Clustering Neighborhoods for Relocation

Albert Olszewski

IBM Data Science Capstone

June 19, 2019

Overview

Background

- Average USA citizen moves 11 times in their lifetime
- Relocation is prominent in large cities due to work opportunity
- Machine learning techniques can be used to cluster similar neighborhoods from different cities
- This information would be useful for:
  - Persons relocating to new city
  - Real estate brokers wishing to provide best service

Problem Statement

Machine learning techniques will be used to group like neighborhoods in Toronto, Chicago, and New York City in order that persons moving between them can find the most familiar and comfortable living situations without exhaustive research.

Data Collection

- Neighborhood names and boroughs were scraped off Wikipedia using BeautifulSoup4 package in Python.
- Longitude and Latitude was collected using Geocoder package in Python.
- Venue and amenity data for each neighborhood was collected using the Foursquare Developer API.

Processing

- Foursquare gives us all amenities available in each neighborhood.
- One Hot Encoding to find top 10 most frequent venues and prep for clustering
- Compile into one data frame

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Adelaide,King,Richmond	Coffee Shop	Café	American Restaurant	Steakhouse	Bar
1	Agincourt	Breakfast Spot	Sandwich Place	Lounge	Chinese Restaurant	Food Court
2	Agincourt North,L'Amoreaux East,Milliken,Steel...	Park	Asian Restaurant	Playground	Yoga Studio	Falafel Restaurant
3	Albany Park	Sandwich Place	Cocktail Bar	Pizza Place	Chinese Restaurant	Hookah Bar
4	Albion Gardens,Beaumond Heights,Humbergate,Jam...	Grocery Store	Pizza Place	Coffee Shop	Sandwich Place	Fried Chicken Joint

K-Means Clustering

- Large Data Set: 655 neighborhoods
- Unsupervised learning: we don't know what the groups will be
- Simple

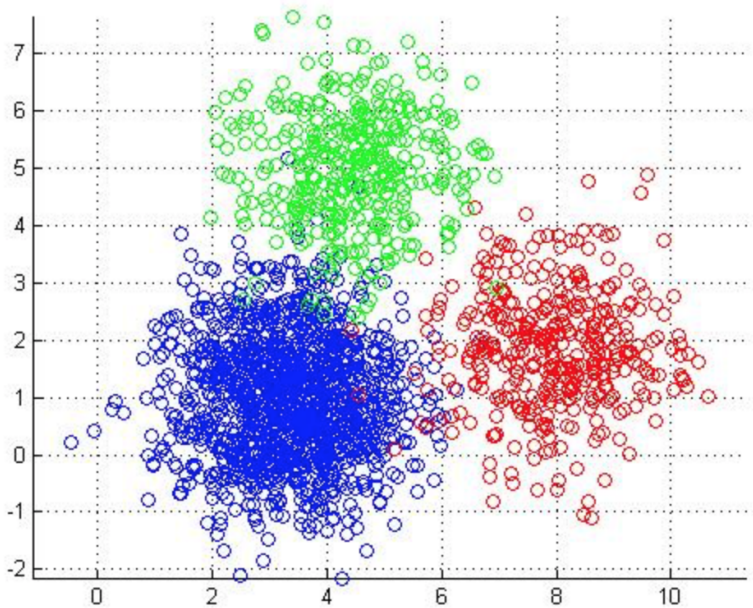


Figure: K-Means example.

