A Venue-Type-Based City Recommender

Xinyue Luo

Mar 03 2019

1. Introduction

The business problem

Looking for a dream city to start a new life or a perfect place to spend your vacation can be a big commitment which usually begins with dedicated reading and research, and even in-person visits. This learning process could be tiring and time-consuming. Can we make this process a little less stressful by generating recommendations based on local information available online, along with customized user preferences of city features? With classic machine learning techniques such as classification, we can easily solve this problem.

Objectives

Primary objective

A K-Means classification model will be constructed which can be used to provide recommendations of cities that one would like to live in or visit. With a customized input of a user's current favorite city, the tool will return a list of cities included in the model that are similar to the user's city as a recommendation of cities the user would like.

Secondary objective

To further pinpoint which part of a potential city on the recommendation list the user would like, a second K-Mean classification model will be built to cluster venues in the city based on geographic location. Each cluster will be fed into the classification model trained in the previous step, and a list of clusters that fit the user's preference will be returned.

2. Data collection

In this project, two types of raw data were used: a list/pool of cities (to pick recommendation from) and existing venues in each city (to train the classification model).

For the list/pool of cities, a list of largest 24 cities by population on the US east coast was scraped from Wikipedia page https://en.wikipedia.org/wiki/Eastern_United_States. Since information on this page is not uniform, i.e., some city names include state while other don't, I further scraped state and area information from each city's main Wikipedia page. During the data cleaning process, city and state names were separated into two columns in a DataFrame, except for Washington, D.C. Area information was further extracted so that each city has area represented in square kilometers. An additional column 'Radius' represented in meter was calculated based on area, with the assumption that each city can be considered as a circle:
$\text{radius}_{(m)}$ = 1000 * square_root($\text{area}_{(km2)}$/pi).

Data of nearby venues (latitude, longitude, category) were obtained from FOURSQUARE API, by using each city's geographic coordinate as the search center, and the radius calculated above as the search radius. The frequencies of each venue type were calculated by one-hot encoding the 'Venue Category' column, summing up and normalizing, which is then used as features to fit a K-Means clustering model.

3. Methodology

3.1 Exploratory data analysis

For sanity check and to understand at a high level the diversity of venues in each city, the number of venue types in each city was counted and plotted in Fig.1. Generally speaking, the venue type diversities are similar among the 24 largest cities on the east coast, with Baltimore, Maryland being the highest (65) and Columbus, Ohio being the lowest (47).

