YELP DATASET VISUALIZATION

ABSTRACT

We designed different visualization tools for the user and the business. For the business, we built a heat map to monitor the flow of customers and built a review river to see the store. For users, we constructed maps and parallel coordinates to help users make rough choices and constructed positive and negative word clouds and radar charts to help users compare in detail.
Project web link https://liyuan5005.github.io.

1 INTRODUCTION

Today, many people tend to use customer review websites, such as Yelp, Google business review and Tripadvisor, to decide where to go and eat. With the emergence of restaurant review websites, people can easily search and get information about price, location, and foods, and check other visitors' reviews. Specially, consumers who leave business reviews on Yelp have become known as "yelpers". Yelpers write their reviews about the business to help others make purchasing decisions, and it becomes to be a really important factor before the decisions. In this project, we will visualize a meaningful yelp dataset that could help business owners and customers. With our visualization, it is expected that business owners could gain meaningful business insights from customers' reviews, ratings, and visit patterns. Furthermore, it is expected to help the customers to decide which restaurant matches their tastes. First, the paper will explain the characteristics of datasets and the methods of data processing. Second, it will show the result and analysis of data visualization about the business-aspect and customer-aspect respectively. The project is implemented on HTML using Plotly, Python, and Javascript.

2 DATASET

This dataset is a subset of Yelp's businesses, reviews, and user data with 8GB. Yelp is a website and mobile app that connects people with great local businesses. In the dataset, there is information about businesses across 5,200,000 user reviews,174,000 businesses and 11 metropolitan areas in 4 countries. The dataset includes user's information, user's review to the local businesses, the business's information just like location, name, star and something else.

Data Sources	Columns
business.json	business_id, latitude, longitude, stars, review_count, categories
checkin.json	business_id, date
review.json	user_id, review_count, text
tip.json	user_id,business_id,text,data
user.json	user_id, review_count, friends, average_stars, compliment

Figure 1: Yelp Dataset

3 BUSINESS ANALYSIS

3.1 TASK

For business, when owners are faced with a large amount of user data, the problem often faced is that it is difficult to monitor the operation situation of his/her business.

Task1, how to monitor the users' check-in in a year trend. So we want to build visualization tools to help owners. It can intuitively visualize the number of check-in in one day of a year and show the difference, which will help them to make an adjustment.

Task2, how to monitor the customers' stars and comments text in 1.5 years. So we want to build visualization tools to help owners. It can intuitively visualize the average of stars, a number of different stars and the main comments from customers every 15 days of 1.5 years and show the difference, which will help them to make an adjustment.

Therefore, we decided to visualize the user data through 1 interactive visualization to help owners to get more details about his/her business.

1. Calendar heatmap shows the users' check-in in a year, marked with different color which represents different crowd degrees.

2. The bidirectional histogram shows the half-month emotional opinions from customers for stores during the past1.5 years.

3. River Metaphor with WordCloud shows the counts of half-month for each star interval and the main words of comments for those intervals.

3.2 DATA PROCESSING

For calendar heatmap, we take the checkin.json in Yelp Dataset to process. The data structure of checkin.json is shown below:


```
{
  "business_id": "tnhfDv5Il8EaGSXZGiuQGg"

  "date":
        "2016-04-26 19:49:16,
        2016-08-30 18:36:57,
        2016-10-15 02:45:18,
        2016-11-18 01:54:50,
        2017-04-20 18:39:06,
        2017-05-03 17:58:02"
        ......
}
```


Figure 2: original check-in Data

The business id is the business we want to visualize, and the date represents there is a user checking on that day. First, we select a specific business. And then, sort the checkin into date-value pairs, which is shown below:

```
{
    “business_id”: “tnhfDv5Il8EaGSXZGiuQGg”

    "data":
        [ ["2018-01-01", 217],
          ["2018-01-02", 106],
          ["2018-01-03", 245],
          ["2018-01-04", 157],
          ["2018-01-05", 114],
          ["2018-01-06", 41],
          ……]
}
```


Figure 3: Count check-in Data

For bidirectional histogram and river Metaphor with WordCloud, we take the data from reviews.json in the yelp dataset from 2017-06-14 to 2018-11-14. Then group by the date for each 15days.

