Las Vegas Companion
Coursera Capstone
Applied Data Science Specialization IBM

Aim

This project is aimed to make your travel experience to Las Vegas Simple by providing all nearby locations of nightlife fun spots with help of foursquare datasets.

Problem Statement

Many people come to Las Vegas each day and experience difficulty in choosing between the best and the not so known night out places. Here I and other foursquare users would try and help them with their tips and ratings for each place and help people enjoy most of their time rather fear losing something in the fake casinos (casinos which cheat you,which can be easily seen by the ratings).

Objectives

1. The geojson data for Las Vegas's neighborhood 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	formattedAddress
0	Golden Nugget Hotel & Casino	Casino	129 Fremont St	US	Las Vegas	United States	at S Casino Center Blvd	494	[129 Fremont St (at S Casino Center Blvd), Las...
1	The D Las Vegas Casino Hotel	Casino	301 Fremont St	US	Las Vegas	United States	NaN	264	[301 Fremont St, Las Vegas, NV 89101, United S...
2	Fremont Hotel & Casino	Casino	200 Fremont St	US	Las Vegas	United States	btwn N Casino Center Blvd & N 3rd St	318	[200 Fremont St (btwn N Casino Center Blvd & N...
3	Four Queens Hotel & Casino	Casino	202 Fremont St	US	Las Vegas	United States	at Casino Center Blvd	378	[202 Fremont St (at Casino Center Blvd), Las V...
4	The Plaza Hotel & Casino	Casino	1 S Main St	US	Las Vegas	United States	at Fremont St	640	[1 S Main St (at Fremont St), Las Vegas, NV 89...
5	LONGBAR at the D Casino Hotel	Hotel Bar	301 Fremont St	US	Las Vegas	United States	NaN	276	[301 Fremont St, Las Vegas, NV 89101, United S...

Data set processed for Clustering

	lat	long	Cluster Labels	Name
0	36.170300	-115.145285	0	Golden Nugget Hotel & Casino
1	36.169655	-115.142732	2	The D Las Vegas Casino Hotel
2	36.170871	-115.143137	2	Fremont Hotel & Casino
3	36.170266	-115.143992	1	Four Queens Hotel & Casino
4	36.171398	-115.146687	2	The Plaza Hotel & Casino
5	36.169804	-115.142873	0	LONGBAR at the D Casino Hotel
6	36.171004	-115.146532	2	Golden Gate Hotel & Casino
7	36.171440	-115.145727	1	Las Vegas Club Hotel & Casino
8	36.168892	-115.139025	1	El Cortez Hotel & Casino
9	36.171239	-115.146849	1	William Hill Sports Book - The Plaza Hotel and...
10	36.170845	-115.143974	2	Sports Book at Fremont Casino

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 assigment 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

$$\underset{c_i \in C}{\arg\min}\ dist(c_i,\ x)^2$$

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).

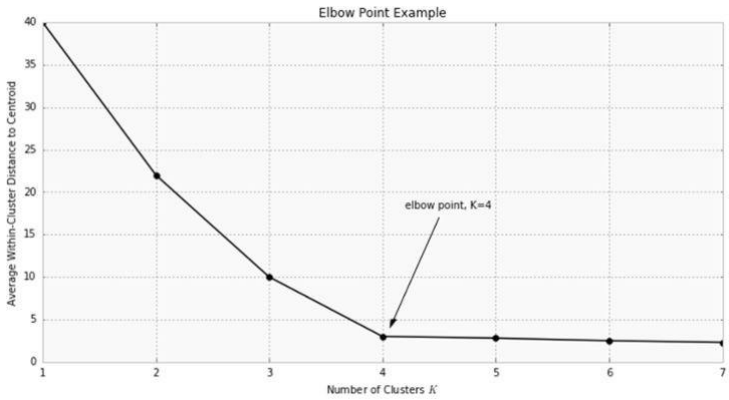
This algorithm is guaranteed to converge to a result. The result may be a local optimum (i.e. not necessarily the best possible outcome), meaning that assessing more than one run of the algorithm with randomized starting centroids may give a better outcome.

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*.

Results

The interesting and fun filled casinos based on their ratings have been clustered and plotted on a map

Conclusions

What we conclude is that mostly all casinos and fun filled places at Las Vegas are genuine and the ones which are not have been provided with bad ratings