**Toronto Neighborhoods Data Clustering**

Dixon Harley

April 10, 2021

**Introduction**

Toronto is the capital city of Ontario, Canada with 103 neighborhoods and the most populous city in the country with a population of 2.7 million. I decided to do this project based of this city owing to the good data availability of this city.

An American company is looking to open a branch of their baseball accessories/souvenirs shop in Toronto, Canada. They are looking for the most ideal location of their new shop to maximize visitors and profits. To do this, they decided to hire a team of data scientists to leverage the power of data to gain insights about where them should open their branch of baseball shop in the city

**Data**

There are three sources of data that we are going to be using for this project. The first data is a list of neighborhoods, boroughs, and postal code of Toronto. We obtained the dataset by scraping this Wikipedia page <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M> using BeautifulSoup.

Text

Description automatically generated

Figure 1. Preview of the dataset.

The next data we will be using is geospatial data which contains coordinates for each neighborhood in Toronto. The data is available in this link <https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs_v1/Geospatial_Coordinates.csv>

Table

Description automatically generated

Figure 2. Coordinates data

Both of these dataframes are combined to create a single data frame, in order to query information from our third data source, Foursquare API. Foursquare Developer API provides the means to query a list of trending venues within the radius of specific coordinates. In the notebook, a function to obtain JSON values from the API into a dataframe is created.

**Methodology**

1. **Data Preparation (pre-processing)**

Before modelling, the data is pre-processed and prepared to suit our goals. First, we combine the scraped neighborhood data with the coordinates data which contains both latitude and longitude values for different postal codes. Table

Description automatically generated

Figure 3. Combined Data Frame

With each neighborhoods coordinate values ready, we can begin to query the top trending venues in each neighborhood in Toronto from Foursquare Developer API. The resulting dataframe is then converted to binary values, enabling for easier aggregation and analysis.

Table

Description automatically generated

Figure 4. Data frame from Foursquare API converted into binary values

The data frame is then aggregated by the neighborhood columns with mean values, obtaining the mean value of different types of venues occurrences in each neighborhoods. **This data will later be used for clustering.**

Graphical user interface, text, email

Description automatically generated

Figure 5. Aggregated toronto\_onehot

1. **Exploratory Data Analysis**

We can further analyze the data by ranking the most common venues in each neighborhoods. For this specific purpose, a function to sort the columns is created.

Graphical user interface

Description automatically generated

Figure 6. Ranking the most common venues in all neighborhoods.

1. **Clustering**

Utilizing sklearn’s KMeans library, we can proceed clustering the data that has been obtained and processed. Five number of clusters are assigned to this dataset.Graphical user interface, text, application, email

Description automatically generated

Figure 7. Obtaining the cluster labels for the neighborhoods in Toronto

**Results**

In order to effectively analyze the results, resulting cluster labels are merged to the previous venues ranking for each neighborhoods dataframe, and again joined with the toronto\_merged dataframe we created earlier. I found out that Foursquare did not detect any venues for some neighborhoods, resulting in NA values. To avoid errors, we removed the neighborhoods without venues information with pandas dropna() function.

Graphical user interface

Description automatically generated with low confidence

Figure 8. Resulting dataframe (note cluster labels in place)

**Cluster 1 (*Cluster label = 0*)**

Cluster 1 seems to be describing the downtown areas of Toronto, dominated by restaurants and cafes.

Graphical user interface

Description automatically generated with medium confidence

**Cluster 2 (*Cluster label = 1*)**

Cluster 2 has only three neighborhoods with consistently similar trending venues, with all of its first trending venues being Baseball Fields and Accessories Stores in 2nd or 3rd most common venues. It can be argued that neighborhoods in this cluster is well suited for the business problem that we are going to solve (opening up a baseball related store).

Graphical user interface

Description automatically generated

**Cluster 3 (*Cluster label = 2*)**

Cluster 3 appears to be describing the suburbs of Toronto, dominated by recreational venues such as park, rivers, playgrounds, etc.

Graphical user interface

Description automatically generated with medium confidence

**Clusters 4&5**

Both of these clusters are not significant, consisting of only one neighborhood each, so we are not going to use these clusters in our findings analysis.

Below are the results of our clustering visualized in a map. The map is generated using python’s folium library.

Map

Description automatically generated

Figure 9. Clustering results plotted in a map

**Discussions**

Based on our results, arguably cluster 2 contains the most ideal neighborhoods to open up a baseball store. Neighborhoods in cluster 2 shows enthusiasm in baseball-related venues, as reflected by trending venues in the areas that are mostly related to baseball and accessories stores.

In the last part of our notebook, I created a data frame that shows the mean values of each venue types for the five different clusters. This data frame could be useful for other applications, i.e comparing which cluster is the best place to open a BBQ Joint based on the trending venues. It allows for this project to be beneficial for other use cases.

Table

Description automatically generated

**Conclusions**

Data Science aims to solve problems in different areas by utilizing the abundance of data we have today and the modern day computational power. I successfully built a clustering model to cluster neighborhood data in Toronto based on similar attributes, in this case, trending venues. This model can be useful by offering potential insights and solution to the business problem that we have.