The Battle of Neighborhoods - Search for Similar Cities for Graduating University Students

Dexin Wang for Applied Data Science Capstone Final Assignment

March 2020

Problem Statement and Project Objective

- ## Problem Statement

  Graduating students are looking for similar cities to New York for searching jobs. The first step was to find out which cities are similar in terms of nearby venues.

- ## Project Objective

  To help the students make decision of relocation, the project objective is to collect and analyze city or neighborhood venue data, from which draw insights about which cities are more similar to the current city they live in. In addition to the lists of similar cities, the most popular venues are also presented for further information along with grouped city visualization.

Finding Data from Analytic Approach & Requirement

- Analytic Approach:

  Based on the needs of the students, it is appropriate to find city neighborhood similarity by clustering the similar cities as groups.
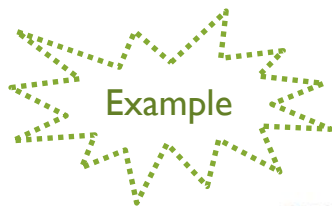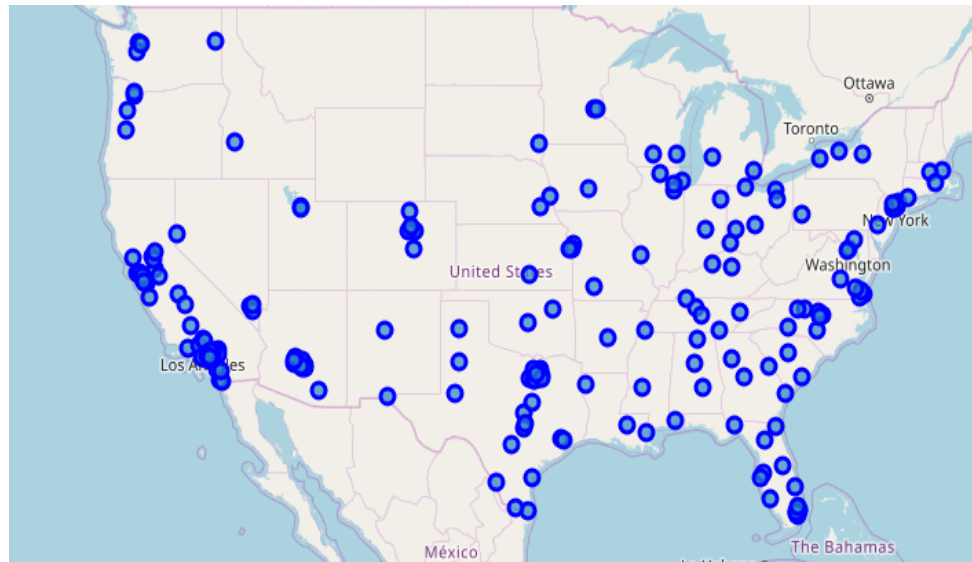
- Data Requirement:

  ➢ Major cities across the US with large populations

  ➢ All searchable venues in neighborhoods defined by given location latitude and longitude coordinates

- Data Sources:

  ◦ http://worldpopulationreview.com/us-cities/#cities

  ◦ https://foursquare.com

City Data Description

- City Location & Population – A list of 200 large cities are included

Example

Here is the table read from the http://worldpopulationreview.com/us-cities:

	Rank	Name	State	2020 Population	2010 Census	Change	2020 Density	Latitude/Longitude	Area (km²)
0	1	New York	New York	8,622,357	8,175,133	0.25%	11,084/km²	40.66/-73.94	778
1	2	Los Angeles	California	4,085,014	3,792,621	0.67%	3,365/km²	34.02/-118.41	1,214
2	3	Chicago	Illinois	2,670,406	2,695,598	-0.32%	4,535/km²	41.84/-87.68	589
3	4	Houston	Texas	2,378,146	2,099,451	0.79%	1,443/km²	29.79/-95.39	1,649
4	5	Phoenix	Arizona	1,743,469	1,445,632	1.88%	1,300/km²	33.57/-112.09	1,341