Elbow Method

- Loop through values of K to minimize Sum of Squared Distances

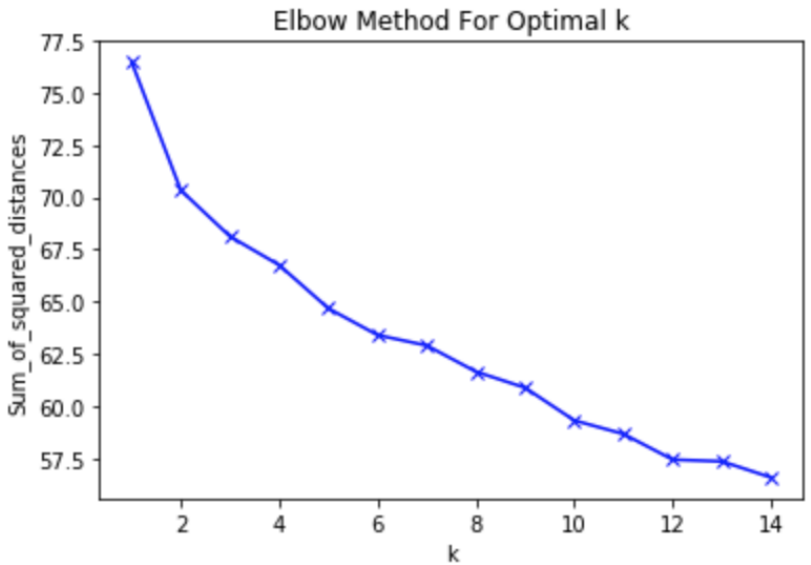
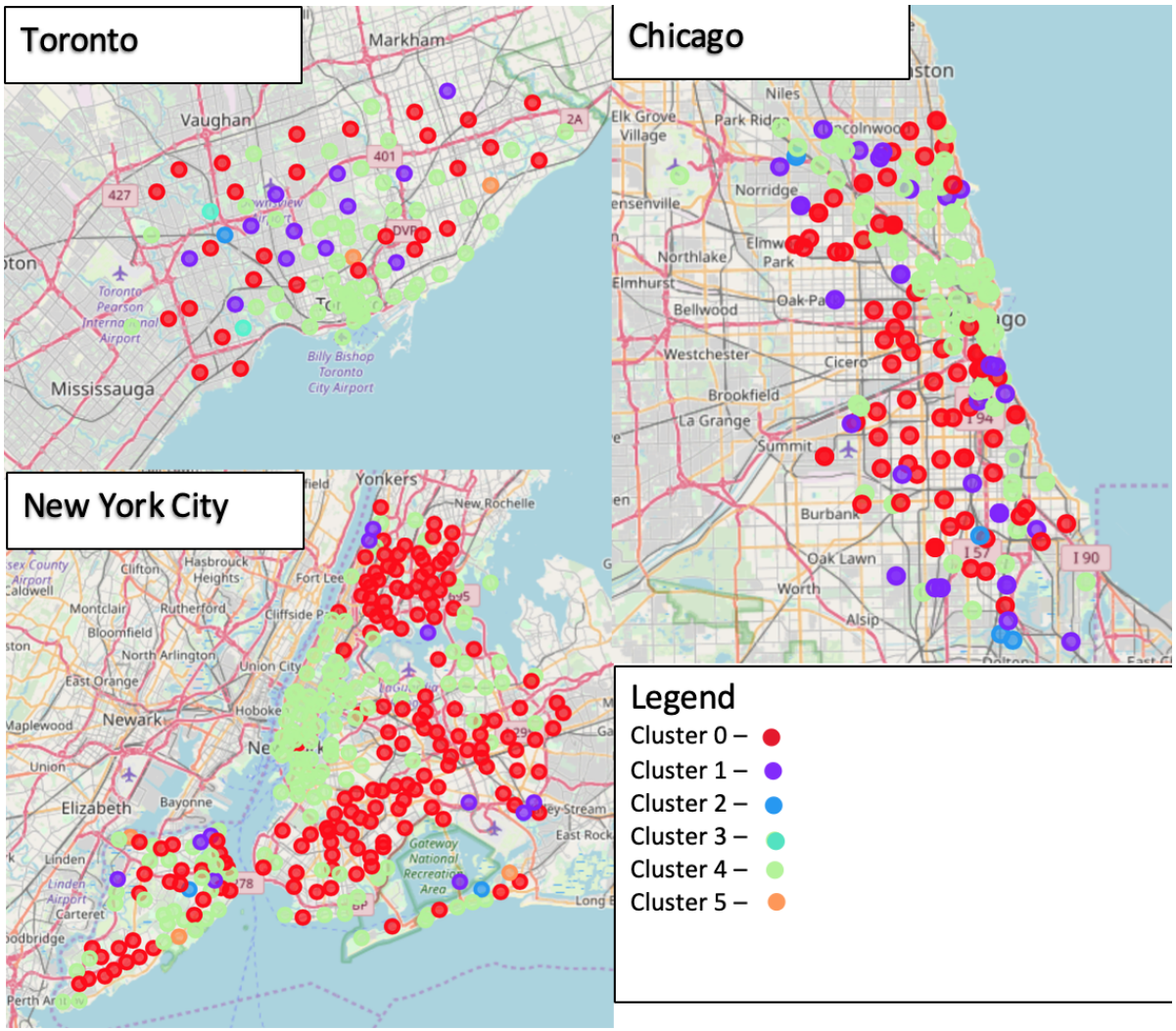


