Exploring Regional-based Patterns of Restaurants Using Yelp Data

Dongjie Fan
df1676@nyu.edu

Xinshi Zheng
xinshi.zheng@nyu.edu

Ziman Zhou
zz1598@nyu.edu

June 11, 2017

Abstract

In this paper, we focus on Chinese and Japanese restaurants that are highly-rated by Yelp users. We develop a reproducible machine learning framework to uncover the patterns behind these popular restaurants. By combining both text analysis and anomaly detection techniques, we are able to highlight both main types of successful restaurants, as well as unique ones. We are confident that this procedure will become a useful yet inexpensive tool to understand a city's restaurant profile.

1 Introduction

Food services have always played a major role in an urban system. According to the Restaurant Industry Pocket Fact-book published by National Restaurant Association, there are over 1 million restaurant locations in the United States, and the restaurant industry sales are projected to total $798.7 billion in 2017. In particular, among the Commercial Restaurant Services with an estimated total of $736.3 billion, the sales of Eating Places account for $551.7 billion [1].

The Fact-book also indicates that "9 in 10 consumers say they enjoy going to Restaurants" [1]. As the types of cuisines are no longer restricted in their origins, we wonder how different or similar in people's preferences on the sub-cuisines (typical regional dishes) of a cuisine in different regions. This drives us to dig deeper into the factors in terms of cultures, spatial locations, and other characteristics.

Yelp is a popular on-line database for local business listings. Businesses can create and update their information, and highlight their products or services on the Yelp application to attract users. Users are also able to rate and review businesses to provide feedbacks to the business owners and other potential Yelp users. This interactive platform provides additional references for researchers and business analysts to answer questions such as what characteristics a successful restaurant business shall have.

We are interested in using Yelp data to explore different cuisines in different cities in North America. Specifically, our analysis studies the highly rated Chinese and Japanese restaurants in three city regions and draws insights by applying statistical learning methods. We believe that a deeper understanding of these businesses will be beneficial for restaurant business owners seeking success or effective improvement in a specific city.

2 Related Work

Various researches have been conducted using Yelp data [13], [4]. Previously, Jack Linshi worked on Personalizing Yelp's star ratings by topic modeling processes (a modified latent Dirichlet allocation (LDA)) on Yelp reviews that allowed users to learn the latent subtopics in review text [8]. Another research used Latent Dirichlet Allocation method to identify customer demand from Yelp reviews [5]. On the other hand, Anomaly detection was utilized for the time series pattern study of Yelp rating changes [3]. However, so far none of them has applied their techniques in combination with the restaurant businesses' self-described information in the Yelp database to get deeper insights.

GitHub Repository: https://github.com/djfan/yelp-challenge

3 Dataset

The dataset we use in our study comes from the Round 9 of Yelp Data Challenge [6]. The entire dataset includes information about local businesses in 11 cities across 4 countries. It contains 4.1 million reviews by 1 million users, and 1.1 million business attributes for 144 thousand businesses. For our study, we focus on the business and review subsets of the dataset.

The full description of the business and review subsets can be found in Appendix A. Apart from basic information, each business has an aggregated star rating from Yelp users, which is rounded to half-star intervals in the range of 1 to 5. The business dataset has information of each business' categories, provided by the business owners, describing the types of products or services they provide. Each business also has a list of attributes describing its characteristics, for example, whether a dining place is upscale or not. The review subset contains text reviews written by Yelp users for Yelp businesses.

4 Methodology

We define our project scope as following:

- We specifically study the Chinese and Japanese restaurants in three different cities in North America, namely Toronto, Phoenix and Las Vegas, as they are the top 3 cities within the provided Yelp dataset in terms of total number of restaurants.
- We assume the star rating for a restaurant is a reasonable indicator of its success.
- We will use only the Yelp data to conduct analysis in this project.

We firstly apply clustering methods to divide each city's restaurants into sub-groups, and our overall workflow can be summarized by Figure 1.

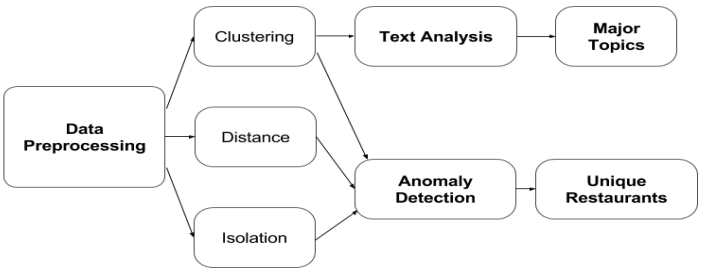


Figure 1: Overall Work Flow

4.1 Data-preprocessing

Our data-preprocessing is comprised of three major parts. To identify restaurants of our interest, we use the key word "restaurant" with selected words related to"Chinese" or "Japanese" in each business' category to acquire only the respective Japanese or Chinese restaurants. A list of selected words can be found in Appendix B.

