Berlin Museums

Coursera Capstone Applied Data Science Specialization IBM

Aim

This project is aimed to help tourists who come to Berlin to find the Museums near to.

Problem Description

In Berlin there are many museums that you can visit and each one has special features and each Museum give you new experience. When a tourists come to Berlin he/she may not be able to decide which Museum he/she should visit first, especially if the time is limited.

Objectives

1- the geojson data for Berlin is required.

2- Then, analyzing the data using the Foursquare API.

3- Use clustering to identify each area and its category.

Data Source we are going to use is from Foursquare. Foursquare is a technology company that uses location intelligence to build meaningful consumer experiences and business solutions. So We are going to build a project with the help of Foursquare location data, Foursquare API provides great amount of quality data's about locations.(cafe, restaurant etc) Using this data will allow tourists to easily decide where to go when they are in a specific city. Using techniques such as K-means clustering, I was able to get results about common venues in city. This information can be really helpful to tourists since they can focus on what they are trying to experience most during their Travel. (Food, culture, sport etc) These techniques also provides the visualization of clustering of city. k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells

Dataset (From foursquare json)

	name	categories	address	cc	city	country	crossStreet	distance
0	Pergamonmuseum	History Museum	Am Kupfergraben 5	DE	Berlin	Deutschland	NaN	79
1	Museum für Islamische Kunst	Art Museum	Am Kupfergraben 5	DE	Berlin	Deutschland	NaN	58
2	Neues Museum	History Museum	Bodestr. 1-3	DE	Berlin	Deutschland	Museumsinsel	71
3	Alte Nationalgalerie	Art Museum	Bodestr. 1-3	DE	Berlin	Deutschland	NaN	75
4	Museum für Naturkunde	Science Museum	Invalidenstr. 43	DE	Berlin	Deutschland	NaN	1627
5	Deutsches Historisches Museum	History Museum	Unter den Linden 2	DE	Berlin	Deutschland	NaN	335
6	Berliner Unterwelten e.V.	History Museum	Brunnenstr. 105	DE	Berlin	Deutschland	NaN	3059
7	Dokumentationszentrum | Documentation Center (...	History Museum	Bernauer Str. 111	DE	Berlin	Deutschland	Ackerstr.	1711
8	Besucherzentrum | Visitor Center (Besucherzent...	Tourist Information Center	Bernauer Str. 119	DE	Berlin	Deutschland	Gartenstr.	1598
9	Berlinische Galerie	Art Museum	Alte Jakobstr. 124-128	DE	Berlin	Deutschland	NaN	1924
10	Deutsches Technikmuseum (Deutsches Technikmuse...	Science Museum	Trebbiner Str. 9	DE	Berlin	Deutschland	NaN	2852
11	Schloss Charlottenburg	Palace	Spandauer Damm 20	DE	Berlin	Deutschland	NaN	6901
12	DDR Museum	History Museum	Karl-Liebknecht-Str. 1	DE	Berlin	Deutschland	NaN	352
13	Martin-Gropius-Bau	Art Museum	Niederkirchnerstr.	DE	Berlin	Deutschland	NaN	1865

Data set processed for Clustering

	lat	long	Cluster Labels	Name
0	52.520843	13.396395	1	Pergamonmuseum
1	52.520709	13.396697	1	Museum für Islamische Kunst
2	52.520158	13.397838	1	Neues Museum
3	52.520796	13.398673	1	Alte Nationalgalerie
4	52.530271	13.379281	1	Museum für Naturkunde
5	52.517788	13.396948	1	Deutsches Historisches Museum
6	52.547745	13.388864	2	Berliner Unterwelten e.V.
7	52.535386	13.389668	1	Dokumentationszentrum | Documentation Center (...
8	52.533767	13.387485	1	Besucherzentrum | Visitor Center (Besucherzent...
9	52.503494	13.398403	1	Berlinische Galerie
10	52.498460	13.376881	1	Deutsches Technikmuseum (Deutsches Technikmuse...
11	52.520895	13.295667	1	Schloss Charlottenburg
12	52.519404	13.402239	2	DDR Museum
13	52.506996	13.381886	0	Martin-Gropius-Bau
14	52.519537	13.398803	1	Altes Museum
15	52.528119	13.372576	1	Hamburger Bahnhof – Museum für Gegenwart
16	52.506884	13.383601	1	Topographie des Terrors
17	52.506784	13.330789	1	C/O Berlin
18	52.514367	13.487327	0	Stasi-Museum
19	52.534896	13.390140	1	Gedenkstätte Berliner Mauer
20	52.501946	13.395241	2	Jüdisches Museum (Jüdisches Museum Berlin)

K-Means methodology

Algorithm

The *K*-means clustering algorithm uses iterative refinement to produce a final result. The algorithm inputs are the number of clusters *K* and the data set. The data set is a collection of features for each data point. The algorithms starts with initial estimates for the *K* centroids, which can either be randomly generated or randomly selected from the data set. The algorithm then iterates between two steps:

1. Data assignment step:

Each centroid defines one of the clusters. In this step, each data point is assigned to its nearest centroid, based on the squared Euclidean distance. More formally, if c_i is the collection of centroids in set C, then each data point x is assigned to a cluster based on

$$\underset{c_i \in C}{\operatorname{argmin}} \, dist(c_i, x)^2$$

where $dist(\cdot)$ is the standard (L_2) Euclidean distance. Let the set of data point assignments for each i^{th} cluster centroid be S_i.

2. Centroid update step:

In this step, the centroids are recomputed. This is done by taking the mean of all data points assigned to that centroid's cluster.

$$c_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i$$

The algorithm iterates between steps one and two until a stopping criteria is met (i.e., no data points change clusters, the sum of the distances is minimized, or some maximum number of iterations is reached).

This algorithm is guaranteed to converge to a result. The result may be a local optimum (i.e. not necessarily the best possible outcome), meaning that assessing more than one run of the algorithm with randomized starting centroids may give a better outcome.

Choosing K

The algorithm described above finds the clusters and data set labels for a particular pre-chosen K. To find the number of clusters in the data, the user needs to run the K-means clustering algorithm for a range of K values and compare the results. In general, there is no method for determining exact value of K, but an accurate estimate can be obtained using the following techniques.

One of the metrics that is commonly used to compare results across different values of *K* is the mean distance between data points and their cluster centroid. Since increasing the number of clusters will always reduce the distance to data points, increasing *K* will *always* decrease this metric, to the extreme of reaching zero when *K* is the same as the number of data points. Thus, this metric cannot be used as the sole target. Instead, mean distance to the centroid as a function of *K* is plotted and the "elbow point," where the rate of decrease sharply shifts, can be used to roughly determine *K*.

A number of other techniques exist for validating *K*, including cross-validation, information criteria, the information theoretic jump method, the silhouette method, and the G-means algorithm. In addition, monitoring the distribution of data points across groups provides insight into how the algorithm is splitting the data for each *K*.