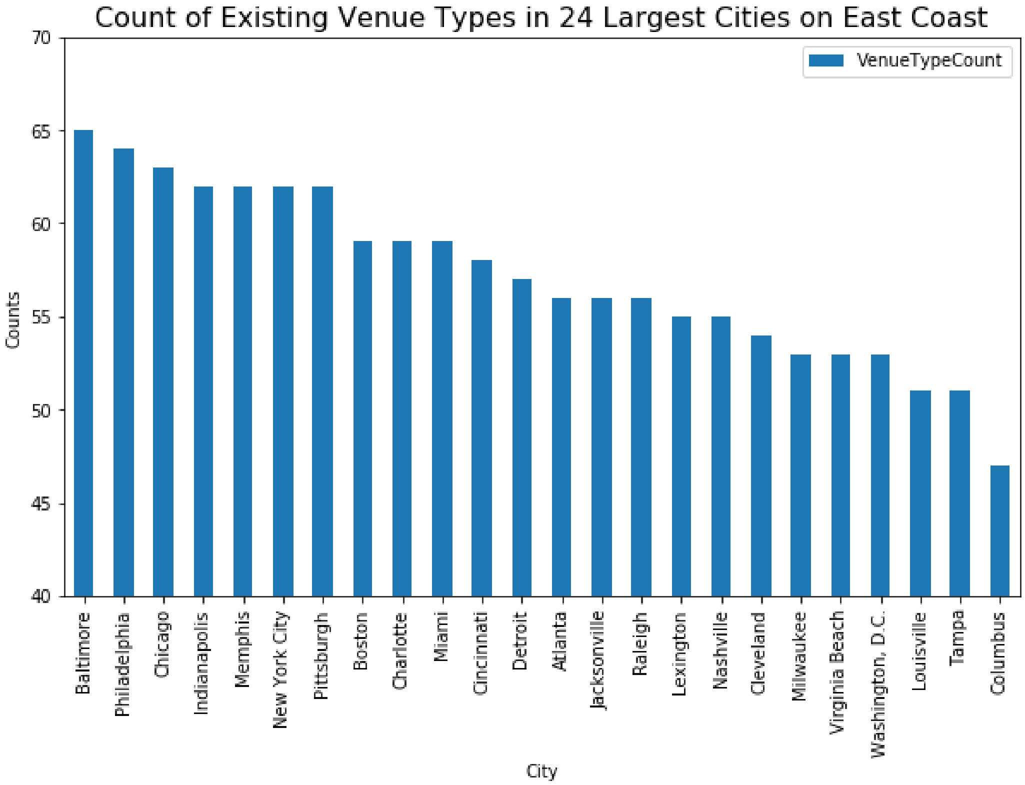
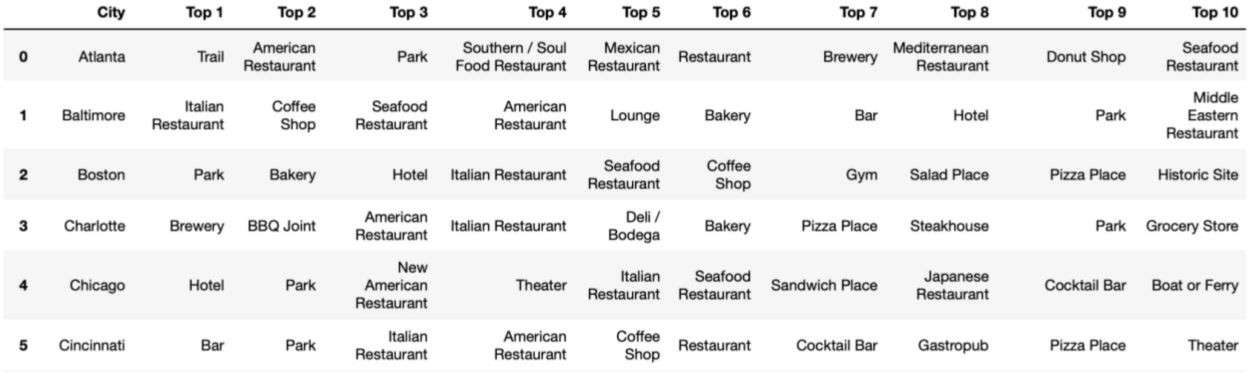


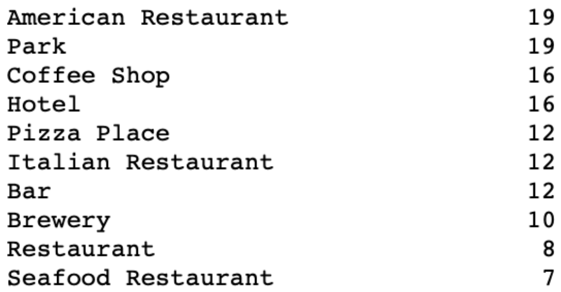
Figure 1. Count of venue types in 24 largest cities on US east coast

The top 10 venue types in each city were also selected (the first 5 are shown in Table 1). Although each city has its unique combination of top venue types, a frequency count on the whole DataFrame (Table 2) showed that 'American Restaurant' and 'Park' are the most popular venue types in all 24 cities – 19 out of 24 cities have these two on their top 10 venue type list.



	City	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6	Top 7	Top 8	Top 9	Top 10
0	Atlanta	Trail	American Restaurant	Park	Southern / Soul Food Restaurant	Mexican Restaurant	Restaurant	Brewery	Mediterranean Restaurant	Donut Shop	Seafood Restaurant
1	Baltimore	Italian Restaurant	Coffee Shop	Seafood Restaurant	American Restaurant	Lounge	Bakery	Bar	Hotel	Park	Middle Eastern Restaurant
2	Boston	Park	Bakery	Hotel	Italian Restaurant	Seafood Restaurant	Coffee Shop	Gym	Salad Place	Pizza Place	Historic Site
3	Charlotte	Brewery	BBQ Joint	American Restaurant	Italian Restaurant	Deli / Bodega	Bakery	Pizza Place	Steakhouse	Park	Grocery Store
4	Chicago	Hotel	Park	New American Restaurant	Theater	Italian Restaurant	Seafood Restaurant	Sandwich Place	Japanese Restaurant	Cocktail Bar	Boat or Ferry
5	Cincinnati	Bar	Park	Italian Restaurant	American Restaurant	Coffee Shop	Restaurant	Cocktail Bar	Gastropub	Pizza Place	Theater