Figure: Elbow method for K-Means clustering.

- Choose K=6 because it is at elbow

Map


Toronto
Markham
Vaughan
Mississauga
Toronto Pearson International Airport
Billy Bishop Toronto City Airport
Chicago
Niles
Elk Grove Village
Norridge
Northlake
Elmhurst
Bellwood
Westchester
Cicero
Brookfield
La Grange
Summit
Burbank
Oak Lawn
Worth
Alsip
New York City
Yonkers
New Rochelle
Clifton
Hasbrouck Heights
Fort Lee
Montclair
Rutherford
Bloomfield
North Arlington
Union City
East Orange
Newark
Hoboke
Maplewood
Union
Elizabeth
Bayonne
Linden
Carteret
Gateway National Recreation Area
Legend
Cluster 0 –
Cluster 1 –
Cluster 2 –
Cluster 3 –
Cluster 4 –
Cluster 5 –

Cluster 0

Venue	Count
Pizza Place	80
Chinese Restaurant	48
Mexican Restaurant	34
Fast Food	32
Pharmacy	31

Table: A count of how many times a venue is the most common in a neighborhood for Cluster 0.

Cluster 1

Venue	Count
Park	51
Fast Food	7
Bus Station	6
Deli / Bodega	6
Convenience Store	5

Table: A count of how many times a venue is the most common in a neighborhood for Cluster 1.

Cluster 2

Venue	Count
Park	8
Yoga Studio	8
Fountain	4
Campground	2
Grocery Store	1

Table: A count of how many times a venue is the most common in a neighborhood for Cluster 2.

Cluster 3

Venue	Count
Electronics Store	2
Yoga Studio	2
Baseball Stadium	2

Table: A count of how many times a venue is the most common in a neighborhood for Cluster 3.

Cluster 4

Venue	Count
Coffee Shop	103
Bar	53
Italian Restaurant	41
Sandwich Shop	40
Cafe	34

Table: A count of how many times a venue is the most common in a neighborhood for Cluster 4.

Cluster 5

Venue	Count
Yoga Studio	3
Playground	3
Bar	2
Tennis Court	2
Park	1

Table: A count of how many times a venue is the most common in a neighborhood for Cluster 5.

Conclusion

- Collected data through: web scraping, geocoders, and Foursquare Developer API.
- Processed and cleaned data to prep for clustering.
- Performed K-Means clustering with an optimal value of K=6
- Provided maps and information for a consumer to make the most familiar and comfortable choice when relocating between cities.

Future Work

- More cities: San Francisco, Seattle, Houston, etc.
- More sources of information: cultural, pricing, religious, etc.
- Develop a user friendly tool for cross-referencing neighborhoods.

References

 Chandler, Adam. "Why Do Americans Move So Much More Than Europeans?" October 2016. Web.

 Hacker Noon. "What Is One Hot Encoding? Why And When Do Your Have To Use It?" August 2017. Web.