3.3 CALENDAR HEATMAP

In the calendar heatmap, what we used is a python module named piecharts. It can pass the date-value pairs into a calendar. And then, we set five levels represents different crowd degrees, which is shown below:

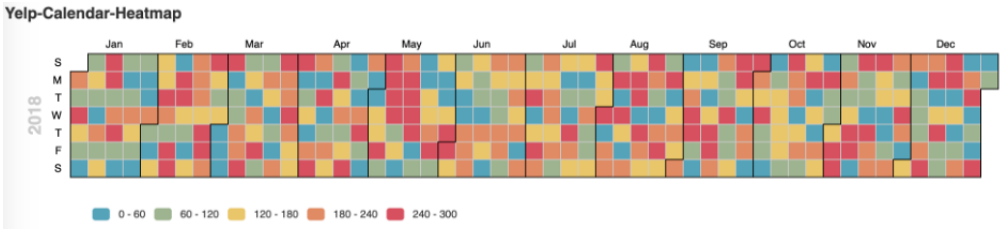


Figure 4: Calendar heatmap

For interactive usage, there are two methods. First, owner can select one specific day to know the number of checkin in that day:

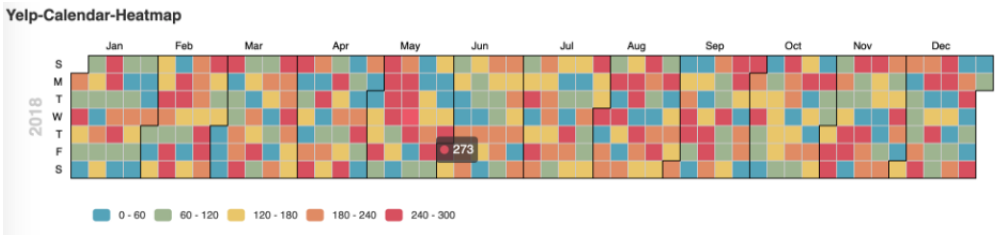


Figure 5: Remark: May 4th Tuesday has been chosen.

Second, the owner can select one color level to know the distribution of this crowd degree in the whole year:

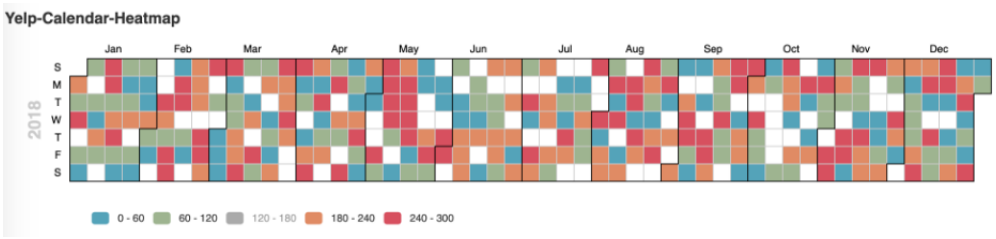


Figure 6: Remark: Crowd degree of 120-180 level is been chosen.

3.4 Sentiment Analysis of Rating

3.4.1 Bidirectional histogram

In the bidirectional histogram, what we used is a python module named pygal. we Analyze the half-monthly emotional opinions from customers for stores during the past 1.5 years.

