Project Report

Members

Haotian Wu, **Desheng Zhang**, **Wenjun Shen**

Project Narrative

As one of the most famous city all over the world, Boston is a popular city for people to travel. People might use website like booking or tripdavisor to find out the suitable hotel from them. However, most of the grades and comments concentrates on the hotels themselve, like the quality of service, cleaningness, etc. Obviously, those grades and comments ignore the surroundings, like crime rates, transportations. Hence we try to evaluate the hotels of Boston by some other factors, like crime rate, transportation information, foods and gardens. Based on the evaluation, we aimed to find a new protential coordinate to build new hotel with high score, and we can also recommend hotels to customers based on their preferences. Besides that, we also set up an interactive web-based visualization. In the website, users can sort the factors with their own preference. Our backend will send the recommended hotels for users, pointed on a map.

Datasets

In this project, we use 5 original datasets. Our core dataset is the hotels of Bostons, which we scraped using BeautifulSoup with Python. Besides that, we have crime data, MBTA data, restaurant data and garden data of Boston.

1. **Hotels in Boston**
2. **Boston Crime Data**
3. **MBTA Data**
4. **Restaurants in Boston**
5. **Gardens of Boston**

Data Transformation

Orginal Dataset	Transformation Description	New Dataset
Hotels in Boston, Boston Crime Data, MBTA Data,Restaurants in Boston, Gardens of Boston	Start from each hotel, set a radius(ex. 0.5 miles), count number of crimes, mbta stops, restaurants and gardens	Boston Hotel Data
BostonHotelData	Apply our custom algorithm to give a new score to each hotel	BostonHotelCustomScore
BostonHotelCustomScore	Apply Correlation coefficient to figure out the most related factor to our custom score system	BostonHotelCorrelation
BostonHotelCustomScore, BostonHotel	Use K-means to cluster hotels based on new custom score and select best coordinates to build a new protential hotel	BostonHotelProtential
BostonHotelCustomScore, BostonHotel	run K-means on every possible permutation, then caculate the score and ranking to form a new dataset	BostonHotelPotentialPermutation

Custom Rating System

In our new scoring system, we calculate the number of gardens, crimes, MBTA stops, restaurants and cafes near each hotels(within certain distance). Then we use the normalize formula below to scale the original sorce and datas gathered together to calculater the new score.

Normalize Formula

$$S_{normalized} = \frac{S - S_{min}}{S_{max} - S_{min}}$$

Custom Score Formula

$$S_{custom} = \frac{(S_{origin} + S_{garden} + S_{food} + S_{mbta} + (1 - Scrime))}{5}$$

Correlation Coefficient

Formula:

$$corr(x,y) = \frac{cov(x,y)}{std(x) - std(y)}$$

Calculate the correlation coefficient between each factor and custom score of each hotel using the formula above.

_	Coefficient	p_value
Crime	-0.09448652255976357	0.3984638884028119
MBTA Stops	0.5511472545817303	8.065664653031119e-8
Gardens	0.8712154440427992	1.9353053027735632e-26
Foods	0.6383325374109514	1.1093179778057194e-10

Result

We find out that the coefficient between crime and custom hotel score is extremely close to zero, and the p-value is relatively large when comparing to other factors. Thus, we conclude the crime factor is not significant and no longer consider crime as a factor at next step.

Potential Best place to build a hotel

Based on the effort we made above, we try to find a best place to build to hotel. In the part, we first discard useless factors and recalculate the custom score of each hotel. Then we use the coordinates of the hotel and the new score to do K-means. We choose the cluster with highest average score, the calcuate the center coordinate of this cluster, which will be the best place to build a new hotel.

Factor Filter

Learnt from the correlation coefficient, we find out that our custom score is nearly not related to crime. Thus we

discard crime factor first.

New Custom Score

We discard the crime factor, thus the new score will be:

