Report

2018

To find out best place for chenzhu to date

czhubj@cn.ibm.com

目录

Content

INTRODCTION

It is difficult to find a suitable One day's theme dating venue, such as the need to find a place in the middle of the week, you can have a quiet place to date dinner, but press the road to the subway station, hope to find a lively place on the weekend, in addition to eating and shopping, at the same time for the choice of cuisine for the choice The choice for single is a test. And now most of the software, such as the public comment, can only be based on the recommendations of the location and play, it is difficult to know with the interest of single men and women, this experiment will take me the subway station often visited as location information, go to foursquare I explored the types of play in the street, intuitively sorted the data from the places I used to go, extracted features, and helped me to visualize the characteristics of the date.

BUSINESS PROBLEM

Can cluster the place we dating and help us to learn about them for making decision with different purpose?

DATA

Data Collection

An open source data set of a foreign institution. Foursquare is a technology company that uses location intelligence to build meaningful consumer experiences and business solutions.The content of the data includes Pinpoint by Foursquare, Attribution by Foursquare, Place Insights by Foursquare and Pilgrim SDK by Foursquare（https://foursquare.com/） Further more, we also need to draw a map for the place I often vistied with the help of the tool (http://www.gpsspg.com/maps.htm)

Data Understanding and Preparation

This data is all marked and unstructure data ,and there is not training set and test set. It is all about exploring and to see the feature. As for the data ,what we should do is to standardlize and visualize for overview understanding. Maybe we should pay more attention to Feature selection and the processing of missing values.

METHODOLOGY

Analytic Approach

First of all, the data we get is the street comprehensive data based on the latitude and longtitude from foursquare. This data contains many dimensions. The problem to be solved is to explore the feature in cluster. There are not limited standard for weather it is right or false. Because we do not need to predict at this stage. We just want to find out the result of clusters from the data in entertainment facilities. The algorithm selected in this project is K-means.).Clustering is the division of data into groups such that data points in the same group are more similar than data points in other groups. In short, clustering is the division of data points with similar characteristics into groups, that is, clusters. The goal of the K-means algorithm is to find a group in the data, the number of groups being represented by the variable K. Each data point is assigned to one of the K groups by an iterative operation based on the characteristics provided by the data.

As for the advantages and disadvantages:

Advantages:

1. The algorithm is simple and easy to implement;
2. .For processing large data sets, the algorithm is relatively scalable and efficient because its complexity is approximately O(nkt), where n is the number of all objects, k is the number of clusters, and t is the number of iterations. Usually k<<n. This algorithm usually converges locally.
3. The algorithm attempts to find the k partitions that minimize the value of the squared error function. When the clusters are dense, spherical or lumpy, and the difference between clusters and clusters is obvious, the clustering effect is better.

Disadvantage:

1. High requirements on data types, suitable for numerical data;
2. May converge to a local minimum and converge slowly on large-scale data.
3. K value is more difficult to select;

4. Sensitive to the cluster value of the initial value, which may result in different clustering results for different initial values;

5. Not suitable for finding clusters with non-convex shapes, or clusters with large differences in size.6.Sensitive to "noise" and outlier data, a small amount of this type of data can have a significant impact on the average.

Conclusion for selecting K-means

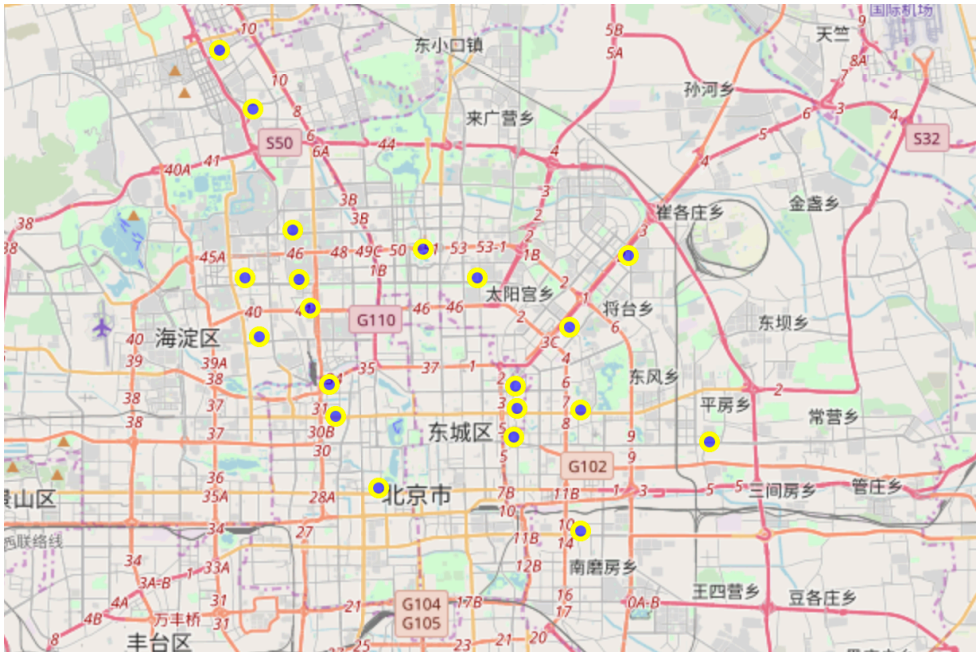
Compared to the DBscan, DBscan is based on density calculation clustering, which will eliminate exceptions (noise points). The noise point will be deleted and clustered by the dbscan algorithm (not in the core point and not in the neighborhood of the core point). As for the station, most of them are the same, we should pay more attention finding most common features in different station to help chenzhu to descide