We then apply a spatial K-Means clustering [7] using geographic coordinates of restaurants to refine the metropolitan area for each of the three cities. Finally, we construct features using restaurant-relevant attributes from the filtered business subset. Figure 2 shows the portion of Chinese and Japanese restaurants in three cities. Figure 3 shows the distribution of restaurants with different ratings across the three cities.

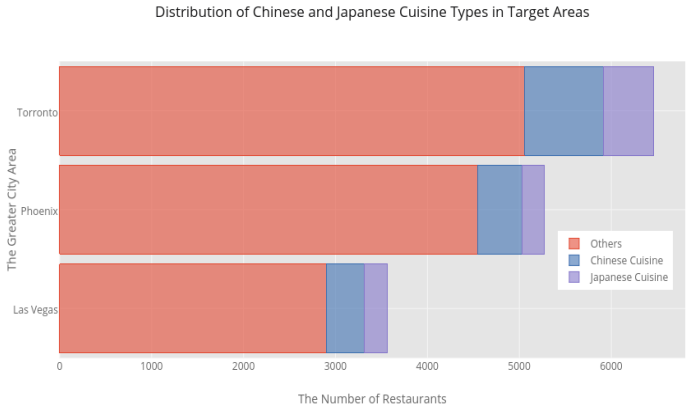


Figure 2: Portion of Chinese and Japanese Restaurants in Target Cities

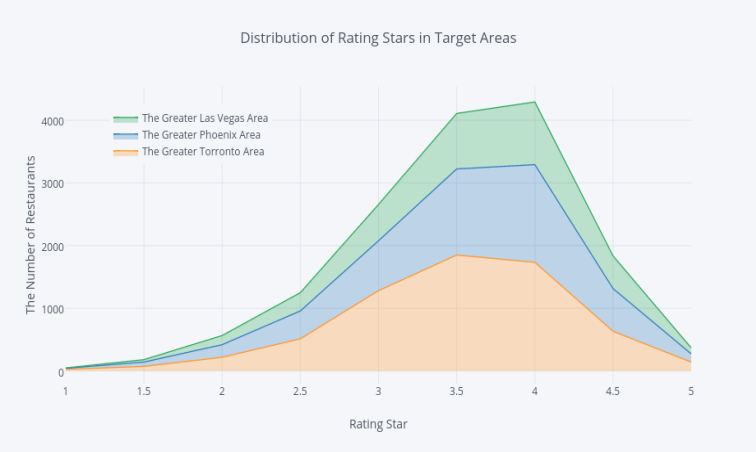


Figure 3: Distribution of Rating Stars in Target Cities

4.2 Restaurant Clustering

We perform three clustering methods to group restaurants in each city per cuisine type for further analysis. Different clustering techniques we use include K-means, Gaussian Mixture [11] and spectral clustering [15].

K-Means clustering computes the cluster centers and assigns each data point to its closest cluster based on distance to the cluster centers repeatedly until it reaches an optimal and stable state. Gaussian Mixture clustering computes the likelihood of each data point being in a specific Gaussian distribution with known parameters. With the assumption that all the data points are generated from a mixture of normal distributions with unknown parameters, the model captures the various distributions that the input data points could belong to. In our case, we expect the result of this modeling method to reflect the different patterns of the study restaurant groups and help to identify the most distinct ones. Spectral clustering computes the differences between instances by firstly using a similarity matrix. The similarity matrix used in this project is a standard sigmoid kernel.

These clustering methods are applied to the attribute features previously constructed for each city. The optimal number of clusters from each method is determined using Silhouette scores [14].

4.3 Review Topic Modeling

Latent Dirichlet Allocation (LDA) is a popular unsupervised topic model method in text analysis which has a probabilistic procedure to generate topics [2]. LDA views each document as a bag of words and draw a topic distribution θd $\sim$ Dir(α) over each document, where Dir($\cdot$) is a draw from a symmetric (uniform) Dirichlet distribution with scaling parameter, α. Besides, LDA uses

a symmetric Dirichlet prior on word distributions over each topic with scaling parameter, β. LDA model returns a per-topic word distribution and a per-document topic distribution, which can infer the theme of each topic and the most representative (most related) topic for each document.