City Data Description

- City Venues –
  - Up to 100 venues for each city neighborhood are included
  - There are 434 unique categories

Example
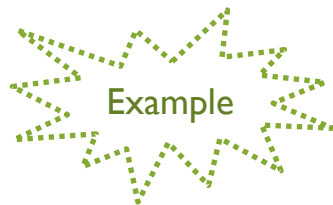
	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	New York	40.789624	-73.959894	North Meadow	40.792027	-73.959853	Park
1	New York	40.789624	-73.959894	Central Park Tennis Center	40.789313	-73.961862	Tennis Court
2	New York	40.789624	-73.959894	East Meadow	40.790160	-73.955498	Field
3	New York	40.789624	-73.959894	Central Park - Woodman's Gate	40.787786	-73.955924	Park
4	New York	40.789624	-73.959894	The Jewish Museum	40.785276	-73.957411	Museum

# Method - Extracting Features to Characterizing the City Similarity

- Charactering cities by numbers of venues in venue categories. For example, a city neighborhood may have more coffee shops while another city may have more parks

- Adding population density as another factor

- Features need to be normalized for calculating "distance" measuring city similarity

Example

	City	ATM	Accessories Store	Adult Boutique	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	Airport Service	Airport Terminal	American Restaurant	Antique Shop
0	Akron	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.031250	0.0
1	Albuquerque	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.041667	0.0
2	Alexandria	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
3	Amarillo	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.035714	0.0
4	Anaheim	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.034483	0.0

Method - City Clustering to Group the Similar Cities
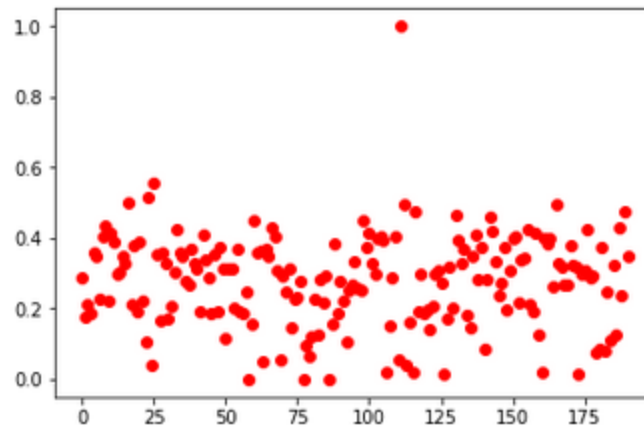
- Using Silhouette and Dendrogram to support cluster number selection

- Cross-check K-means & agglomerative results – close results were achieved
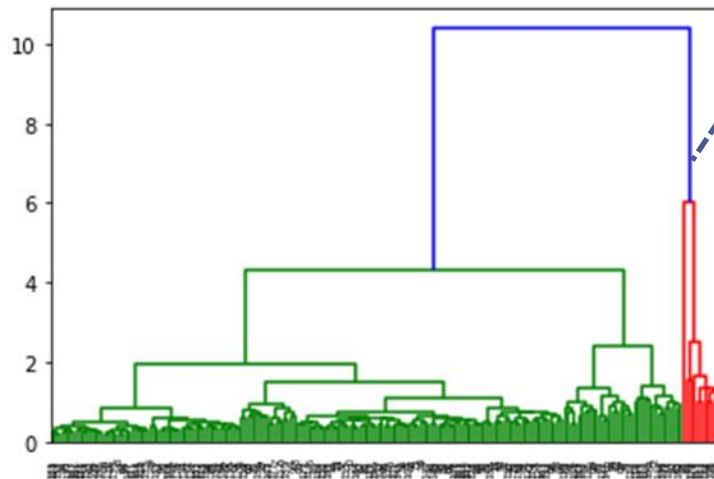
Dendrogram – dissimilarity indicated by Y-value

Results

- There are 121 cities in the same cluster as New York
- Checking cosine similarity indicates the highest similar city is Chicago

Plot of Cosine Similarity Scores (Y) vs. City Indices (X)