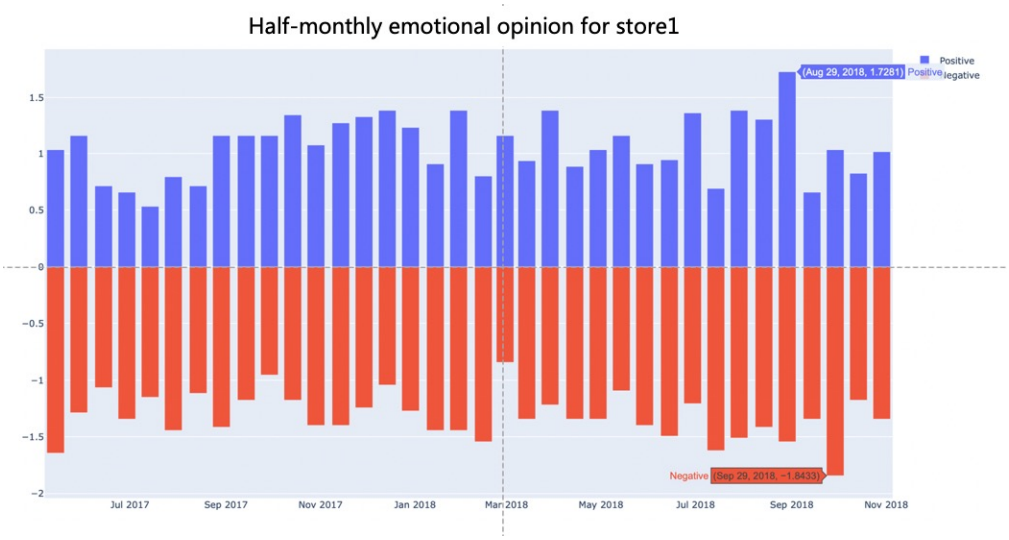


Figure 7: Bidirectional histogram

We can make some reviews from this graph. For example,Negative opinion is more than Positive but it is not bad as a whole look. And, 2018.08.14-2018.08.29 had the most positive opinion; 2018.09.14-2018.09.29 had the most negative opinion. We can help owners to modify their operating policies.

3.4.2 River Metaphor with WordCloud

In the river Metaphor with WordCloud, what we used is a python module named plotly. we will calculate the average stars of one specific business in the past 1.5 years, and show it by half-month time which means owners can make reviews every 15 days, with the comparison of positive and negative stars and then making some modifications.

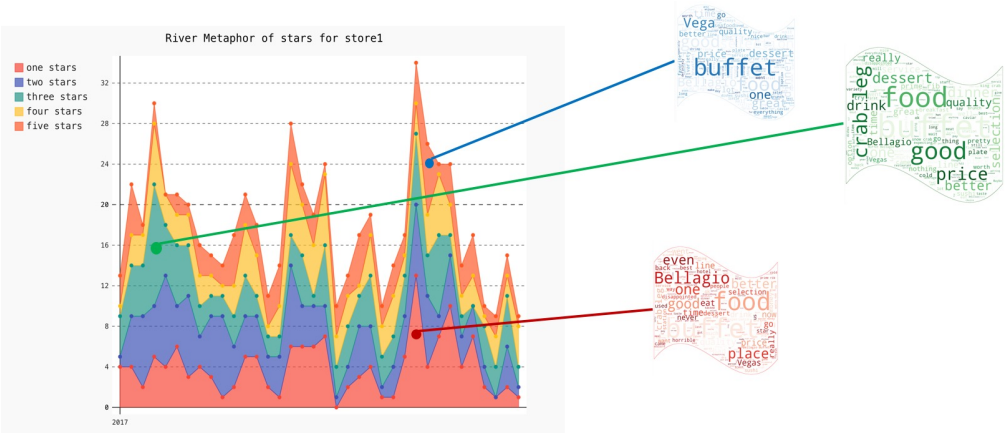


Figure 8: Bidirectional histogram

For example, this business did not have a good performance in the past 1.5 years. The stars from customers are concentrated on 1-3 stars, and comments are talking about food, price, Bellagio, baffle and so on. So the owner can strengthen their performance form those aspects. And the comments from 4-5 stars are talking about Vega, dessert, food and so on. Thus we can see that for this business, Food is the most influenced aspect of the business.

4 User Part

4.1 Task

For users, when users are faced with a large amount of business data, the problem often faced is that it is difficult to intuitively compare and select the business they are satisfied with. So, we want to build visualization tools to help users. It can intuitively understand the differences between different stores and can choose the store that suits them according to their location and preferences. It can also analyze the text data of reviews to help users extract valid information from reviews. Therefore, we decided to compare the differences in ”restaurants” in all categories through 4 interactive maps to help users choose the restaurant they want to go to.

1. Maps to show the location of restaurants around the user. (Using a restaurant in Las Vegas as an example)

2. The parallel coordinate chart to display high latitude data of all restaurants around the user, such as price, taste, rating star, etc.

3. Use word clouds to show positive and negative reviews for each restaurant.

4. Use radar charts to help users compare detailed data for specific restaurants.

4.2 Data Processing

Because in the yelp dataset, each business just has some basic information, just like business id, review count and categories. Obviously, we need more data to help us to introduce each business.

	business_id	name	address	city	state	postal_code	latitude	longitude	stars	review_count	is_open
0	1SWheh84yJXfytovILXOAQ	Arizona Biltmore Golf Club	2818 E Camino Acequia Drive	Phoenix	AZ	85016	33.522143	-112.018481	3.0	5	0
1	QXAEGFB4oINsVuTFxEYKFQ	Emerald Chinese Restaurant	30 Eglinton Avenue W	Mississauga	ON	L5R 3E7	43.605499	-79.652289	2.5	128	1
2	gnKjwL_1w79qoiV3IC_xQQ	Musashi Japanese Restaurant	10110 Johnston Rd, Ste 15	Charlotte	NC	28210	35.092564	-80.859132	4.0	170	1
3	xvX2CttrVhyG2z1dFg_0xw	Farmers Insurance - Paul Lorenz	15655 W Roosevelt St, Ste 237	Goodyear	AZ	85338	33.455613	-112.395596	5.0	3	1

