**Battle of The Neighborhoods**

Yang Liu

May 14 2019

1. **Introduction**

1.1 Business Problem:

As an Asian who is planning to move to Toronto from another city, how can I find the most Asian-friendly neighborhoods to settle down in Toronto?

1.2 Interest:

Asians who are planning to move to Toronto but are not familiar with Toronto.

1. **Data acquisition**

**2.1** Postcode, Borough and Neighborhood information of Toronto is retrieved from Wikipedia webpage <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>. BeautifulSoup is used to do scraping and some data clean is done to get only the Postcodes, Borough and Neighborhood information.

|  | **Postcode** | **Borough** | **Neighborhood** |
| --- | --- | --- | --- |
| **0** | M3A | North York | Parkwoods |
| **1** | M4A | North York | Victoria Village |
| **2** | M5A | Downtown Toronto | Harbourfront , Regent Park |
| **3** | M6A | North York | Lawrence Manor, Lawrence Heights |
| **4** | M7A | Queen's Park | Queen's Park |
| **5** | M9A | Etobicoke | Islington Avenue |
| **6** | M1B | Scarborough | Rouge, Malvern |
| **7** | M3B | North York | Don Mills North |
| **8** | M4B | East York | Woodbine Gardens, Parkview Hill |
| **9** | M5B | Downtown Toronto | Ryerson, Garden District |

**2.2** Geo-spatial data is retrieved from <http://cocl.us/Geospatial_data> to get Latitude and Longitude for all the Neighborhoods in Toronto

| **Postcode** | **Borough** | **Neighborhood** | **Latitude** | **Longitude** |
| --- | --- | --- | --- | --- |
| M3A | North York | Parkwoods | 43.7532 | -79.3296 |
| M4A | North York | Victoria Village | 43.725882 | -79.3155 |
| M5A | Downtown Toronto | Harbourfront , Regent Park | 43.654260 | -79.3606 |
| M6A | North York | Lawrence Manor, Lawrence Heights | 43.718518 | -79.4647 |
| M7A | Queen's Park | Queen's Park | 43.662301 | -79.3894 |

.

**2.3** Venue information is retrieved from FourSquare

| **Neighborhood** | **Neighborhood Latitude** | **Neighborhood Longitude** | **Venue** | **Venue Latitude** | **Venue Longitude** | **Venue Category** |
| --- | --- | --- | --- | --- | --- | --- |
| Parkwoods | 43.753259 | -79.329656 | Brookbanks Park | 43.751976 | -79.332140 | Park |
| Parkwoods | 43.753259 | -79.329656 | PetSmart | 43.748639 | -79.333488 | Pet Store |
| Parkwoods | 43.753259 | -79.329656 | KFC | 43.754387 | -79.333021 | Fast Food Restaurant |
| Parkwoods | 43.753259 | -79.329656 | Variety Store | 43.751974 | -79.333114 | Food & Drink Shop |
| Parkwoods | 43.753259 | -79.329656 | GreenWin pool | 43.756232 | -79.333842 | Pool |

**2.4** Ratio of Asian Restaurant among all venue categories is calculated for all the neighborhoods since it is used as a factor in k-means.

| **Neighborhood** | **Neighborhood Latitude** | **Neighborhood Longitude** | **Ratio** |
| --- | --- | --- | --- |
| Adelaide, Richmond, King | 43.650571 | -79.384568 | 0.080000 |
| Agincourt | 43.794200 | -79.262029 | 0.125000 |
| Alderwood, Long Branch | 43.602414 | -79.543484 | 0.000000 |
| Bathurst Manor, Downsview North, Wilson Heights | 43.754328 | -79.442259 | 0.047619 |
| Bayview Village | 43.786947 | -79.385975 | 0.250000 |

**2.5** The other factor for the k-means algorithm is the Top 5 venue categories for each neighborhood.

| **Neighborhood** | **1st Most Common Venue** | **2nd Most Common Venue** | **3rd Most Common Venue** | **4th Most Common Venue** | **5th Most Common Venue** |
| --- | --- | --- | --- | --- | --- |
| Adelaide, Richmond, King | Asian Restaurant | Café | Coffee Shop | American Restaurant | Steakhouse |
| Agincourt | Coffee Shop | Pool Hall | Motorcycle Shop | Badminton Court | Asian Restaurant |
| Alderwood, Long Branch | Pizza Place | Coffee Shop | Gas Station | Gym | Pharmacy |
| Bathurst Manor, Downsview North, Wilson Heights | Coffee Shop | Frozen Yogurt Shop | Asian Restaurant | Supermarket | Bank |
| Bayview Village | Bank | Asian Restaurant | Grocery Store | Skating Rink | Skate Park |

**2.6** The two data frames from 2.4 and 2.5 are merged to get the final data frame for the k-means algorithm. And since k-means can’t be applied to process float and string data at the same time, the categories are mapped into corresponding numbers,

|  | **Ratio** | **1st Most Common Venue** | **2nd Most Common Venue** | **3rd Most Common Venue** | **4th Most Common Venue** | **5th Most Common Venue** |
| --- | --- | --- | --- | --- | --- | --- |
| **Neighborhood** |  |  |  |  |  |  |
| **Adelaide, Richmond, King** | 0.080000 | 124 | 12 | 8 | 108 | 106 |
| **Agincourt** | 0.125000 | 8 | 277 | 280 | 279 | 124 |
| **Alderwood, Long Branch** | 0.000000 | 10 | 8 | 265 | 71 | 49 |
| **Bathurst Manor, Downsview North, Wilson Heights** | 0.047619 | 8 | 185 | 124 | 111 | 48 |
| **Bayview Village** | 0.250000 | 48 | 124 | 57 | 96 | 224 |

1. **Methodology**

Use k-means algorithm to cluster the neighborhoods based on 1) ratio of Asian Restaurant among all venue categories, and 2) top 5 venue categories in each neighborhood.

1. **Result**

The neighborhoods in Toronto are clustered into 10 clusters. And Cluster #7 (the yellow circles shown below) has the highest ratio of Asian restaurant and the most numbers in top 5 venue categories. So neighborhoods in cluster 7 are recommended.