RESULT

We have selected about 20 subway station where chenzhu oftenly visit and explore the possbilities around the subways.

	Station	Latitude	Longitude
0	XIERQI	40.052243	116.306144
1	SHANGDI	40.032958	116.320519
2	WUDAOKOU	39.992833	116.337780
3	ZHICHUNLU	39.976424	116.340141
4	DAZHONGSI	39.966923	116.345126
5	XIZHIMEN	39.941856	116.353234
6	CHEGONGZHUANG	39.931445	116.356180
7	DONGZHIMEN	39.941255	116.433859
8	HUIXINXIJIENANKOU	39.976998	116.417644
9	SANYUANQIAO	39.960879	116.457055
10	TUANJIEHU	39.933450	116.461705
11	CHAOYANGMEN	39.924540	116.433433
12	HAIDIANHUANGZHUANG	39.976909	116.317043
13	AOTIZHONGXIN	39.986407	116.394155
14	WANGJINGNAN	39.984704	116.482311
15	DONGSISHITIAO	39.933852	116.434333
16	XIDAN	39.907769	116.374751
17	SHUANGJING	39.893557	116.461962
18	QINGNIANLU	39.923018	116.517454
19	WEIGONGCUN	39.957765	116.323136

And we visualize them on map.



After we requested from foursquare for the facilities around station we choose, there are 1468 results there. And we check the categories in 1468 items.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
AOTIZHONGXIN	63	63	63	63	63	63
CHAOYANGMEN	100	100	100	100	100	100
CHEGONGZHUANG	67	67	67	67	67	67
DAZHONGSI	55	55	55	55	55	55
DONGSISHITIAO	100	100	100	100	100	100
DONGZHIMEN	100	100	100	100	100	100
HAIDIANHUANGZHUANG	100	100	100	100	100	100
HUIXINXIJIENANKOU	57	57	57	57	57	57
QINGNIANLU	40	40	40	40	40	40
SANYUANQIAO	100	100	100	100	100	100
SHANGDI	32	32	32	32	32	32
SHUANGJING	84	84	84	84	84	84
TUANJIEHU	100	100	100	100	100	100
WANGJINGNAN	100	100	100	100	100	100
WEIGONGCUN	49	49	49	49	49	49
WUDAOKOU	76	76	76	76	76	76
XIDAN	100	100	100	100	100	100
XIERQI	17	17	17	17	17	17
XIZHIMEN	46	46	46	46	46	46
ZHICHUNLU	82	82	82	82	82	82

There are 155 uniques categories.

RESULT

We changed the data of string for categories in to number and show the frequency.

