BU CS506 Building Justice Maps
Final Report

Ziran Li
Boston University
zrli@bu.edu

Yize Liu
Boston University
lyz95222@bu.edu

Zisen Zhou
Boston University
Jason826@bu.edu

GitHub Link:
https://github.com/ZiranLi/Building-Justice-Maps

Abstract

In this report we will discuss about the main progress in our Building Justice Maps project. We first obtained all the data of four different categories (Education, Health, Finance and Community) by calling google map api and writing web scrapers. Then for some ambiguous data, we trained a Naïve-Bayes model to classify them. The result shows that the classification correction rate was below 50%. So to improve it, we write keyword matching program to make classification, which have over 60% correction rate for classification. Also we build maps for the distribution of different resources and user activities.

1. Project Description

Our project goal is to use the data from Google maps, UCB (Union Capital Boston)'s own internal data and 9 scraped websites' data to help UCB create maps for Education, Health, Finance, Community resources, prioritize resource need categories, build a powerful resource distribution map.

In data obtaining, we use web scraping and google map API. In data processing, we use K-means clustering, Naive Bayes classifier and keyword match. To evaluate their performance, we also calculate its Correction Rate.

Currently we have successfully obtained the data and classified them to different categories. Note that for some vague data that are not clearly categorized, we tried k-means clustering algorithm to cluster them into the target subcategories. To improve the result of categorization, we use Naive-Bayes classifier to classify all the subcategories into four main categories (Education, Health, Finance, Community). However, the Correction Rate for this model is below 0.5, which is not accurate enough. So we write program for keyword matching.

Finally, we built maps to project the resource data with user data, to show the distribution of resources on user needs.

Questions to be answered:

- ✓ What resources do union capital members need the most (based on analysis of member engagement records?
- ✓ What resources are available?
- ✓ Where are the resource gaps (resources vs. needs of Union Capital Members)?
- ✓ How to prioritize the importance in each resource?
- ✓ How to link the dataset with different categories? How to classify these data?
- ✓ How far are people going to get to resources? Are there resources that are closer?

2. Data Description

As our project is to build the map, the main part is to collect data. We have three main sources for us to obtain the data. The websites provided by the UCB, the transaction files from UCB and the google map API.

UCB provide us some csv files containing the transactions from Civic Engagement App and files contain the user location and event locations. The file contains the time of the events, location of the events, categories of the events and description of the events. However, the category in the file is really mess. The detail will be mentioned in the data analysis part.

We look up the four main categories provided by UCB. For each subcategory under main categories, we go to the corresponding websites provided and web-scrapes the corresponding data in Boston area. Due to the different structure of different websites, we rewrote the program for each website. The data we scraped include the name of the resource, the address of the resource, the contact and the open hour of the resources. We scribed ten websites in total, including boston government websites for food pantry resources and government departments, public library websites for homework help and adult education resources, social security offices, tax offices, community center, etc.

Besides, we use the Google Map API to collect data. We use the place search API in google map to help us search all the events around the Boston. We set the searching radius to 50km so that we could search all the events happening in the

whole great Boston area. Every time we search through the Google Map API, we use the type value that is supported by the API to filter the searching places. Thus, we could better classify the data we searched. As a result, we called the API for each searching type independently (about 100 times) and later define the categories and combine all the data together to merge into an integrated csv file. Finally, we classify the categories into the four main categories and the required subcategories by analyzing the origin types and descriptions of each places.