Table 1. Top 10 venue types in the first 5 cities (by alphabet)



```
American Restaurant          19
Park                         19
Coffee Shop                  16
Hotel                        16
Pizza Place                  12
Italian Restaurant           12
Bar                          12
Brewery                      10
Restaurant                    8
Seafood Restaurant            7
```

Table 2. Top 10 most popular venue types in the 24 largest cities on the east coast

3.2 Modeling: K-Means classification

The normalized frequency counts of venue types were fit into the K-Means model as training features with K ranging from 1 to 12. The distortion of each fitting was calculated and plotted as a function of K. Using the elbow method, the K which has the biggest difference in the slopes at (K-1) and (K+1) was chosen as the optimal K value used to build the final model (Fig.2). Here optimal K is 8, which means the 24 cities are classified into 8 different clustered based on venue types (Fig.3). Interestingly, Washington D.C. is its own cluster with no other city out of the 24 being similar. Similar observations apply to Philadelphia, Atlanta and Virginia Beach. On the contrary, the biggest cluster contains 11 cities (Detroit, Cleveland, Columbus, Cincinnati, Louisville, Lexington, Milwaukee, Memphis, Raleigh, Jacksonville, Tampa). The other three clusters contain 2 (Nashville, Indianapolis), 3 (New York City, Boston, Chicago), 4 (Miami, Charlotte, Baltimore, Pittsburg) cities, respectively.

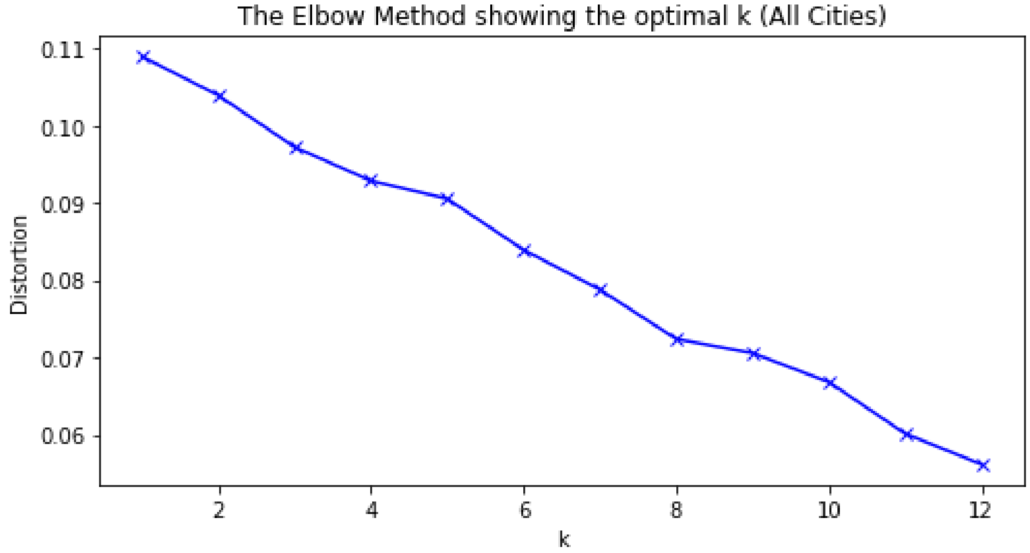


Figure 2. Using the elbow method to choose optimal K for K-Means model of 24 cities