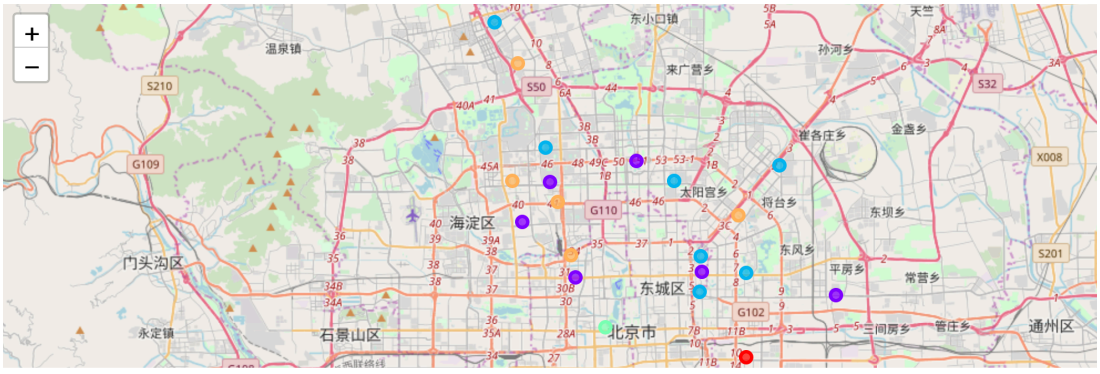
	Neighborhood	American Restaurant	Antique Shop	Aquarium	Arepa Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	BBQ Joint	...	Vegetarian / Vegan Restaurant	Vietnamese Restaurant	Wine Bar	Wings Joint	Xinjiang Restaurant	Yoga Studio	Yunnan Restaurant	Zhejiang Restaurant	Zoo	Zoo Exhibit
0	AOTIZHONGXIN	0.000000	0.00	0.000000	0.00	0.015873	0.00	0.00	0.047619	0.000000	...	0.000000	0.00	0.00	0.00	0.000000	0.00	0.000000	0.00	0.000000	0.000000
1	CHAOYANGMEN	0.000000	0.00	0.000000	0.01	0.000000	0.00	0.00	0.010000	0.000000	...	0.000000	0.00	0.00	0.00	0.020000	0.00	0.010000	0.00	0.000000	0.000000
2	CHEGONGZHUANG	0.014925	0.00	0.000000	0.00	0.000000	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.00	0.00	0.014925	0.00	0.000000	0.00	0.000000	0.000000
3	DAZHONGSI	0.000000	0.00	0.000000	0.00	0.000000	0.00	0.00	0.000000	0.018182	...	0.000000	0.00	0.00	0.00	0.018182	0.00	0.000000	0.00	0.000000	0.000000
4	DONGSISHITIAO	0.000000	0.00	0.000000	0.01	0.000000	0.00	0.00	0.010000	0.000000	...	0.000000	0.01	0.00	0.00	0.010000	0.01	0.010000	0.00	0.000000	0.000000
5	DONGZHIMEN	0.000000	0.00	0.000000	0.01	0.000000	0.00	0.00	0.010000	0.000000	...	0.010000	0.01	0.01	0.00	0.010000	0.01	0.020000	0.00	0.000000	0.000000
6	HAIDIANHUANGZHUANG	0.010000	0.00	0.000000	0.00	0.000000	0.00	0.00	0.010000	0.020000	...	0.010000	0.00	0.00	0.01	0.000000	0.00	0.020000	0.00	0.000000	0.000000
7	HUIXINXIJIENANKOU	0.017544	0.00	0.000000	0.00	0.000000	0.00	0.00	0.070175	0.017544	...	0.000000	0.00	0.00	0.00	0.000000	0.00	0.000000	0.00	0.000000	0.000000
8	QINGNIANLU	0.000000	0.00	0.000000	0.00	0.000000	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.00	0.00	0.000000	0.00	0.000000	0.00	0.000000	0.000000
9	SANYUANQIAO	0.010000	0.00	0.000000	0.00	0.000000	0.00	0.00	0.020000	0.020000	...	0.000000	0.00	0.00	0.00	0.000000	0.00	0.000000	0.00	0.000000	0.000000
10	SHANGDI	0.000000	0.00	0.000000	0.00	0.000000	0.00	0.00	0.031250	0.000000	...	0.000000	0.00	0.00	0.00	0.000000	0.00	0.000000	0.00	0.000000	0.000000
11	SHUANGJING	0.000000	0.00	0.000000	0.00	0.023810	0.00	0.00	0.011905	0.000000	...	0.000000	0.00	0.00	0.00	0.000000	0.00	0.000000	0.00	0.000000	0.000000
12	TUANJIEHU	0.000000	0.00	0.000000	0.01	0.000000	0.00	0.00	0.000000	0.000000	...	0.000000	0.01	0.01	0.00	0.000000	0.00	0.000000	0.00	0.000000	0.000000
13	WANGJINGNAN	0.040000	0.00	0.000000	0.00	0.040000	0.01	0.01	0.010000	0.000000	...	0.000000	0.00	0.00	0.00	0.000000	0.00	0.010000	0.01	0.000000	0.000000
14	WEIGONGCUN	0.020408	0.00	0.020408	0.00	0.000000	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.00	0.00	0.020408	0.00	0.000000	0.00	0.000000	0.000000
15	WUDAOKOU	0.000000	0.00	0.000000	0.00	0.000000	0.00	0.00	0.013158	0.013158	...	0.026316	0.00	0.00	0.00	0.052632	0.00	0.000000	0.00	0.000000	0.000000
16	XIDAN	0.020000	0.01	0.000000	0.00	0.000000	0.00	0.00	0.010000	0.000000	...	0.000000	0.00	0.00	0.00	0.000000	0.00	0.010000	0.00	0.000000	0.000000
17	XIERQI	0.000000	0.00	0.000000	0.00	0.000000	0.00	0.00	0.058824	0.000000	...	0.000000	0.00	0.00	0.00	0.000000	0.00	0.000000	0.00	0.000000	0.000000
18	XIZHIMEN	0.000000	0.00	0.021739	0.00	0.000000	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.00	0.00	0.021739	0.00	0.021739	0.00	0.021739	0.021739
19	ZHICHUNLU	0.000000	0.00	0.000000	0.00	0.000000	0.00	0.00	0.024390	0.024390	...	0.000000	0.00	0.00	0.00	0.024390	0.00	0.000000	0.00	0.000000	0.000000

