# Group 9 Yelp Data Analysis

Fangfei Lin, Lu Chen, Yezhou Li, Chushi Shi

November 20, 2019

## 1 Background

The Yelp dataset has both user and business data. There are 3172 Asian restaurants of five categories, Chinese, Japanese, Korean, Thai and Vietnamnese, in the Greater Toronto Area. Our goal is to provide the business insights from both the perspectives of business owners and users, so that business owners are able to improve their business from data-driven decisions. Our analysis will focus on three parts: social network, sentiment analysis and yelp rating prediction.

Nowadays, social media becomes a trend. Instagram, Facebook, Snapchat, the social media impact is ubiquitous. By reconstructing and analyzing the social network of Yelp, we aim to search for target influencers for business in distinct categories. Moreover, connecting with customers is about knowing what they're thinking and what they want. The first step to improve the business is to fully understand customers' behaviours by social network.

Moreover, the most valuable information is hidden in the review texts. By breaking down people's comments, people's sentiments will be naturally disclosed, enabling us to learn their preference and disfavour.

## 2 Social Network Analysis

In order to effectively and effciently expand influence of each asian restaurant, we goal to build a user-to-user and user-to-business social network to find their top target customers for each business. Specifically, we see each user and business as a node in the network. And for the edge, we build user-to-user network with friend relationships between users, and user-to-business network with relations that a user had a review or gave a tip to a business.

| User Id | PR |
|---|---|
| iLjMdZi0Tm7DQxX1C1_2dg | 1.632320e-03 |
| ACUVZ4SiN0gni7dzVDm9EQ | 7.491643e-04 |
| CoS9VbTzh74Hh7FSlAKKVA | 6.944753e-04 |
| GFyA9ULGAeD-xZEPto2y7A | 6.649296e-04 |
| 58yXn5Y4409kc9q88YwU6w | 6.330321e-04 |

Table 1: Descending Sorted User PageRank

### 2.1 Popularity Computation

PageRank[1] is a very classic and accurate rank algorithm used by Larry Page to rank web pages and it works by counting the number and quality of links to a page to determine a rough estimate of how important the website is[2]. Here is a key formula below:

$$PR\left(p_i\right) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR\left(p_j\right)}{L\left(p_j\right)}$$

where $p_1$, $p_2$, ..., $p_N$ are the pages under consideration, $M(p_i)$ is the set of pages that link to $p_i$, $L(p_j)$ is the number of outbound links on page $p_j$, and $N$ is the total number of pages. In our situation, we treat each user and business as a "web page" to calculate the $PR$ and rank them. And the followed Table 1 and Table 2 show results.

### 2.2 Insights from PR

In the PR graph 1 of user-to-user, it indicates that in the social network there are a little few

| Business Id | PR |
|---|---|
| r_BrIgzYcwo1NAuG9dLbpg | 2.442723e-04 |
| aLcFhMe6DDJ430zelCpd2A | 1.577780e-04 |
| RtUvSWO_UZ8V3Wpj0n077w | 1.498646e-04 |
| N93EYZy9R0sdlEvubu94ig | 1.048958e-04 |
| RwRNR4z3kY-4OsFqigY5sw | 9.358741e-05 |

Table 2: Descending Sorted Business PageRank

powerful users can affect a large number of people. We define these users as $Influencer$[3] and main foucs on $Top$ 100 $Influencer$.

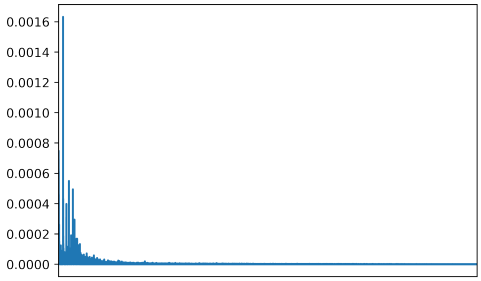


Figure 1: User PR Graph

We visualize 2 the social network of these influencer at a density of 2%. Most edges(relations) point to a few nodes(users) which are influencer. In other words, if the business advertises these influencers specifically, it can increase the exposure and impact area of the advertisement, which will be more efficient.