The LDA analysis is conducted as described below. First of all, for each certain cuisine type, we aggregate all the reviews for each restaurant of a cuisine type. Then we pre-process text by lowering all words, removing punctuations and 'stop words' of English, German and French (Yelp data includes restaurants in cities where people speak above three languages). Also we remove other common words, which are not contained in the cuisine field. Then we build a document-term matrix. Here, each document (row) represents reviews of one restaurant. Next, we run LDA models by setting the number of topics (K) as 5 and 15, and iteration times as 500. From the result, we infer the theme of the each topic based on the most frequent words in per-topic word distribution. In addition, we find the most related (highest probability) topic for each restaurant based on the per-document topic distribution.

Besides, for each cluster in a city, we identify the most representative topic by selecting the most frequent topic which corresponds to the restaurants in this cluster. Then we combine the main topics of all the clusters for each city. Furthermore, we conclude the similarities and differences by comparing the topic combination of different cities. From these findings we are able to interpret the city-wide differences of a certain cuisine type.

4.4 Restaurant Anomaly Detection

In order to identify popular Chinese and Japanese Restaurants that are also unique, we apply several different anomaly detection methods. The methods we use include Distance-based anomaly detection (K nearest neighbor), Clustering-based anomaly detection, and Isolation Forest [9]. Specifically, We set $k = 5$ for the K nearest neighbor distance calculation. Reusing the K-Means clustering results obtained earlier, we compare the euclidean distances between each restaurant and its assigned cluster center. For the Isolation-based method, we set the number of base estimator in the ensemble of Isolation Forest to 100. Unique restaurants are then selected by intersecting the sets of results from the three anomaly detection methods.

5 Results

5.1 Topic Model

Our Analysis focuses on two cuisine types, Japanese and Chinese cuisine. Also we set areas below as our targets:

Area	City1	City2	City3	City4	...
The Greater Torronto Area	Toronto	Mississauga	Markham	Vaughan	...
The Greater Phoenix Area	Phoenix	Scottsdale	Mesa	Tempe	...
The Greater Las Vegas Area	Las Vegas	Henderson	North Las Vegas	Boulder City	...

Figure 4: Target Areas and Cities

For the Japanese cuisines, the top ten words for each topic (set the number of topics K = 5) are shown as below. We can infer that Topic 2 is mainly related with sushi bar because of words like 'sushi', 'roll', 'fish', 'ayce' and 'salmon' in the list; Topic 3 is relevant to Ramen Noodle House, because words like 'ramen', 'pork', 'noodles', 'broth', 'soup', 'spicy' are often listed on the menu of a Traditional Ramen restaurant.

Based on the clustering results, there are four clusters in each area. Detailed results are listed below:

From the results above, we conclude that for Japanese cuisines, there exists differences among different target areas according to their corresponding topic distributions. In greater Toronto, "Ramen Noodle House" is the most representative type of Japanese cuisine. In greater Las Vegas, "sushi bar" is much more likely to be the most typical type of Japanese restaurants. And the greater Phoenix Area stays in the middle.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	chicken	sushi	ramen	menu	korean
2	happy	roll	pork	japanese	bbq
3	hour	rolls	noodles	dishes	sushi
4	ordered	fresh	broth	dish	burrito
5	rice	fish	bowl	amazing	meat
6	table	eat	chicken	delicious	bulgogi
7	lunch	quality	japanese	ordered	montreal
8	shrimp	ayce	spicy	experience	bibimbap
9	pretty	salmon	ordered	vegas	meats
10	sauce	menu	soup	night	kalbi

Area	Result	Summary
The Greater Toronto Area	Cluster1: Topic 3 Cluster2: Topic 3 Cluster3: Topic 3 Cluster4: Topic 3	Topic 3 * 4
The Greater Phoenix Area	Cluster1: Topic 3 Cluster2: Topic 2 Cluster3: Topic 3 Cluster4: Topic 3	Topic 3 * 3 Topic 2 * 1
The Greater Las Vegas Area	Cluster1: Topic 2 Cluster2: Topic 2 Cluster3: Topic 2 Cluster4: Topic 3	Topic 2 * 3 Topic 3 * 1

For the Chinese Cuisines, the top ten words for topics (set the number of topics K = 15) are shown below. Based on these top ten words for each topic, Topic 13 is all about general descriptions of food and restaurants. We infer that Topic 12 is more likely to be categorized as 'American Chinese Food', which is a type of Chinese cuisine adapted for American tastes. Topic 15 is more likely to be traditional Chinese food or East-Asian food. Besides, it is evident to infer that Topic 4 relates to drinks and desserts.