Figure 9: Original Data of Business

We want to use a feedback system build by Huang et al. (2014) based on running Latent Dirichlet Allocation (LDA) algorithm on those textual data. LDA help us extract latent topics hidden inside tips and reviews, we call them subtopics. Then we rank every restaurant under each subtopic and generate a baseline. For restaurants under a specific subtopic baseline, we can give them corresponding feedback. According to this theory, we can pinpoint the pain point of each restaurant.

We use the python library NLTK to clean data of review text. Firstly, we use tokenization to act of separating a sentence into individual word tokens. It's the very first step towards processing individual words in a piece of text.

[they,greeted,us,with,warm,smiles,as,we,came,in]

Secondly,we removed the stop words in English. Stop words are words that appear extremely common but do not convey much meaning nor sentiment of bodies of text. We used the bag of stop words gathered by the programming package nltk and eliminated the words that falling into this bag from each piece of text.

[they,greeted,we,warm,smile,come]

Thirdly, we stemmed from the individual words in each piece of text using Porter Stemmer. Stemming refers to "the process of reducing inflected (or something derived) words to their word stem, base or root form generally a written word form". In our analysis, we used the nltk implementation of Porter Stemmer to bring individual words from text to its stem form. Note that during the processing steps, words with the same stem are assumed to convey roughly the same meaning, and therefore are weighted equally in sentiment analysis.

[they,greet,we,warm,smile,come]

Then we use the topic model of Latent Dirichlet Allocation (LDA). LDA imagines a fixed set of topics. Each topic represents a set of words. And the goal of LDA is to map all the documents to the topics in a way, such that the words in each document are mostly captured by those imaginary topics. LDA is a relatively complex generative statistical model. It can automatically extract N topics by extracting important information, and each topic has T words. Blei et al. (2003) In our experiments, we took 8 topics, operation, open, environment, value, taste, food, service, feeling. There are some keywords in each topic.

```
{'location': ['place','way','location','area','places','spot','parking','street', 'local', 'v
'open':['time', 'minutes', 'times', 'hour', 'home', 'today', 'later'],
'environment':['restaurant', 'table', 'atmosphere', 'dish', 'dishes', 'room','restaurants',
'value':['price','worth','quality','prices','high','pay','expect','average','cheap','expensiv
'taste':['sweet','spicy','garlic','chili','sauces','honey','flavor','salty'],
'food':['food','beef','pork','shrimp','egg','noodles','tofu','dumplings','rice','soup','duck
'survice':['service','friendly','server', 'servers','served','wait','ordered', 'order'],
'feeling':['great','best','delicious','amazing','happy','worth','fantastic','incredible','yum
```

Figure 10: LDA Output

For each business's each review, we count the keywords appearing in this review, add his corresponding score to the topic corresponding to this business, and finally calculate the average score. Then we get 8 new topics' score for each business.

business_id	stars_y	categories_count	location	open	environment	value	taste	food	survice	feeling
15KlfPDU5lQ07l79Wn1Eew	4.0	2	4.25	1.00	3.93	4.00	4.00	5.00	4.50	4.50
1Df5WnLX3DqN6ymlhqznaQ	4.5	6	3.89	3.97	3.93	4.01	4.19	3.83	3.81	3.80
1Oa6PpE1MYDT8OLRm8SBig	4.5	3	4.83	4.00	4.40	5.00	2.00	5.00	5.00	5.00
1YxLacCdn4yYQDPUzdye8g	4.0	4	4.83	4.88	4.40	5.00	2.00	5.00	5.00	5.00
1bfG-RJrbTmPBPHXdTx8ww	3.5	3	5.00	5.00	4.40	5.00	2.00	5.00	4.00	5.00
...	...	...	...	...	...	...	...	...	...	...
QXbCdmyDFHFK_cbG4TPVQ	2.0	3	3.14	3.50	3.75	3.00	4.33	2.00	3.00	4.00
7Qt8-mOdqkAri_9t74h_EA	3.0	9	2.50	2.50	3.75	1.00	3.33	4.00	3.00	2.43
7W8S1-A8sB3ngsoHyNfIGA	5.0	2	3.75	4.50	3.75	5.00	3.33	3.83	5.00	5.00
7Wr0piSMGiq5qy5opiLNrA	3.5	4	3.75	4.50	3.75	4.80	3.33	3.83	3.45	3.25
7aHzNulGbSgDpllHT3Qfiw	4.0	2	1.50	1.00	3.55	4.00	3.33	4.50	2.50	1.00

Figure 11: Result of Data Processing

After that, we got enough data to describe a restaurant.