3. Data Analysis

The main data we can use to analysis the behavior of UCB's user is the transaction file UCB provided. So to have a better understanding on the data, we have to explore it. There are 24 columns in the original file. After look into it, we keep 14 useful columns and delete the not approved data. The useful information includes event address, event name, event time, category, etc. Then we count the occurrences of each category.

```
Walking/In-home Exercise: 6948
OTHER: Type in Description: 6617
Adult Education Class: 4330
Community Meeting: 4100
Reading/In-home learning with child: 3577
Health (Physical & Mental): 3151
Education (Child/Adult): 2503
Community & Service: 2441
Gym/Fitness Center Exercise: 2213
Health Center Appointment: 2156
Volunteer: Helping Others: 2121
Hospital Visit: 1762
Library/Education Center: 1546
Performance/Festival: 1456
Volunteer: Organization: 1262
Fitness Class: 1061
Donating clothing/goods: 1048
Workshop/Info Session: 1004
Chaperone Field Trip/Sport Activity: 930
In-school Meeting: 927
In-school Event: 871
Volunteer: Event: 858
Cooking for an Event: 722
Farmers Market: 650
Health Workshop: 564
Early Childhood Playgroup: 520
In-school Volunteer: 502
Leading/Teaching an Event/Workshop: 405
Advocacy Event/Rally: 366
Political Activity: 257
Financial Advisor Meeting: 160
In-school Meeting: 927
In-school Event: 871
Volunteer: Event: 858
Cooking for an Event: 722
Farmers Market: 650
Health Workshop: 564
Early Childhood Playgroup: 520
In-school Volunteer: 502
Leading/Teaching an Event/Workshop: 405
Advocacy Event/Rally: 366
Political Activity: 257
Financial Advisor Meeting: 160
Financial Workshop: 156
FII Monthly Meeting: 127
Voter Registration/Engagement: 123
Opening New Bank Account: 111
Finances/Employment: 87
EBNHC - Asthma Management: 46
Home Buying Workshop/Class: 40
Tax Services: 37
Servicio comunitario: 13
La salud (f_sica y mental): 13
Educaci_n (niio / adulto): 12
Lending Circles Meeting: 7
Finanzas / Empleo: 1
```

Figure 1: Category Occurrences

In Figure 1, the second largest occurrence of category is "other", so we considering it is necessary to decide which type it supposed to be based on the description. However, when analyze the description in type other, there are a lot of description wrote in Spanish which made it difficult for us to extract features from the text. Besides, there are time of the events which make us consider the relationship between the time of the events. As for the event name and the event description, some of the description is the same as the event name, which makes the description useless. Because of the quality of the description, it is difficult to choose the train set and test set for classification.

Moreover, after the discussion with our contactor from UCB, we found that the location of the event may not be accurate. With the situation that some user may submit the information when they are back home rather than at the location of the event, the geolocation of the event cannot show the right location. To clarify the wrong location, we obtain the address of the user, change the address to geocode and eliminate the event location which is less than 100 meter away from the user address. By eliminate the user location, we have the better event location for the future use.

4. Algorithms

4.1. Data obtaining

For web scrapers, we followed BU cs506 homework 3. We parse the HTML code provided by different websites and extract the classes we need. For some websites, we extract the subdomain urls and reprocess the step mentioned ahead. And this step may repeat many times. Finally, we get desired information and combine them into CSV files.

We also use the Google map api to grasp data of events for each supporting type. Since the total amount of types are more than a hundred, it is difficult for use to manually change the type each time we run the api and write all the types into all program by ourselves. Thus, we firstly write a web scraper to scrap all the types supported by google map place search api in the official document website of google map and store all the types into a list. Then, we iterate all the types in the list and download the information of places respectively. Finally, we classify the data by its types and description.

