**Coursera Capstone**

**IBM Applied Data Science Capstone**

**Finding a new home in Toronto, Canada**

By: Brian Chan

January 2020

A small boat in a body of water with a city in the background

Description automatically generated

**Introduction**

I am a person living in Hong Kong. I am living in the Kennedy Town region, where it is very close to the underground station. The region has a balance of the eastern and western culture. It is easily accessible to the central downtown, but far enough to enjoy a quiet and relaxed environment. Recently, I have been invited by my boss to work in Toronto. The package is a nice deal and I decided to accept it. I am very excited, and at the same time very busy at the preparation work. I am looking for an apartment in Toronto which has a similar ambience compared to my current living environment. The question is, which Neighborhood should I look for?

**Business Problem**

The main problem is to find a neighbourhood in Toronto that exhibits the closest characteristics compared to my current home: Kennedy town. To find out the answer, we can break it down to two questions:

1. Getting the characteristics of Kennedy Town
2. Matching the characteristics of Kennedy Town to a neighborhood (or a few neighborhoods) in Toronto for consideration.

**Data**

We will use the Toronto Data we have prepared in week 3, and get the characteristics of Kennedy Town from FourSquare.

Toronto Data

A screenshot of a cell phone

Description automatically generated

Kennedy Town Data

Let’s take a look at the Kennedy Town.

A close up of a map

Description automatically generated

**Methodology**

For the Toronto data, we majorly relied on Foursquare API to retrieve all venues of each neighborhoods. After that, we calculate the number of different types of venues in each neighborhoods. From the frequency of each type of venues appearing in the neighbourhood, we can infer the characteristics of each neighborhoods.

After that, we will use the K-Means clustering algorithm to infer 10 clusters in Toronto neighborhoods. This will give us an idea of 10 different neighbourhood characteristics in Toronto.

For the Kennedy Town data, we will process the Kennedy Town data in the same way, and infer the frequencies of each type of venues in Kennedy Town. We will use the K-Means clustering algorithm to predict which cluster does Kennedy Town belongs to. The neighborhoods in that cluster will exhibit a similar characteristic compared to Kennedy Town.

**Analysis**

**Step 1: Get the Toronto Data**

We first used the foursquare API to get all venues around 500 meters of each Toronto Neighborhoods. Below shows a snapshot of the data:

A screenshot of a cell phone

Description automatically generated

We take a look at the Neighborhoods that return most amount of venues

A screenshot of a cell phone

Description automatically generated

Let's find out how many unique categories can be curated from all the returned venues. There are **271** unique categories in total.

We then one hot encoded the data into a dataframe like following format:

A screenshot of a cell phone

Description automatically generated

The one hot encoding technique is a technique commonly used in machine learning to preprocess categorical data.

However, the one hot encoded dataframe is quite hard for human to interpret. To make it easier to draw insights, we retrieve the top 10 most common venues for each neighbourhood.

A screenshot of a cell phone

Description automatically generated

**Step 2: Get the Kennedy Town Data**

We will use the foursquare API to get the venues around Kennedy Town. More specifically, we will get the top 100 venues within a radius of 500m around Kennedy Town.

The below table shows the first 5 records returned by the FourSquare API.

A screenshot of a cell phone

Description automatically generated

In total, 78 venues are returned by foursquare. Let’s take a look at the most common venues too.

A screenshot of a cell phone

Description automatically generated

**It seems that Kennedy Town is filled with a lot of restaurants!**

Step 3 – Group Toronto Neighborhoods into 10 clusters.

We apply the k-means clustering algorithm to cluster the Toronto neighborhoods into 10 clusters. The map below shows the clustering results.

A picture containing text, map

Description automatically generated

We see that a lot of the neighborhoods are in the yellow cluster (cluster 7).

The following table shows a snapshot of the top neighborhoods in cluster 7.

A screenshot of a cell phone

Description automatically generated

Upon closer examination of these cluster, we can see a lot of restaurants. Indeed, the environment resembles what Kennedy Town is!

**Results**

We found that Kennedy Town belongs to Cluster 7 of the Toronto Neighborhoods. Some of the Cluster 7 Neighborhoods are:

* Highland Creek
* Rouge Hill
* Port Union
* Scarborough Village
* Cliffcrest, Cliffside,
* Scarborough Village West
* Birch Cliff
* Cliffside West

**Discussion**

Taking a look at the most common places in Cluster 7 neighborhoods, we see that most of them really contains a lot of restaurants: e.g. Bakery, Hakka Restaurant, Pizza Place, Caribbean Restaurant, Thai Restaurant etc.

Therefore, from this perspective, the algorithm really works as expected.

**Conclusion**

Most of the restaurants are concentrated in the Cluster 3 Neighborhoods in Toronto. I will start searching my new home in those Cluster 7 neighborhoods.