### 2.3 Top 100 Influencers Label

After we have selected the top 100 users, we can assign labels to each of them based on their preference toward various category. We are curious about their taste perference. Thus, we draw a plot of the number of reviews of each user in each category of asian restaurant. The figure 3 is as below.

In order to help business owners find their target influencers, we label each of the top users 3, which represents the category of their favorite asian restaurants. By doing so, a business owner can find the most influential users by searching for its restaurant category. These users may do some help if business owners want to do some advertisements to stimulate consumption.

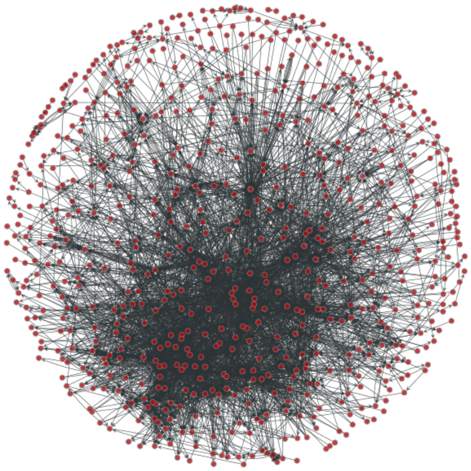


Figure 2: User Network

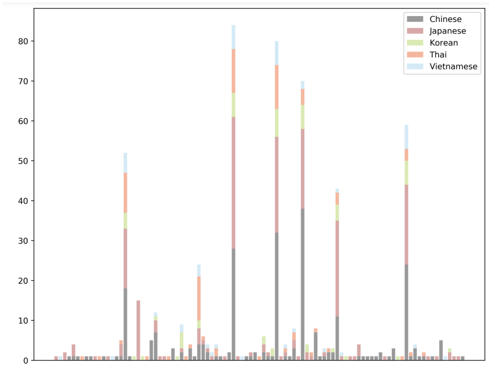


Figure 3: Asian Restaurant Visited by Top 100 Influencers

## 3 NLP: Sentiment Analysis

In order to gain insights of various types of restaurants, we aim to learn the characteristics of each category from the positive words and negative words in the user's reviews based on each words' polarity score.

### 3.1 Preprocessing the review texts

We remove all the reviews written in foreign language. First, we labeled each review text. If the user rating stars are larger than 4 (includ-

| User Name | CN | JP | KR | TH | VT |
|---|---|---|---|---|---|
| Ryan Gallagher | 1 | 1 | 1 | 1 | 1 |
| Tracy Swanson | 1 | 1 | 1 | 1 | 1 |
| William Porter | 1 | 1 | 1 | 1 | 1 |
| Michael Bird | 1 | 1 | 1 | 1 | 1 |
| Scott Blevins | 0 | 1 | 0 | 0 | 0 |

Table 3: Label of Top 100 Influencers

ing 4 stars), we labeled it as 'positive'; else if the user rating stars are below 3 (including 3 stars), we labeled it as 'negative'. Since the number of reviews above 4 stars and the number of review below 3 stars are balanced. Second, we calculate the number of words in each review, and use this as a feature to train our model. Third, we convert all the letters to lower case and remove all the punctuations. Also, we lemmatize each word. Fourth, I used NLTK stopwords dictionary. Also, I created a stop word list specifically for restaurants reviews, including 'restaurants', 'toronto', 'korean', 'chinese'. By removing all the stop words, we wish increase the accuracy of our model.