4.3 Rough Selection

Rough selection is the first step, which means that help user to select several restaurants to compare. For example, the user needs to decide the basic distance range, food category, and price range. So we need to show all the candidate restaurants. 1.For a user to select an ideal restaurant, location is one of the most important factors. So we decide to use a map to show all the restaurants' locations first. We use the visualization tool Plotly, which supports zoom in and zoom out. And it also supports to use "lasso select" to select a range on the map. And we can get a subset of all candidate restaurants.

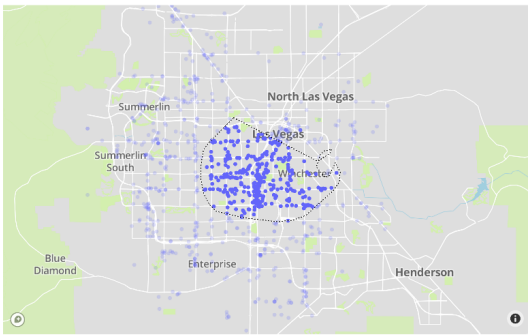
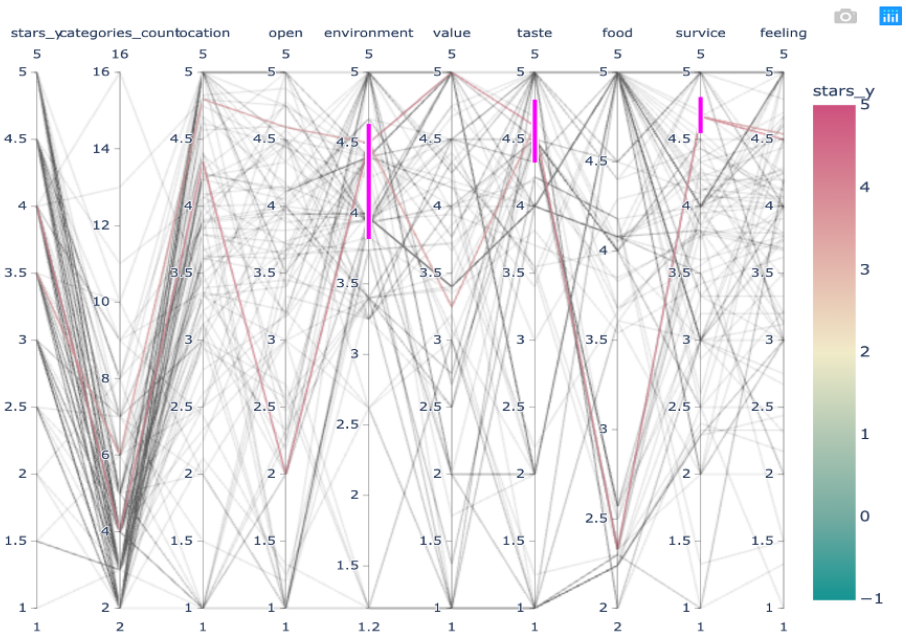


Figure 12: Map of Candidate Restaurants

2. We want to show the restaurant we extract from the review text. Because it is high dimensional data, we used parallel coordinates to visualize it. Each line in the graph is a specific restaurant. And each axis is an attribute. The color of each line means star of this business, red represents 5 stars and green represents 1 star.

User can select a range on each axis, to filter some important attributes he thinks. After this process, the user can get still do more filter until leave 2 or 3 restaurants. Then we can do a detail compare. For example, in the image below, we can only select the restaurant in which environments score higher than 4, taste's and service's score higher than 4.5.