4.2. Data preprocessing and parsing

To better classify and analyze the data, we first parse all the data in UCB events. We remove the invalid data which lacks the essential information, such as locations, users' information. Also, we extract the features form the description of each data for later classification. We first use the natural language toolkit package(nltk) in python to parse all invalid word (like nonsense symbols, punctuations and spanish words) and lemmatize all the words to the normal form. Then, we extract the features from the description words we have parsed by using Tf-Idf to transform the words into vectors, which represents the frequency of each word appears in each description and could be later used in our cluster and classification. The result is shown as below:

```
[[ 0.         0.         0.         ...,  0.         0.1553171  0.        ]
 [ 0.         0.         0.         ...,  0.         0.         0.        ]
 [ 0.         0.         0.         ...,  0.         0.         0.        ]
 ...,
 [ 0.         0.         0.         ...,  0.         0.         0.        ]
 [ 0.05499061  0.         0.         ...,  0.         0.         0.        ]
 [ 0.16153968  0.         0.         ...,  0.         0.         0.        ]]
{'natural': 9242, 'ronke': 11359, 'cerebral': 3304, 'presentar': 10401, 'hiset': 6833, 'cartoon': 3192, 'assistive': 2194, '30a': 566, 'ldbpeaceinstitute': 8062, 'parkman': 9841, 'essaibi': 5435, 'anytime': 1995, 'navigator': 9252, 'mifflin': 8899, 'entrance': 5359, '30th': 570, 'disaster': 4747, 'brasil': 2868, 'disciplined': 4755, 'codmancouncil': 3637, 'hand': 6606, 'resourves': 11148, 'cancer': 3099, 'pnf': 10198, 'octubre': 9533, 'dance': 4357, 'negociacion': 9287, 'viet': 13542, 'gibson': 6274, 'indicators': 7203, 'englandfor': 5320, 'haitiams': 6587, 'ruff': 11424, 'mula n': 9139, '_join': 1421, 'humanity': 6994, 'entre': 5360, 'remaing': 11044, 'healey': 6704, 'minimal': 8934, 'socieda dlatina': 12063, 'fiibostonwinterfest': 5849, 'celebra': 3270, 'braden': 2845, 'haiti': 6584, 'kimodoml': 7850, 'alle ghany': 1832, 'dreaming': 4967, 'accomodation': 1545, 'conditioning': 3857, 'race': 10731, 'informacion': 7256, 'clar ke': 3539, 'tavares': 12676, 'angle': 1951, 'opportunitiesto': 9613, 'entradas': 5358, 'lkrov': 8278, 'disadvantage d': 4744, 'brooklineteencenter': 2927, '6176202497': 972, 'seek': 11675, '6376': 1006, 'ubicacion': 13227, 'showing': 11883, 'directly': 4730, 'aromatherapy': 2105, 'song': 12107, 'thank': 12804, 'mandell': 8515, 'streets': 12378, 'men torship': 8831, 'ruffen82': 11426, 'constant': 3938, 'setting': 11771, 'professional': 10518, 'strange': 12364, 'work
```

Figure 2: Feature vectors extracted from the description (The 2-D lists shown above represents the extracted vector features from the description, the dictionary printed shows the word represented in each index)

4.3. Data classification

4.3.1 Classification criteria

We are required to classify all the data into four main categories which are Education, Health, Finance and Community. For each main category, we need to further classify the data into required sub-categories, which are Education (School, Class, Training, Read, Tutor, Program), Health (Gym, Therapy, Class, Counseling, Fitness, Appointment, Fruits, Vegetables), Finance (Home, Job, Tax, Budget, Financial, Bank, Savings, Employment, Entrepreneurship, Debt) and Community (Volunteer, Conference, Meeting, Program, Discussion, Church).

4.3.2 Training data obtaining

To make classification of the data, we need training dataset to train our machine learning algorithm. However, since USC does not provide us data that have already been classified into required categories, we decide to build the training and test dataset by ourselves. To achieve this, we first tried to match the keyword for each data. we match the keywords in the original categories and in the description. Since all the data contains both category and description. We first match the original category with each subcategory, if we could not find the match, we then try to match the description with each subcategory by searching the keyword by matching regular expression. By going through this whole process, we have classified part of the data, the category of which is apparent to match, into the required main category and sub category. As for the left data, we manually set the categories after discussing with our contactor. Finally, we classify about 8000 rows of data in this way. After checked by our UCB contactor, the classification result fits for their requirement so that we use these data as our training and testing dataset.

4.3.3 Clustering by K-means

As discussed above, the type "other" takes a huge count of the total category. Thus, we first tried to use the Tf-Idf to extract features from the description and then use the K-means to cluster the features. After the testing, we found the result is not satisfied. As can be seen from the figure from the result, the result of the cluster cannot be categorized. The poor performance of the cluster may cause by the features Tf-Idf extracted is not accurate enough. Besides, cluster may not be a good way to categorize the label.

4.3.4 Classification

Later, we use different supervised classification algorithms to make different classifiers. Since we have already obtained 8000 rows of classified data, we use the first 1000 data as the testing dataset and the left as training dataset. We train and test the performance of five different classification algorithms, which are logistic regression, k neighbors, decision tree, random forests and linear SVM. The results are shown below in experimental results. And after comparing the accuracy of each classifier, we choose to use the SVM classifier to classify the left data because it shows the highest accuracy and performs better in our relative small dataset.