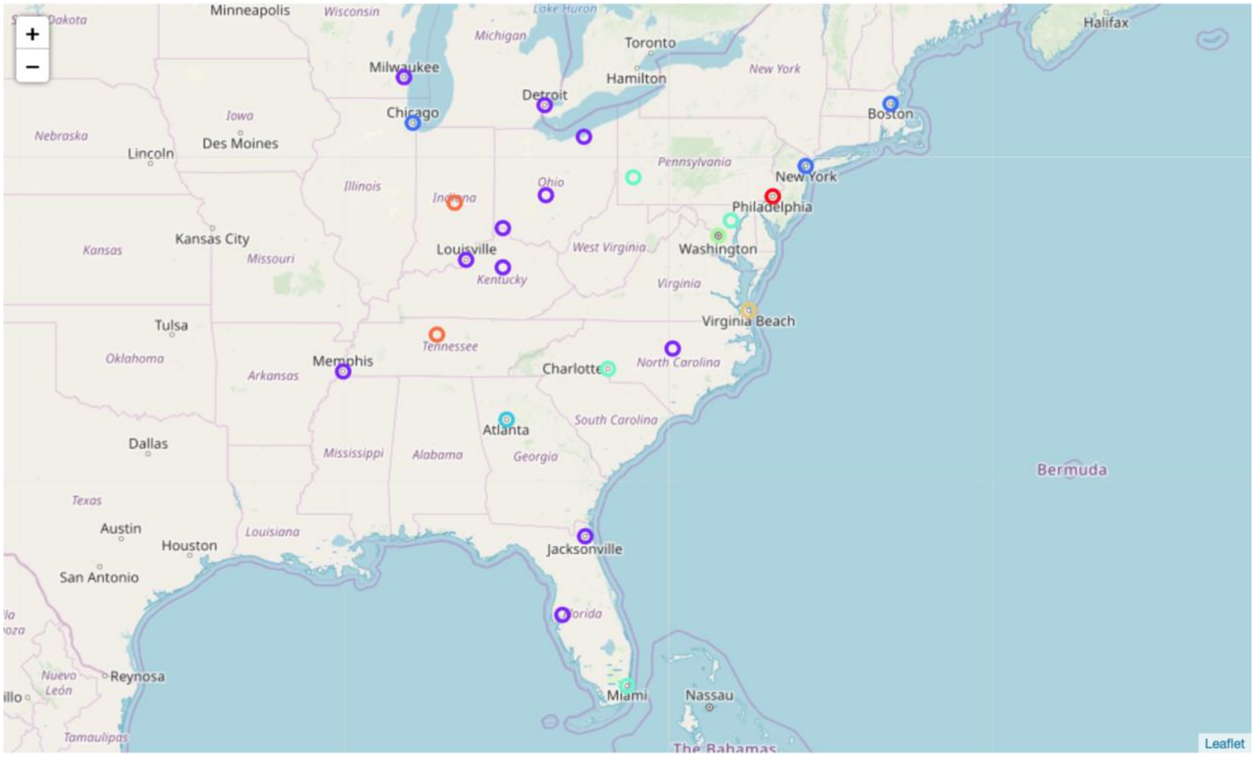
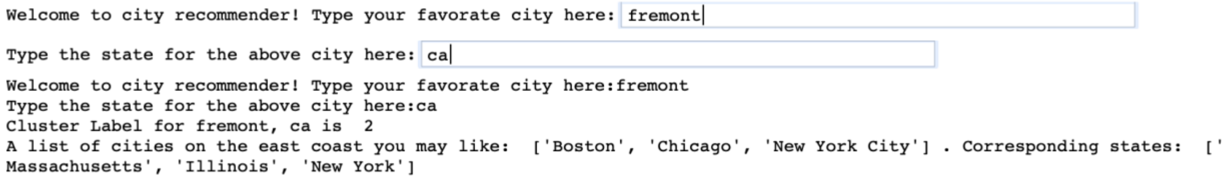


Figure 3. Map visualization of 24 cities labeled with cluster (by different colors)

4. Results

Using the trained K-Means model on the 24 cities, the tool was able to return a list of recommended cities once a user types in any city of choice. In the example shown in Fig.4, ‘Fremont’, ‘CA’ was a user input of a random city outside of the 24 cities list. The output returned ‘Boston’, ‘Chicago’ and ‘New York City’ as recommended cities, since ‘Fremont’ was predicted to belong to the same cluster as these three using the model.