### 3.2 Calculating the weights of each word

We used bag of words approach to tokenize our review texts. It simply calculated the number of apperance of each word in the review texts. Since each word was treated as an individual feature, our feature matrix would be a high dimensional sparse matrix. By using Linear SVM, it will calculate the weights of each word in the review text in affecting the positivity/negativity of each review. The formula is as below:

$$min_{w \in r^N} \frac{\lambda}{2} \|w\|^2 + \frac{1}{m} \sum_{i=1}^{m} max\{0, 1 - y_i w^T x_i\}$$

$w$ represents the weight of each words, and $x_i$ are feature, $y_i$ is the labeled class.

### 3.3 Calculating the polarity of each word

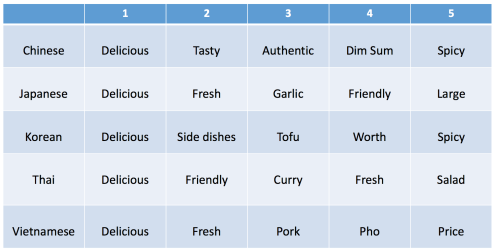
To explore insights of each category of restaurants, we calculate the polarity of sentiments of each word. The range of polarity is between 0 and 1. Closer to 1 means that the positive sentiment is stronger. Closer to -1 means that the negative sentiments is stronger. The formula is as below:

$$Polarity(t,c) = score(t) * \frac{totalfrequency(t,c)}{numberofreviews(c)}$$

where t refers to word, and c refers to category. Total Frequency(t,c) means the number of word t in the reviews of type c restaurants; number of reviews(c) refers to total number of reviews in category c; score t refers to the weight of each word t.

### 3.4 Result: Top words in each category

The top positive 4 and negative 5 words are as below. People explicitly concern more about the taste of food than service. The taste of food directly impact the star ratings of restaurants, since 'delicious','taste' and 'food' are of high rank in both top positive and negative words. The polarity score of word 'delicious' is way higher than the polarity score of word 'friendly' though they are both in the top positive words in each category. In addition, we can tell people's preference and disfavour toward specific dishes and taste in each category. The detailed analysis result is in our Shiny App.

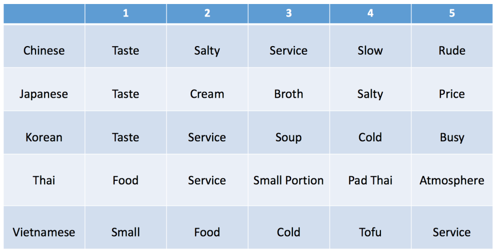


| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Chinese | Delicious | Tasty | Authentic | Dim Sum | Spicy |
| Japanese | Delicious | Fresh | Garlic | Friendly | Large |
| Korean | Delicious | Side dishes | Tofu | Worth | Spicy |
| Thai | Delicious | Friendly | Curry | Fresh | Salad |
| Vietnamese | Delicious | Fresh | Pork | Pho | Price |

Figure 4: Positive Words

### 3.5 Example: Thai

The chart of top ten positive 4 and negative 6 words are shown as below. Top positive words are 'delicious','friendly','curry', 'fresh', 'salad'. From their polarity score, the taste of food always come to the first place and is clearly



|  | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Chinese | Taste | Salty | Service | Slow | Rude |
| Japanese | Taste | Cream | Broth | Salty | Price |
| Korean | Taste | Service | Soup | Cold | Busy |
| Thai | Food | Service | Small Portion | Pad Thai | Atmosphere |
| Vietnamese | Small | Food | Cold | Tofu | Service |

Figure 5: Negative Words

| Word | Score | Freq | P Score |
|---|---|---|---|
| delicious | 0.646394 | 1098 | 0.124189 |
| friendly | 0.380387 | 647 | 0.0430639 |
| curry | 0.126946 | 1790 | 0.0397607 |
| fresh | 0.296449 | 595 | 0.0308639 |
| salad | 0.166103 | 630 | 0.0183106 |
| khao | 0.120993 | 688 | 0.0145657 |
| make | 0.202987 | 408 | 0.0144915 |
| new | 0.295324 | 279 | 0.0144174 |
| packed | 0.439906 | 172 | 0.0132395 |
| tea | 0.171127 | 423 | 0.0126661 |