5. Experimental Results

5.1. K-means result

```
['en', 'meeting', 'school', 'food', 'community', 'class']
['kids', 'story', 'time', 'took', 'park', 'program']
['boston', 'family', 'night', 'carol', 'day', 'ucb']
['class', 'group', 'program', 'survey', 'autism', 'home']
['class', 'program', 'kids', 'story', 'time', 'autism']
['class', 'family', 'boston', 'program', 'home', 'autism']
```

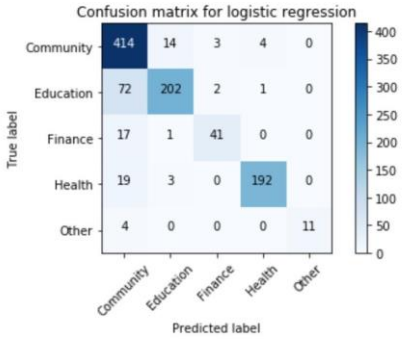
Figure 3: K-means result

Figure 2 shows the result of clustering when setting k=6. It is obvious in figure 2 that cluster 4-6 are much similar. Besides, the result in the first cluster, "en" is not a English word which is affected by the Spanish in the description. Due to the poor quality of the cluster result, we considering to use other ways of categorize the events.

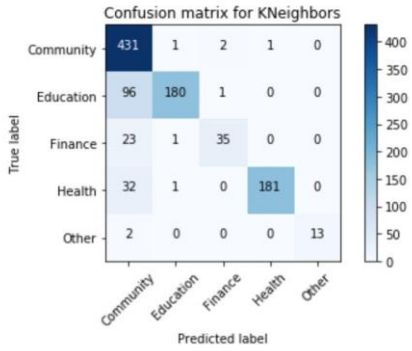
5.2. Classification result

We test the performance for each classifier and make the confusion matrix to display the results.

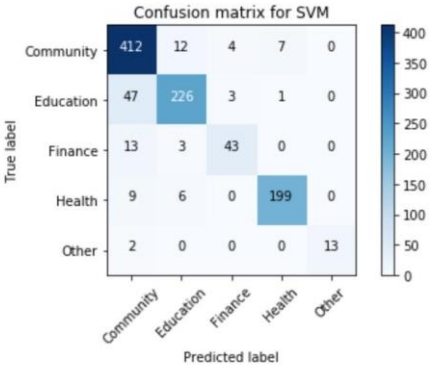
```
correct rate for logistic regression:0.86
```



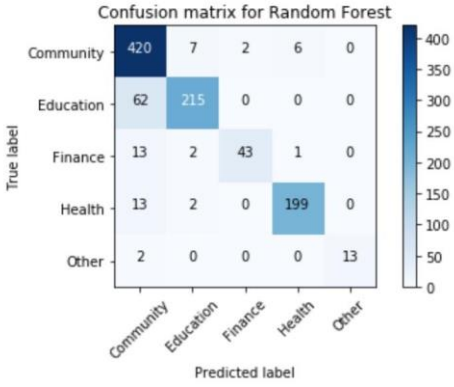
correct rate for KNeighbors:0.84



correct rate for SVM:0.893



correct rate for Random Forest:0.89



correct rate for Decision Tree:0.873

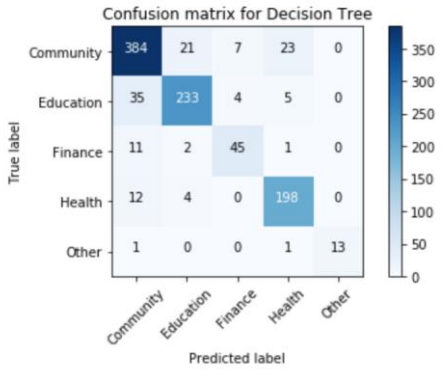


Figure 4: Confusion matrix for different classifier

From the confusion matrix of each algorithm shown above. The SVM shows the highest correct rate. And, the K-neighbors works the best in classifying the Education, Finance, Health and Other labels, however, it works not well when classifying the Community events. Finally, after comparison, we choose the SVM as our classification algorithm.