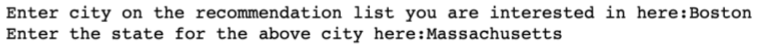
```
Welcome to city recommender! Type your favorate city here: fremont

Type the state for the above city here: ca

Welcome to city recommender! Type your favorate city here:fremont
Type the state for the above city here:ca
Cluster Label for fremont, ca is  2
A list of cities on the east coast you may like:  ['Boston', 'Chicago', 'New York City'] . Corresponding states:  ['
Massachusetts', 'Illinois', 'New York']
```

Figure 4. Example of using a customized input city to generate a city recommendation list

A second K-Means classification model was built to cluster venues in a city of choice from the recommendation list based on their geographic location. In the example shown in Fig.5 and Fig.6, ‘Boston’, ‘Massachusetts’ was typed in as a city of interest out of the recommendation list for ‘Fremont’, ‘CA’. Similar to fitting the K-Means model for the 24 cities, the elbow method was used to select the optimal K, which is 2. The venue type frequencies in each cluster was then counted and normalized to feed into the first K-Means model for the 24 cities for cluster prediction. As shown in Fig.7, both clusters are of the same cluster as ‘Boston’ as a whole as well as ‘Fremont’. This result suggests that the distribution of venue types in Boston is rather uniform, and that the user who likes Fremont is likely to enjoy visiting both clusters in Boston (considering local venue types).



```
Enter city on the recommendation list you are interested in here:Boston
Enter the state for the above city here:Massachusetts
```

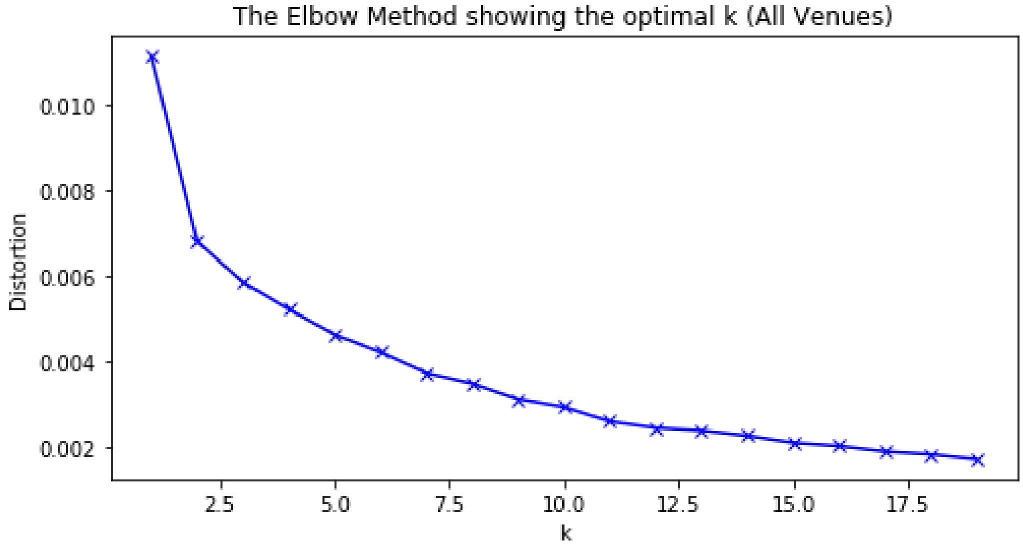


Figure 5. Using the elbow method to choose optimal K for K-Means model of venues in Boston