20 rows × 156 columns

We decided to choose TOP 10 features to cluster.

	Neighborhood	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	10th Most Common Venue
0	AOTIZHONGXIN	Chinese Restaurant	Hotel	Coffee Shop	Fast Food Restaurant	Asian Restaurant	Park	Stadium	Lounge	Shopping Mall	Bus Stop
1	CHAOYANGMEN	Hotel	Shopping Mall	Italian Restaurant	Cocktail Bar	Chinese Restaurant	Café	Japanese Restaurant	Coffee Shop	Eastern European Restaurant	Peking Duck Restaurant
2	CHEGONGZHUANG	Coffee Shop	Hotel	Chinese Restaurant	Department Store	Fast Food Restaurant	Café	Theater	Hunan Restaurant	Shopping Mall	Hostel
3	DAZHONGSI	Fast Food Restaurant	Chinese Restaurant	Café	Coffee Shop	Restaurant	Shopping Mall	Convenience Store	Grocery Store	Movie Theater	Hotel
4	DONGSISHITIAO	Hotel	Shopping Mall	Japanese Restaurant	Cocktail Bar	Chinese Restaurant	Peking Duck Restaurant	Szechuan Restaurant	Café	Brewery	Coffee Shop
5	DONGZHIMEN	Hotel	Japanese Restaurant	Brewery	Chinese Restaurant	Peking Duck Restaurant	Café	Cocktail Bar	Shopping Mall	Coffee Shop	Spanish Restaurant
6	HAIDIANHUANGZHUANG	Coffee Shop	Chinese Restaurant	Fast Food Restaurant	Café	Hotel	Bar	Department Store	Bakery	Multiplex	Clothing Store
7	HUIXINXIJIENANKOU	Coffee Shop	Chinese Restaurant	Café	Asian Restaurant	Hotel	Gym	Pizza Place	Multiplex	Fast Food Restaurant	Karaoke Bar
8	QINGNIANLU	Chinese Restaurant	Supermarket	Coffee Shop	Hotpot Restaurant	Hotel	Clothing Store	Shopping Mall	Multiplex	Italian Restaurant	Skating Rink
9	SANYUANQIAO	Japanese Restaurant	Hotel	Chinese Restaurant	Coffee Shop	Café	Italian Restaurant	Park	Bakery	Mexican Restaurant	New American Restaurant
10	SHANGDI	Coffee Shop	Chinese Restaurant	Hotel	Fast Food Restaurant	Cantonese Restaurant	Bus Stop	Tea Room	Park	Department Store	Shopping Mall
11	SHUANGJING	Coffee Shop	Chinese Restaurant	Hotel	Café	Fast Food Restaurant	Shopping Mall	Multiplex	Supermarket	Hotel Bar	Italian Restaurant
12	TUANJIEHU	Japanese Restaurant	Hotel	Cocktail Bar	Chinese Restaurant	Café	Brewery	French Restaurant	Shopping Mall	Italian Restaurant	Peking Duck Restaurant
13	WANGJINGNAN	Café	Coffee Shop	Chinese Restaurant	Korean Restaurant	Hotel	Art Gallery	Shopping Mall	American Restaurant	Japanese Restaurant	Cantonese Restaurant
14	WEIGONGCUN	Hotel	Fast Food Restaurant	Coffee Shop	Chinese Restaurant	Café	Pizza Place	Movie Theater	Department Store	Grocery Store	Museum
15	WUDAOKOU	Café	Chinese Restaurant	Fast Food Restaurant	Korean Restaurant	Coffee Shop	Bar	Xinjiang Restaurant	Hotel	Japanese Restaurant	Bakery
16	XIDAN	Historic Site	Coffee Shop	Hotel	Chinese Restaurant	Shopping Mall	Department Store	Café	Hostel	Bookstore	History Museum
17	XIERQI	Coffee Shop	Hotel	Fast Food Restaurant	Hunan Restaurant	Korean Restaurant	Farmers Market	Chinese Restaurant	Food Court	Asian Restaurant	Café
18	XIZHIMEN	Chinese Restaurant	Fast Food Restaurant	Coffee Shop	Café	Shopping Mall	Zoo Exhibit	Stadium	Russian Restaurant	Hotpot Restaurant	Planetarium
19	ZHICHUNLU	Chinese Restaurant	Fast Food Restaurant	Café	Korean Restaurant	Hotel	Coffee Shop	Bar	Bakery	Szechuan Restaurant	Dessert Shop