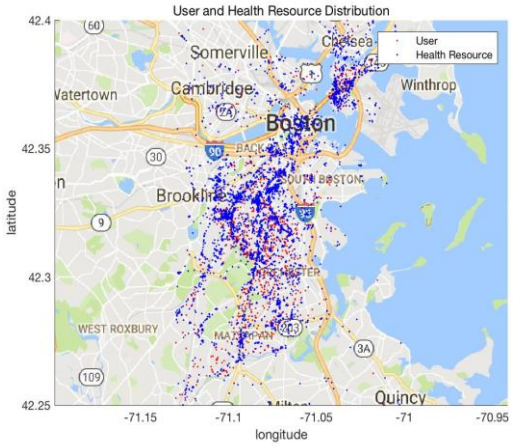
Besides training the model for classifying the five main categories, we also train the model for classifying the sub categories. Since the number of sub-categories is large and the description we have for each data is not enough for us to clearly differentiate each sub-category, the accuracy of the predicating result for each classifier is not high enough. Besides, the meaning of classifying sub-categories is not as important as differentiating main categories. Therefore, we do not use further use machine learning algorithm to further classify the sub-categories. For the sub-categories classification, we just use keyword matching as well as manually work as we have done with the training dataset before.

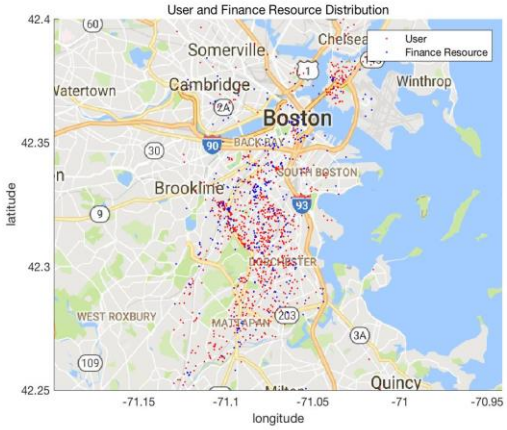
Finally, from the confusion matrix, we could find that for each classifier, differentiating between Community and Education is the most difficult task since the number of items that are misclassified between Community and Education are the most of among all misclassified items. This may be due to the reason that the community and educations events shows some common features which are hard to differentiate. To improve the accuracy of the classifier later, maybe the description of each event should be more clear and special and we need to parse some common words both in these two classes and use the other words to differentiate them.

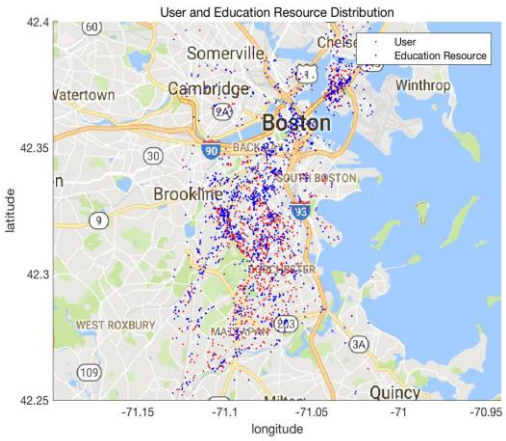
5.3. Final result

By comparing the both csv files for users and trans events. We first acquire the geolocation of the user by using google map api and then filter the original trans.csv data. After filtering, we parse almost half of the events and generate the

new csv file with the required main category and sub-category with about 28000 rows of data. Also, after the classification of the original UCBevents.csv file, we also generate a new modified csv file with about 8000 lines of events. Moreover, by scrapping from the websites and the google map api, we acquire about 3000 lines of events and making them into the right categories. All those data collected above will help UCB for later creating the resource map. Also, to better visualize the collected data and their distribution, we render all the events on the map and the results are shown in figure 5 below:







