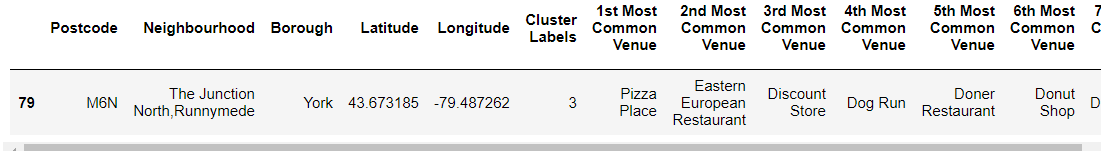
## Validation and Statistical Testing

We choose ‘pizza’ as a test and validation example.

The model give the suggestion is 'The Junction North,Runnymede'.



Then we check this neighbourhood features:



And find “Pizza Place” is the first most common venue.

## Results

After testing 20 key words, the following result is generated.

21% prediction is specific

55% prediction is too general - because they are in a big cluster.

11% prediction is outliner

13% prediction is non-sense like general words “shop” “restaurant”

## Discussion

### Generality of Categories name

From [Section 2.4](#_Modeling - Machine Learning DBSCAN) , the key step is to generate the fake vector due to keywords from matching with Categories of venues.

In this stage, the report only include categories name from Foursquare response, but it contains the other items ‘pluralName’ and ‘shortName’ which could be used in the next stage to cover more keywords. Thus the modeling will be more general to cover the variety of key words.

'name': 'Pub',

'pluralName': 'Pubs',

'shortName': 'Pub',

### Key words generation vector issue

The core in the method is to create the fake category vector from key words. The result is a raw suggestion in a big range sometimes due to the clustering result. So in the next stage, some more intelligent way could be deployed to generate more accurate result, like analysis and generating the vector from discussion (audio or text) with Natural Language Processing techniques.

### Clustering or other method

From previous sections, our clustering way only generate a raw result , no priority , no probability. This could be improved after the clustering , then calculate the distance with other example in a same cluster.

# **Conclusion**

The way of using Machine Learning to predict the suggested business location for small business is a good reference due to the acceptable result.

# **Appendix:**

1. **Foursquare - www.foursqure.com**