$$S_{custom} = \frac{S_{origin} + S_{garden} + S_{food} + S_{mbta})}{4}$$

Apply K-means

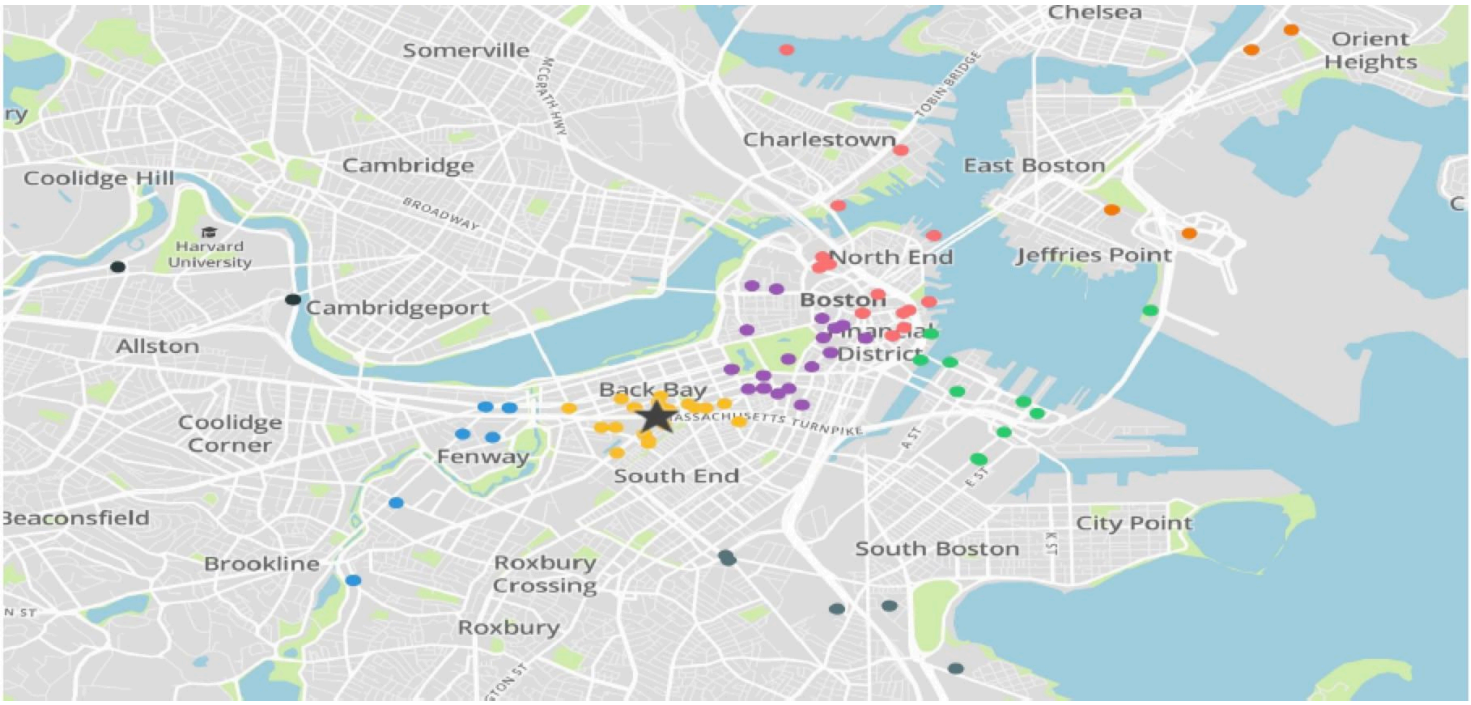
For K-means matrix, we use the coordinates and new custom rates of hotels. First we normalize the coordinates and rates separately. Then we slightly make coordinates with higer weights to make sure the clusters can be clusted based on there location first. We choose number of clusters as 10. For each calculated cluster. We select the cluster with higheset average custom rate and calculate the center coordinate of this cluster, which is [42.347708499999996 -71.0792716]

Protential hotel description

As mentioned above, we select the coordinate [42.347708499999996 -71.0792716] to be the protential best place to build a hotel. The approximate address of this coordinate is 111 Huntington Avenue, where is downtown Boston, really close to Prudential tower, Copley Place, Boston Public Library, etc. It is reasonable to believe the coordinate is a good place to build a new hotel.

Visualization

Based on the K-means algorithm and the calculation of the best place to build our new hotel. we use a library called “plotly” to visualize the clusters and the new place to build the hotel.



As the image shown above, for different cluster, we mark with different color. For the best place to build hotel, we mark it with a black star.

Interactive web-based Visualization

Besides the investigation of the potential best place to build a hotel in Boston, we realize that different people have different preference on hotels. Thus we set up a web-based application to recommend hotels to people with different preference. The basic idea is list out all four factors we include in our project, which are origin hotel score, food, mbta stops and gardens. We let users choose the orders with their own preference.