.png

Figure 13: Parallel Coordinates

4.4 Detail comparison

After checking a user preference by rough selection step, the user can compare details through two visualization graphs, one is word cloud to display positive and negative text from review comment, and another is a radar chart to compare 7 topics of the selected restaurants. Word cloud displays how frequently words appear in reviews of each restaurant, by making the size of each word proportional to its frequency. In this project, we provide two-word clouds, which are positive and negative, to visualize more useful information for the user's decision. To implement this, we classify the review and text with their sentimental value as positive sentiment and negative sentiment by the python TextBlob sentiment method. A radar chart is useful to compare multiple quantitative variables, especially the datasets scored high or low. Additionally, the user can see which variables have similar values or if there are any outliers amongst each variable, such as taste, stars, and feeling.

In figure 14, we use two colors (green and red) to represent each restaurant for word cloud and radar chart identically to make the user perceive each restaurants results easily. To visualize the difference between the positive and negative reviews, we use contrast images by thumbs-up representing positive and thumbs-down representing negative. Furthermore, the radar chart is interactive graph in HTML, it shows detail scores of the variables if clicked.

In this figure, a user find two restaurants considering taste food with the criteria (taste $>$ 4.5, food $>$ 4.5, and stars $>$ 4.0). Firstly, the user can get text information about the menu and feelings which other visitors left in the word cloud. Then, the user can compare the detail scores of other topics that the user doesn't consider important.

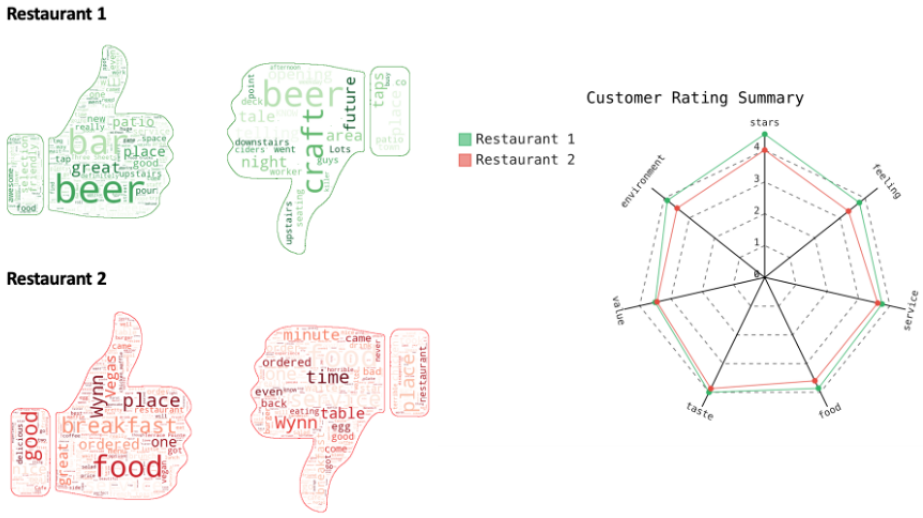


Figure 14: Detail Comparison

5 Conclusion

In summary, we designed different visualization tools for business owners to help their business and for users to provide useful information to decide where to go. For the business owner, we implemented the calendar heat map to monitor the flow of customers, which can interactively visualize the number of check-ins a day via HTML and display different crowdedness with a sequence color scheme. It helps the business owners gain insights about when was the busiest day and when was not. Furthermore, we implemented 3 visualizations to monitor the trend of stars and reviews by customers every 15 days for 1.5 years. If can easily display the difference about the star and review trend with bi-directional histogram, river metaphor, and wordcloud. The owners can consider some improvements by monitoring customers' reactions periodically.

For the users, we implemented the map and parallel coordinates to help users make rough choices from various decision factors such as location, taste, stars. To provide detail information for the users, we implemented positive and negative wordcloud from reviews and radar chart with 7 different variables. With the detail visualization, the user can perceive many text information at first glance, and get the information about frequent reviews about the restaurants. Additionally, the radar chart visualizes the scores of each restaurant to make the users perceive the difference between them. Especially, parallel coordinates and radar charts are effective to display the difference among multidimensional data, and they can be used to compare the restaurants intuitively for the users.

6 Future Work

For the parallel coordinates of the user part, because plotly does not support the display of the filtering results, the user cannot intuitively see the business that he has filtered. Afterward, we hope to use d3 to achieve a table below the map and parallel coordinates to display the current filtering results even.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

James Huang, Stephanie Rogers, and Eunkwang Joo. Improving restaurants by extracting subtopics from yelp reviews. *iConference 2014 (Social Media Expo)*, 2014.