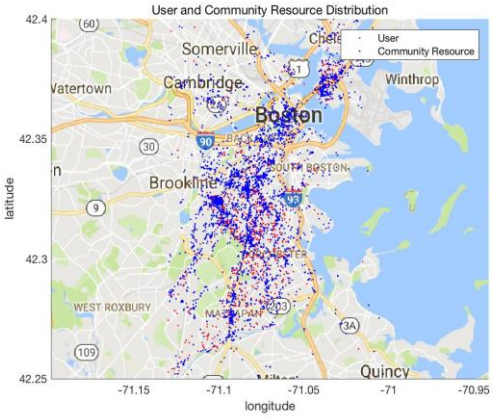
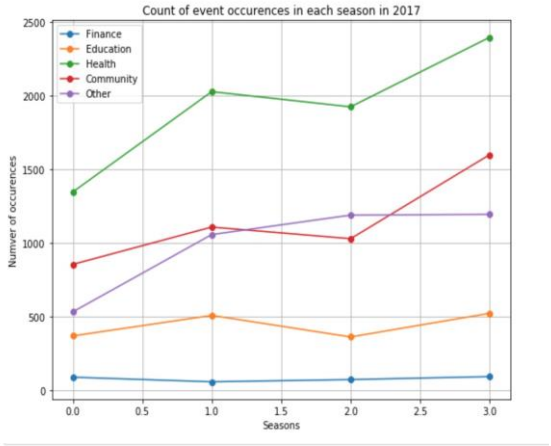
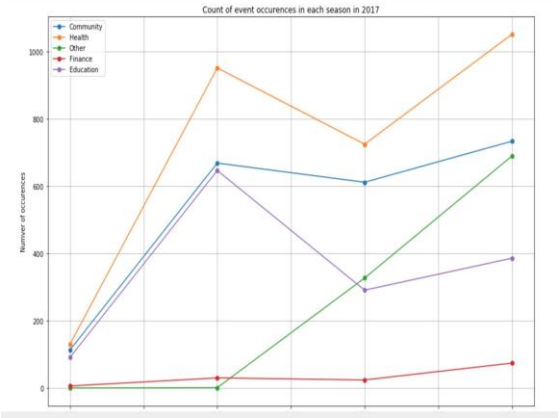
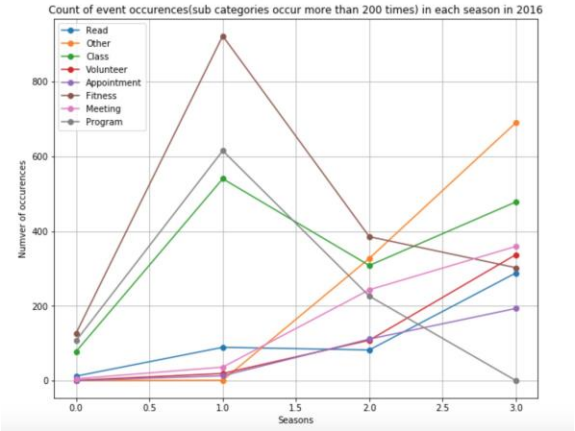


Figure 5: Different Resource vs User activities map

In the last report, we analyze the relationship between the occurrence times of the events and the seasons. However, since the data we use is the trans.csv file, it is not totally correct at that time and the categories we define at that time is not actually what we finally classify, the analyzing result is not quite convincing. Therefore, we re-render the figure to show the relationship of the number of occurrences of each main-category and sub-category with the time of the year in 2016 and 2017 respectively and use the atplot in python to plot the chart. The results are shown in figure 6:







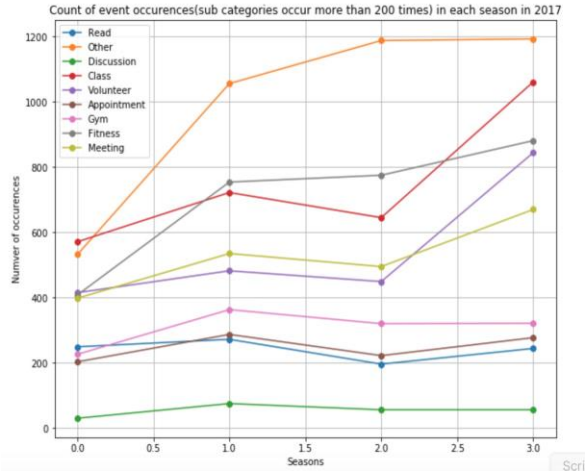


Figure 6: Occurrence Tendency

6. Conclusions

✓ What resources do union capital members need the most (based on analysis of member engagement records)?

The resources union capital members need the most is Health related resources. Based on the analysis on the transaction of 2017, the health events are attended the most. With the categorize the description into different sub-category, we had a better understanding in the records. To be more specific, despite the influence of the "other" type, the "class" in Education and Health and "Fitness" in Health are the most attended events. However, we think it is not enough to get the most need resources just from the data we have. The resources with low occurrence doesn't mean user don't need them. There is a chance that the limited resources lead to the low occurrence. More data are needed to have a better conclusion on this question. This is also discussed in the future work in part 7.