Application Architecture

Backend(Django)

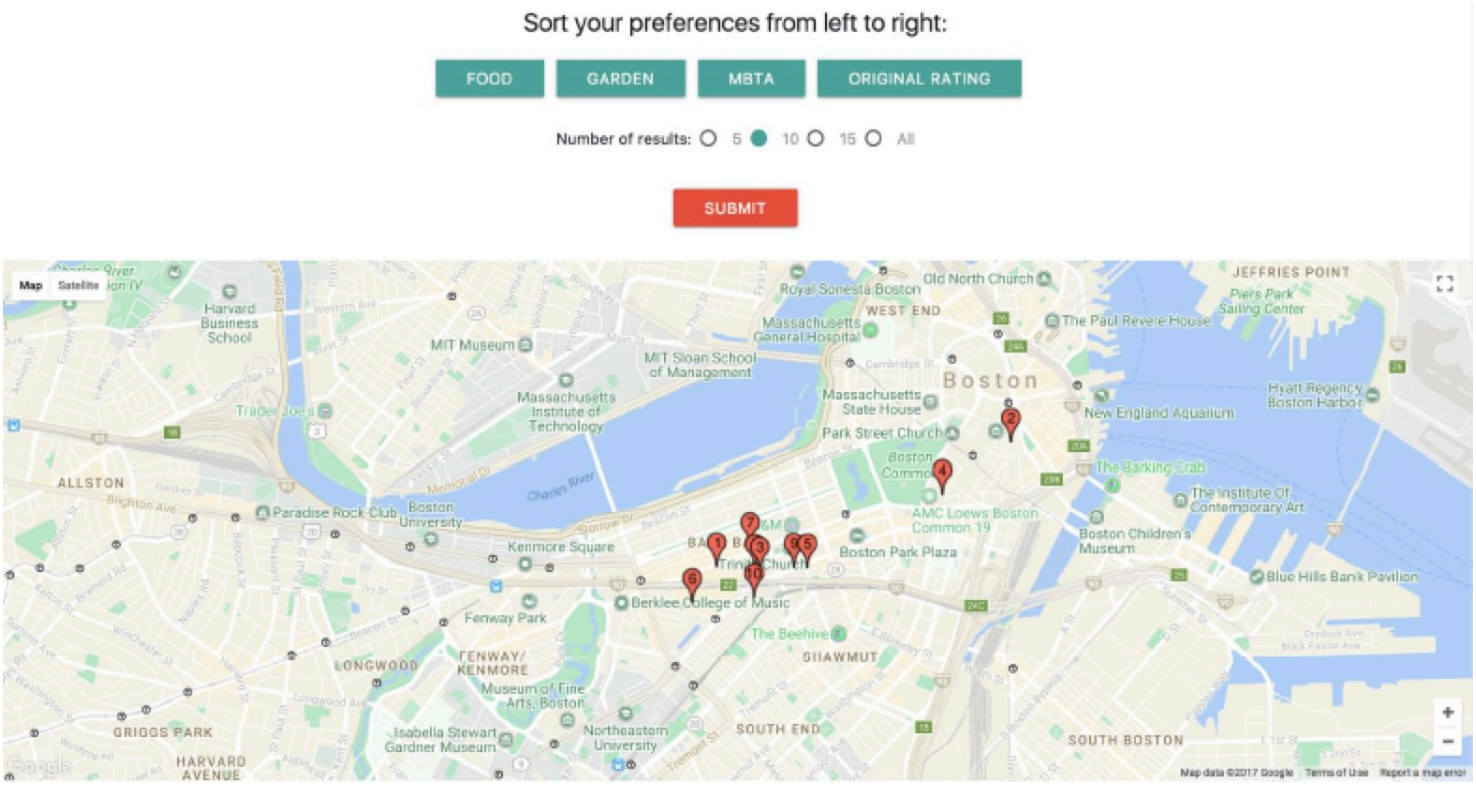
For backend, we use Django framework. Django is a free and open-source web framework, written in Python, which follows the model-view-template (MVT) architectural pattern. We use Django to receive the request from the frontend and interect with MongoDB to retreive the data. Then we send the data back to frontend with JsonResponse.

Frontend

For frontend, we use html and javascript. We let use choose the order of the factors to choose a hotel by their preference. After they submit their preference, the seleted hotels will be shown on a map. They can also

choose the number of hotels to show on the map as the hotels are sort by their scores based on our custom score system.

Visualization



Conclusion

Using the custom score system and K-means clustering, we successfully find out the best potential place to build a new hotel in Boston and make recommendation to users with different preference.

Future Work

- Since food establishment licenses dataset contains many small cafeterias and our project targets hotel customers, we can filter those cafeterias and make the food dataset more target specified.
- We can include more factors such as nearby attractions in the score system to make the score reflect the hotel's quality more accurately.
- When we visualize the result hotel, we only display the hotel name and the a pinned location on the map. we can show more information about the information including photos and external links to book the hotel.