**Analysis of City Similarities**

**Introduction**

The top 10 largest US cities by populations are in Table 1:

|  |  |  |  |
| --- | --- | --- | --- |
| **Cities** | **Latitude** | **Longitude** |  |
| **0** | New York City, NY | 40.712728 | -74.006015 |
| **1** | Los Angeles, CA | 34.053691 | -118.242766 |
| **2** | Chicago, IL | 41.875562 | -87.624421 |
| **3** | Houston, TX | 29.758938 | -95.367697 |
| **4** | Phoenix, AZ | 33.448437 | -112.074142 |
| **5** | Philadelphia, PA | 39.952724 | -75.163526 |
| **6** | San Antonio, TX | 29.424600 | -98.495141 |
| **7** | San Diego, CA | 32.717420 | -117.162773 |
| **8** | Dallas, TX | 32.776272 | -96.796856 |
| **9** | San Jose, CA | 37.336191 | -121.890583 |

Table 1: the top 10 largest US cities by populations

I'm curious about the similarity among them and going to cluster them using venue-based similarity measurement. It will be useful for movers from those 10 cities to find the one that they are already familiar with and like to live.

**Data Description**

I first converted the address to the latitude and longitude for each of the 10 cities using geopy.geocoders.Nominatim. The result is shown in Table 1.

Then I fetched the top 100 venues that are within a radius of 500 meters of each city. (Foursquare API)

**Data Processing Workflow**

* Transform the venue categories to hot code for analysis convenience
* Normalize the data
* Calculate the Euclidean distances to show how different the 10 cities are from each other
* Get the distance matrix

**Analysis of the Distance Matrix**

The distance matrix represents the difference between each pair of cities. Take the 4th column as an example:

|  |  |  |
| --- | --- | --- |
|  | **Cities** | **Difference** |
| **4** | New York City, NY | 0.000000 |
| **1** | Dallas, TX | 6.836431 |
| **7** | San Antonio, TX | 6.843296 |
| **2** | Houston, TX | 7.042819 |
| **6** | Phoenix, AZ | 7.059759 |
| **3** | Los Angeles, CA | 7.160766 |
| **0** | Chicago, IL | 7.287910 |
| **5** | Philadelphia, PA | 7.643115 |
| **9** | San Jose, CA | 7.736015 |
| **8** | San Diego, CA | 7.758686 |

These values stand for the difference between New York City, NY and the rest 9 cities. We can see the values is the least for Dallas, TX, which means it is the most similar city to New York City, NY.

**Clustering the cities**

I then used scikit-learn to generate Hierarchical Clustering of the top 10 cities and visualize the result in Figure 1. It clearly shows the similarities among cities. For example, Houston TX is very similar to Dallas, TX, which is not surprising.

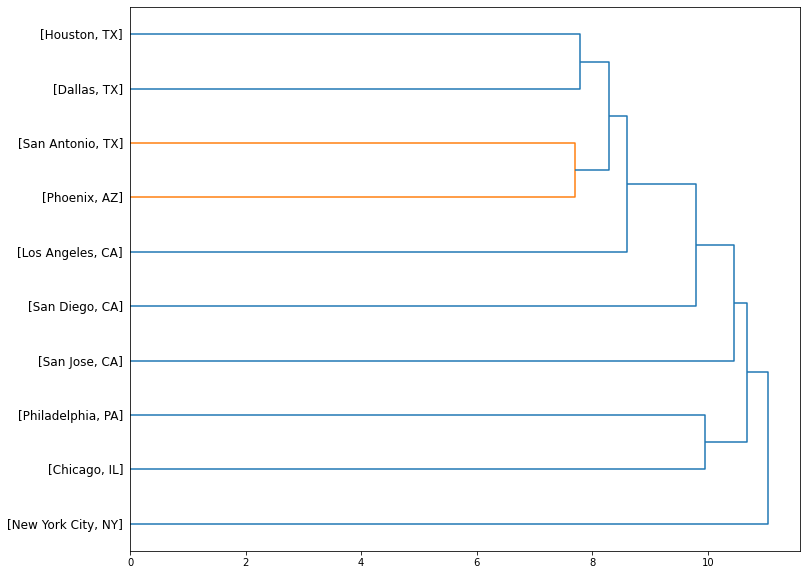


Figure 1: Hierarchical Clustering of the top 10 cities.

**Conclusion**

The Hierarchical Clustering of the top 10 cities show us the similarities among cities. It will be useful for movers from those 10 cities to find the one that they are already familiar with and like to live. For example, I’m now living in Houston, TX. I find that Dallas, TX is very similar to Houston, TX. If I can find a better job in Dallas, I won’t hesitate to move there.