Table 4: Top 10 Positive Words

more important than the service. Some popular dishes are curry and salad. Since there are a lot of fresh fruit and vegetables used in Thai cuisine, the word 'fresh' is not surprising 6.

| Word | Score | Freq | P Score |
|---|---|---|---|
| food | -0.076080 | 3433 | -0.0457012 |
| pad | -0.116050 | 1700 | -0.0345205 |
| portions | -0.325669 | 309 | -0.0176084 |
| soup | -0.156183 | 621 | -0.016971 |
| small | -0.239382 | 352 | -0.0147441 |
| see | -0.352785 | 223 | -0.0137657 |
| roll | -0.416159 | 185 | -0.0134715 |
| give | -0.345666 | 211 | -0.0127621 |
| service | -0.058574 | 1240 | -0.012709 |
| atmosphere | -0.234533 | 307 | -0.0125987 |

Table 5: Top 10 Negative Words

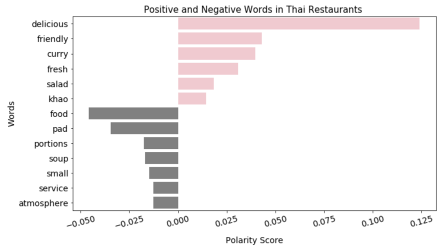


Figure 6: Positive and Negative Words

# 4 NLP: Prediction

## 4.1 Data Preparation

We first clean the review text by the same tricks used in sentiment analysis. However, instead of bag of words, we use tf-idf to reflect the importance of each words in the review texts. Tf-idf will automatically filter the stop words because of their high frequencies in the review texts. We labeled all the reviews above 4 stars as 'positve', otherwise, we labled them as 'negative'.

## 4.2 Model: Naive Bayes Classifier

The accuracy of Naive Bayes is 81%. The accuracy score calculate the proportion of number of correctly predicted positive reviews in total reviews. The precision and recall are 74% and 85%.

# 5 Strength and Weakness

## 5.1 Pros.

1. We analyze the dataset from the perspectives of business owners. We spent efforts in reconstructing the social network to analyze the users' behaviours.

2. Our approach is easy to follow and intuitive.

## 5.2 Cons.

1. In processing the texts, we remove all the reviews in foreign characters. In our analysis, we only focus on five categories of

Asian restaurants, indicating that exclusion of foreign language comments may cause strong bias in the result. Based on our personal experience as Asians, we inclined to use our native language to write the reviews. Our analysis is based on the assumption that sentiments of reviews in English are the same as the sentiments of reviews in other language.

2. The model and text preprocessing in the prediction are very weak. We should try more models to compare their perfomrances, or improve the accuracy of models by tuning the parametres.

3. There are a lot of fake users in Yelp and it is very difficult to filter all the fake users.

## 6 Reference

1. Page, Larry, "PageRank: Bringing Order to the Web". Archived from the original on May 6, 2002. Retrieved 2016-09-11., Stanford Digital Library Project, talk. August 18, 1997 (archived 2002)

2. PageRank. (n.d.). In Wikipedia. Retrieved 2002, from https://en.wikipedia.org/wiki/PageRank

3. KATONA, ZSOLT, et al. “Network Effects and Personal Influences: The Diffusion of an Online Social Network.” Journal of Marketing Research, vol. 48, no. 3, 2011, pp. 425–443. JSTOR, www.jstor.org/stable/23033849.

4. Yu bo, Jiaxu Zhou, Yi Zhang, Yunong Cao,(2017) Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews

## 7 Contribution

- Chushi Shi: data cleaning and network analysis

- Yezhou Li: Shiny App and LDA

- Fangfei Lin: NLP and EDA

- Chen Lu: users’ analysis