And the final result is like this there 5 clusters in the map.



DISCUSSION

We check the details of 5 clusters.

There 2 cluster only contain 1 item, so we just discuss the rest 3 cluster.

We count the second cluster and we found 34 cat here and most of them are Coffee Shop, Chinese Restaurant, Fast Food Restaurant and Café. We call it local style cluster.

```
Out[58]: Coffee Shop              6
         Chinese Restaurant       6
         Hotel                    6
         Shopping Mall            4
         Fast Food Restaurant     4
         Café                     4
         Department Store         2
         Szechuan Restaurant      2
         Cocktail Bar             1
         Pizza Place              1
         Multiplex                1
         Italian Restaurant       1
         Korean Restaurant        1
         Supermarket              1
         Movie Theater            1
         Asian Restaurant         1
         Hunan Restaurant         1
         Theater                  1
         Lounge                   1
```

DISCUSSION

We count the third cluster and seepic hereWe found 29 cat here and most of them are Hotel,Japanese Restaurant,Shopping Mall,Korean Restaurant,Cocktail Bar,Fast Food Restaurant,Peking Duck Restaurant, Italian Restaurant ,Brewery,Asian Restaurant,French Restaurant,American Restaurant,Spanish Restaurant, Eastern European Restaurant,Cantonese Restaurant,Xinjiang Restaurant,Hunan Restaurant and Pizza Place. We call it tasty food in different countries cluster.

Out[60]:

Hotel	7
Chinese Restaurant	7
Café	7
Coffee Shop	6
Japanese Restaurant	5
Shopping Mall	4
Korean Restaurant	3
Cocktail Bar	3
Fast Food Restaurant	3
Peking Duck Restaurant	3
Italian Restaurant	2
Brewery	2
Asian Restaurant	2
French Restaurant	1
Gym	1
American Restaurant	1
Spanish Restaurant	1
Multiplex	1
Eastern European Restaurant	1
Cantonese Restaurant	1
Art Gallery	1
Xinjiang Restaurant	1
Karaoke Bar	1
Hunan Restaurant	1
Farmers Market	1
Food Court	1
Bar	1
Pizza Place	1
Bakery	1

DISCUSSION

We count the fifth cluster and We found 28 cat here and most of them are between the second cluster and the third cluster. We call it choose it when you don't know what to choose cluster.

Out[56]:

Chinese Restaurant	5
Coffee Shop	5
Fast Food Restaurant	4
Hotel	4
Café	4
Shopping Mall	3
Bakery	2
Park	2
Department Store	2
Bus Stop	1
Italian Restaurant	1
Cantonese Restaurant	1
Clothing Store	1
Convenience Store	1
Planetarium	1
Zoo Exhibit	1
Russian Restaurant	1
Stadium	1
Restaurant	1
Mexican Restaurant	1
Japanese Restaurant	1
Hotpot Restaurant	1
Bar	1
New American Restaurant	1
Tea Room	1
Grocery Store	1
Movie Theater	1
Multiplex	1

CONCLUSION

Based on chenzhu's transportation hobby we request the facilities around the subway to cluster it. Because he don't have a car and the only way he go out for date is just subway. And we cluster them into 5 clusters. Besides the other 2 clusters contain only one item so we just put them into choice for weekend. Because there are actually several big shopping malls there.

As for the second cluster(local style cluster.), chenzhu can choose when he want to have the chinese food for date and which is more convinient for him to get there.

As for the third cluster(tasty food in different countries cluster), chenzhu can choose from them when he want to taste something new and walk around for more. It must be more expensive and funny.

As for thr last cluster (choose it when you don't know what to choose cluster.), chenzhu can choose them when he don't have any strong feeling to some places.