Clustering the Fifty Largest Metropolitan Areas in the World

1. Introduction

I am running an immigration agency in Paris, France. Many of my clients who want to start a new life ask me which cities are most culturally similar to Paris. I believe that the types of venues in each city reflect each city's distinct culture. In this lab, clustering is used to find cities similar to Paris.

2. Data

Using the BeautifulSoup library, a list of largest cities in the world was scraped from Wikipedia (https://en.wikipedia.org/wiki/List_of_largest_cities). The list of cities was then populated with each city's respective longitudes and latitudes using the Geocoder package. These coordinates were run in the Foursquare API to get the venue categories in each city. After the data was cleaned, k-means clustering was applied.

3. Methodology

K-means clustering was used in this model. K-means algorithm is fast, robust, and simple algorithm, and given the limited processing power on my CPU, it made the most sense. Since I had a table of frequencies of venues, these frequencies were converted into Euclidean distances that the K-means algorithm minimizes and uses to develop cluster centroids.

4. Results

My home city, Paris, was in a cluster with Tokyo, São Paulo, Istanbul, Rio de Janeiro, Los Angeles, Moscow, Paris, Seoul, Nagoya, Tehran, and Chicago. Below are the complete results:

City	Country	Cluster
Delhi	India	0
Cairo	Egypt	0
Beijing	China	0
Manila	Philippines	0
Kinshasa	DR Congo	0
Bangkok	Thailand	0
Nanjing	China	0
Ho Chi Minh City	Vietnam	0
Kuala Lumpur	Malaysia	0
Hong Kong	China	0
Shanghai	China	1

Chongqing	China	1
Guangzhou	China	1
Jakarta	Indonesia	1
Chengdu	China	1
Xi'an	China	1
Hangzhou	China	1
Tianjin	China	2
Wuhan	China	2
Dongguan	China	2
Tokyo	Japan	3
São Paulo	Brazil	3
Istanbul	Turkey	3
Rio de Janeiro	Brazil	3
Los Angeles	United States	3
Moscow	Russia	3
Paris	France	3
Seoul	South Korea	3
Nagoya	Japan	3
Tehran	Iran	3
Chicago	United States	3
Karachi	Pakistan	4
Dhaka	Bangladesh	5
Luanda	Angola	5
New York City	United States	6
Bogotá	Colombia	6
Lima	Peru	6
London	U.K.	6
Kolkata	India	7
Chennai	India	7
Ahmedabad	India	7
Mexico City	Mexico	8
Mumbai	India	8
Buenos Aires	Argentina	8
Lagos	Nigeria	8
Lahore	Pakistan	8
Bangalore	India	8
Hyderabad	India	8
Osaka	Japan	9
Shenzhen	China	9

5. Discussion

I found that the clustering was pretty arbitrary. However, this is understandable as the human brain can't possibly process 400+ dimensional data is this way. In the future, I would like play with the number of clusters and add other factors like per-capita GDP, urban density, and tourism numbers.

6. Conclusion

Overall, this was a very informative lab that exposed me to the fundamentals of Python and machine learning. It was very helpful to go through the whole data science life cycle, from scraping the data to creating a model. I got very interesting results that I will consider in my career, and I will continue to tweak this model in the future.