✓ What resources are available?

Generally, we did three parts to get the resources. Initially, we received three csv files from UCB, which contained user transaction events in 2016 and 2017. Meanwhile, we worked on the data from Google Map API. We search all the events happening around the great Boston area and we have collected more than 2000 rows of data, which could be categorized into our four main categories for later analysis. Also, we write 10 web scrapers, analyzing data from different websites (including state government websites/ public libraries websites/ national services, etc.). Based on

our analysis which is shown in Fig 5, the resources on most category are adequate for users.

✓ Where are the resource gaps (resources vs. needs of Union Capital Members)?

The quantity of the user is not enough for us to get a clear conclusion on the source gap. However, based on our analysis on the transaction and as can be seen from the Event map (Fig 5) shown above, we could conclude that we still lack enough Finance resources when compared to the data of Health, Community and Education. The blue dots in the Fig 5 are the resources location and red dots are the location of users. As number of resources in other type are even more than the number of users, as can be seen, the figures are cover in blue. However, finance is in between. The finance resources in Boston city center seems enough and the resources in Dorchester seems insufficient especially compare with the number of users.

✓ How to prioritize the importance in each resource?

As for prioritizing the importance in each resource. We count the users visiting times of each event. Also, since some of the event or resources may count as two or more main categories or sub-categories, we have used to make classification (For example, some class for gym could be both counted as resource for Health and Education), we merge all the tags appearing in each event and count the total occurrence times. If more users visit the same event and the event could be classified into more categories, the event owes higher priority. Based on our analysis, the place that have highest priority would be community center. Normally, community center can have resources in Education and Health that are also the resources that UCB members need most.

✓ How to link the dataset with different categories? How to classify these data?

We first create the training dataset by the given ucb_Events.csv. We parse the data and use keyword matching to classify the data into required classes and classify the rest data manually and the result fits well with the expectation. By doing this, we obtain a 8000 rows dataset which could later be use as our training and testing dataset. Then, we lemmatize the words in description of each event by using nltk and extract the features from the modified description by Tf-idf way. Then, we choose the first 1000 rows of data as the testing data and left 7000 data for training. We have trained and tested the performances of five different supervised learning algorithm, which are logistic regression, k neighbors, decision tree, random forests and linear SVM and the unsupervised algorithm K-means. After comparing the performances of each way, we finally choose the SVM, which shows the highest accuracy, to make further classification of other data we acquired from the UCB, google map api and web scratching.

✓ How far are people going to get to resources? Are there resources that are closer?

We have displayed the location of the resources on the map so that people could easily view the distance between their current location and the resource. And as can be seen from the Figure 5, in the center of Boston, the location of the resources and the users are near enough. However, in the finance resources figure, the finance resources seems are enough for users compare with other type of resources. Besides, the events are UCB's original resources, with the resources we collected, it would be easier for people to get to resources.

7. Future Steps

From our last version of the result, we have also implemented other ways to do the classification with better results. Besides what we can do to have a better result, it would be better if the UCB can have better data set. Based on our result, the class and fitness are the resources that been used most. However, the transaction can be hard to reflect the need of frequency. The frequency of attending class and fitness exercise is easy to be more than attending other events. More data are needed to have better result on resources analysis. By collecting more data, the result classification and prediction of can be more accurate.

To predict the future resources, we managed to analysis the activity in different seasons. More factors can be considered to analysis the user activity. For example, the population, the average income or the average age in different city can result in different activity preference. With limited data, those factors are hard to consider. But it definite worth more discovery and discussion.

Moreover, we cannot just focus on the type of activity that most user attend. We need to consider the activity with low occurrence and try to find the reason of its low participation. It would be a problem that UCB need focus to solve, if the reason is the resources are not enough. A survey after user upload their activity would be a good start to find the reasons.

8. Data sources

google search API:
https://developers.google.com/places/web-service/search?hl=zh-en
Other websites we scripted are shown in our GitHub:
https://github.com/ZiranLi/Building-Justice-Maps