	Topic 4	Topic 12	Topic 13	Topic 15
1	tea	chicken	happy	bowl
2	milk	chinese	menu	chicken
3	boba	rice	hour	sauce
4	drink	fried	server	rice
5	drinks	ordered	ordered	fresh
6	ice	egg	table	location
7	bubble	lunch	bar	teriyaki
8	sweet	soup	night	veggies
9	green	beef	drinks	eat
10	taste	shrimp	experience	meat

Based on the clustering results, there are two clusters for each city. Detailed results are as below:

From the results above, we conclude that for Chinese cuisines, there also exists differences among different target areas based on their corresponding topic distributions. In the greater Phoenix Area, "American Chinese food" is the most representative characteristic of the Chinese cuisines, while in greater Toronto area, Chinese cuisines are much more normal and traditional. Besides, in greater Las Vegas, drink shops which sell different kinds of tea is much more likely to be the most representitive types of Chinese restaurants.

In addition, If we take all the Chinese cuisine restaurants without clustering, we find Topic

Area	Result
The Greater Toronto Area	Cluster1: Topic 13 Cluster2: Topic 15
The Greater Phoenix Area	Cluster1: Topic 12 Cluster2: Topic 13
The Greater Las Vegas Area	Cluster1: Topic 4 Cluster1: Topic 13

12 becoming the most representative one. Thus, the method we use that applies clustering first and then combines the topics of each cluster could better reveal the characteristics of the latent subgroups.

5.2 Unique Restaurants: Case Studies

Combining the results from the different anomaly detection methods as discussed previously, unique restaurants are identified successfully. In the greater Toronto area, both Isolation Forests and K-Means clustering anomaly detection methods identify the same Chinese restaurant as the most anomalous one. We find that this restaurant is indeed unique among its peers in Toronto because of two reasons [12]. Firstly, although it labels itself as a Chinese restaurant, no food menu is found on its website, and its main business is alcoholic drink. Secondly, it has a business unit selling gifts and accessories.

For Las Vegas, we are also able to spot the most unique Japanese restaurant [10]. Both K nearest neighbor method and K-Means distance method rank it as the toppest anomalous Japanese restaurant with Isolation Forests ranking it the second anomalous. We find that though it labels itself as a Japanese sushi bar, it also provides therapy and acupuncture services to customers.

6 Evaluation

6.1 Data Preprocessing

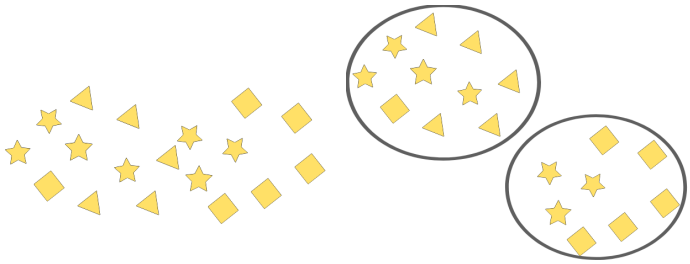
While filtering the businesses in Yelp dataset using the key word "restaurants", we assume that businesses with this word in its "categories" have met certain criteria – they are normal restaurants where customers can dine-in and interact with the restaurant environment and services. However, some businesses such as milk tea shops label themselves as restaurants. This might have a potential negative impact on our study results.

The process also requires some domain knowledge about cuisine types which can be quite subjective in our case when we pick vocabularies for a specific cuisine. For instance, according to our prior observation, we believe that "Asian Fusion" should be dominated by the "Chinese" dishes, so we include these restaurants as a part of Chinese cuisines. However, our model results indicate that such restaurants often provide other popular Asian cuisines such as sushi and bibimbap. Therefore, Asian Fusion restaurants serving these cuisine specific dishes are easily detected as "unique" (anomalous) ones among other Chinese restaurant businesses.

6.2 Topic Model

For each city and a certain cuisine type, we apply clustering methods first, extract the most representative topic of each cluster, and then take the combination of the main topics in all the clusters as the symbolic characteristic of the city, while a simpler alternative is to find the most frequent topics among all restaurants in the city.