Most Common Venues

City	Longitude	Agg Cluster Labels	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
New York	-73.959894	0.0	5.0	Park	Grocery Store	Café	Pizza Place	Pharmacy	Deli / Bodega	Playground	Bakery	Italian Restaurant
Los Angeles	-118.242767	0.0	5.0	Coffee Shop	Sushi Restaurant	Mexican Restaurant	Bakery	Indian Restaurant	Thai Restaurant	Japanese Restaurant	Ice Cream Shop	Italian Restaurant
Chicago	-87.624421	0.0	5.0	Italian Restaurant	Coffee Shop	Hotel	Grocery Store	Pizza Place	Garden	Theater	Gym / Fitness Center	Sandwich Place
Houston	-95.367697	0.0	5.0	Hotel	Mexican Restaurant	Coffee Shop	Bar	Park	Burger Joint	American Restaurant	Southern / Soul Food Restaurant	Cocktail Bar
Phoenix	-112.074142	0.0	5.0	Coffee Shop	Pizza Place	Hotel	Mexican Restaurant	American Restaurant	Fast Food Restaurant	Bar	Music Venue	Convenience Store

Results Discussion – What Insights Drawn from the Data

- It is not intuitive to use the 435 features to interpret why the cities are similar

- Checking the most similar cities in other data source such as www.areavibes.com, indicating they are rated with A+ amenities

- One cluster has overwhelming number of cities, indicating more differentiating features may be helpful
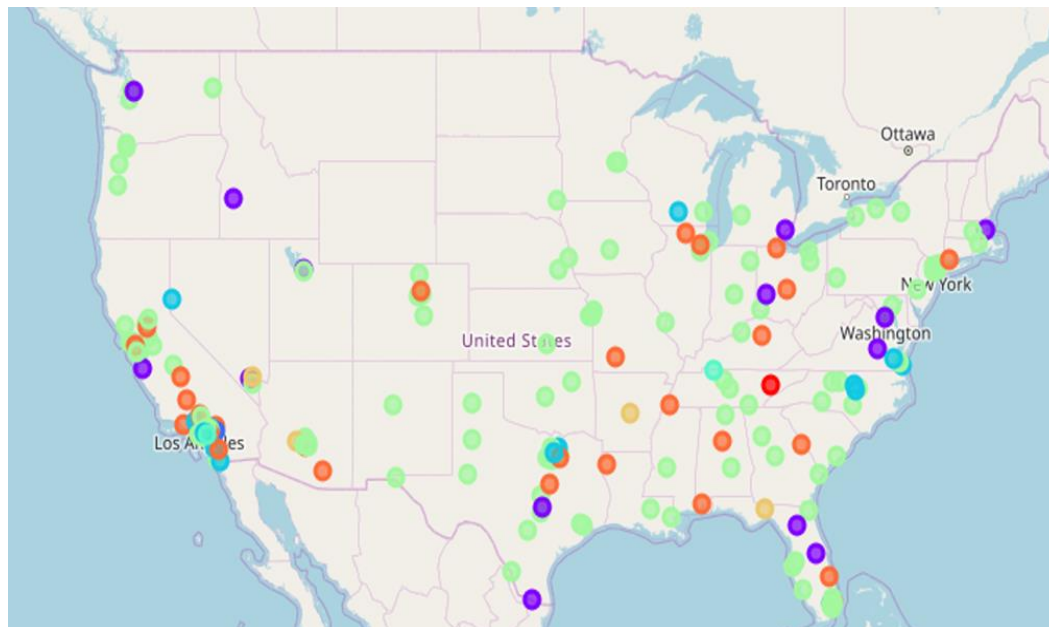
Different cluster are colored

Conclusions

- There are over 120 big city neighborhoods are similar to New York city in terms of venues – good news there are many similar cities to begin with for additional comparison

- Chicago has the relatively higher similarity score to the city

- The features used in this study are unfortunately very limited due to number of venues could be searched in each neighborhood. Therefore results can be improved if more venues could be searched

- Further study could add more features to narrow down the list of similar cities