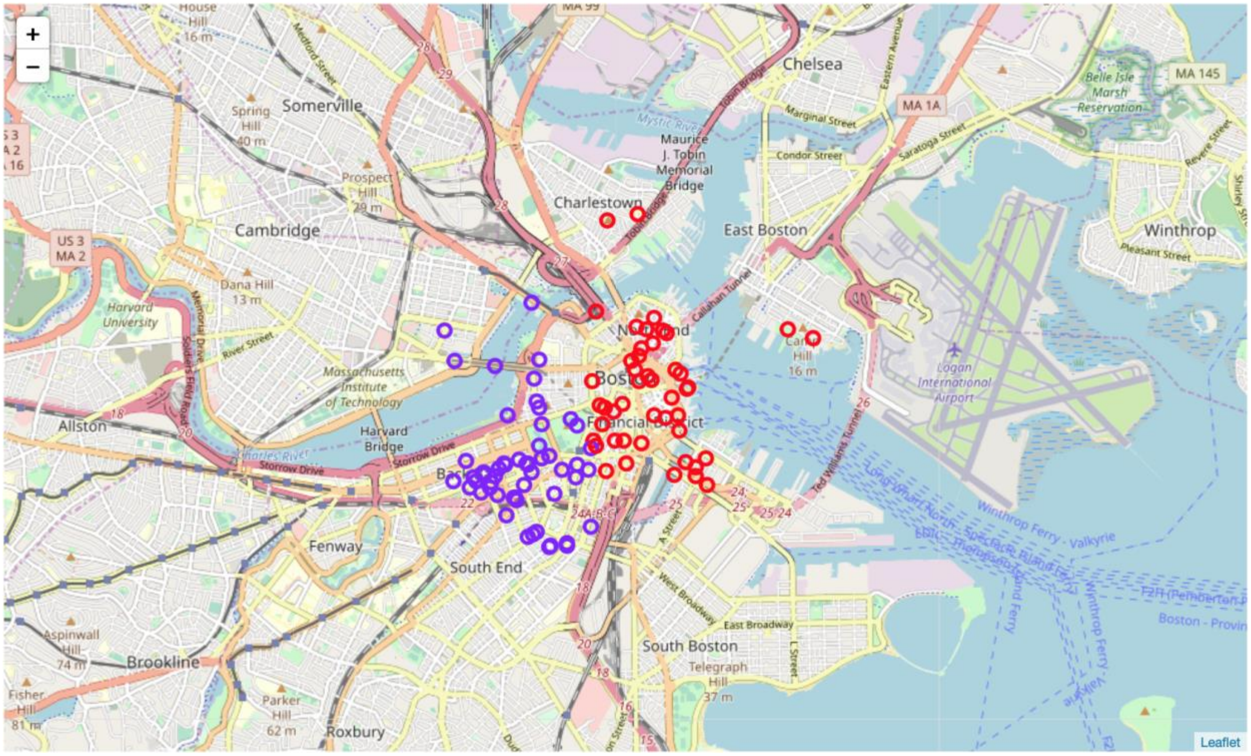
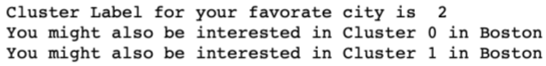


Figure 6. Map visualization of venues in Boston labeled with cluster (by different colors)



```
Cluster Label for your favorate city is  2
You might also be interested in Cluster 0 in Boston
You might also be interested in Cluster 1 in Boston
```

Figure 7. Output of classification of each of the venue clusters in Boston

5. Discussion

One major limitation of this tool is that only limited number of venues (100) were acquired for each city. This is due to the limitation FOURSQUARES API puts on the output of each inquiry. As easily seen in Fig.7, most of the acquired venues are around center Boston area. There are very few venues in boroughs like East Boston, South Boston. There is no venue acquired for boroughs even farther from the center area, such as Roxbury. This limitation would affect especially users who apply this tool to identify regions to live, rather than to visit for short-term, under the assumption that tourism is more popular in the center city area while residence is more popular in the suburb area. Nevertheless, an easy solution could be to initiate multiple inquiries for a single city, either by using a geographic JSON file defined for all boroughs of a city, or randomly picking a few additional geographic coordinates away from the center coordinate. Introducing boroughs can also be used to recommend a specific borough within the city that the user potentially would like.

Another future improvement could be to include other types of training features in additional to venue types, such as demographics, macroeconomics, housing price, employment types etc. On top of that, I would like to further build an interface where a user can select the type of features that matter to him/her for the initial model training as well as the prediction. For example, for a user who is more concerned about potential employers in a region rather than any other factors, the model can be trained and used to generate recommendation purely based on employment-related features. At the same time, the list/pool of candidate cities can also be expanded to include more cities rather than just the largest 24 cities on the east coast.

6. Conclusion

In this study, I applied K-Means model to classify the largest 24 cities by population on the US east coast into 8 different clusters, based on the frequency of different venue types in each city. Using the trained model, I built a tool that is able to generate a list of recommended cities out of the 24 that a user would like to visit, based on the user's input of a known favorite city. Furthermore, an extended part of the tool can group venues based on geographic distribution in a city of interest selected from the recommendation list, and analyze whether each of the clusters within the same city matches the user's preference. This tool can provide helpful insights for anyone who would like recommendations on cities to visit or live in.