We find the pre-clustering procedure necessary for the following reasons: In the cuisine fields, there are some latent sub-structure that we assume can be detected through clustering. For example, there exists restaurants for formal meal or for snacks. So, sometimes the major topic using the whole dataset is not representable enough. However, if there exists any latent subgroup, the combination of major topics of each subgroup could better represent the whole dataset. For example, in 5 ‘star’ is the majority in non-clustered group, which can be the representative one for the whole dataset; However, ‘triangle‘ and ‘rectangle‘ are the majority in two different clusters. Thus the combination (‘triangle’, ‘rectangle’) can be the representative one of the whole dataset.



(a) Nonclustered Group (b) Clustered Group

Figure 5: Two Groups

In the LDA model, some hyper parameters should be tuned including the number of topics K. While it is supposed to test a wide range of K and find the robust one, in practice, we set K as 5 and 15 considering the time consumption of modeling process and interpretation of the topic model.

6.3 Anomaly Detection

The three clustering methods we use (KMeans, Gaussian Mixture, Spectral) deliver quite different outputs depending on the nature of their algorithms. For anomaly detection purposes, KMeans enables us to locate the "unique" restaurants by spotting the largest distances between data points and the assigned clusters. Similarly, Gaussian Mixture is also a feasible method as the businesses with the lowest likelihoods of being in certain known Gaussian distributions are the most "special" ones. The Spectral clustering does not perform as well as we expected. Further tuning on parameters might be helpful to improve this model.

Both the distance-based and isolation-based methods work as desired and provide quite consistent results while identifying the most "abnormal" instances.

7 Future Work

In data preprocessing, we shall select more restricted key words to select "proper restaurants", and modify the list of vocabularies for each cuisine type for more accurate results.

For the text analysis, in order to make the result more robust, we shall keep tuning the hyperparameters for the topic-model (i.e. change the number of topics K). Besides, We can switch to a structural topic-model (STM) method to make the result more accurate and interpretable. Furthermore, it would be a good step for us to compare the differences across different regions for a certain cuisine type.

8 Appendix

A Data Schema

```
yelp_academic_dataset_business.json
{
    "business_id":"encrypted business id",
    "name":"business name",
    "neighborhood":"hood name",
    "address":"full address",
    "city":"city",
    "state":"state -- if applicable --",
    "postal code":"postal code",
    "latitude":latitude,
    "longitude":longitude,
    "stars":star rating, rounded to half-stars,
```

```
    "review_count":number of reviews,
    "is_open":0/1 (closed/open),
    "attributes":["an array of strings: each array element is an attribute"],
    "categories":["an array of strings of business categories"],
    "hours":["an array of strings of business hours"],
    "type": "business"
}

yelp_academic_dataset_review.json
{
    "review_id":"encrypted review id",
    "user_id":"encrypted user id",
    "business_id":"encrypted business id",
    "stars":star rating, rounded to half-stars,
    "date":"date formatted like 2009-12-19",
    "text":"review text",
    "useful":number of useful votes received,
    "funny":number of funny votes received,
    "cool": number of cool review votes received,
    "type": "review"
}
```

B Selected Words Related to Chinese or Japanese Cuisine

Chinese Cuisine	Japanese Cuisine
Chinese	Japanese
Cantonese	Sushi Bars
Szechuan	Tonkatsu
Shanghainese	Udon
Hainan	Ramen
Taiwanese	Japanese Curry
Dim Sum	
Hot Pot	
Hong Kong Style Cafe	
Asian Fusion	

C Group Contributions

Data Downloading, storing	Ziman
Data Pre-processing	Ziman, Xinshi, Dongjie
Text Analysis	Dongjie
Anomaly Detection - Clustering	Ziman, Xinshi
Anomaly Detection - Distance-Based	Dongjie, Xinshi
Anomaly Detection - Isolation Forest	Ziman
Method Evaluation	Ziman, Xinshi, Dongjie
Slides & Report	Ziman, Xinshi, Dongjie

References

[1] National Restaurant Association et al. Restaurant industry pocket factbook, 2017.

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

[3] Nikou Günnemann, Stephan Günnemann, and Christos Faloutsos. Robust multivariate autoregression for anomaly detection in dynamic product ratings. In *Proceedings of the 23rd international conference on World wide web*, pages 361–372. ACM, 2014.

[4] Longke Hu, Aixin Sun, and Yong Liu. Your neighbors affect your ratings: on geographical neighborhood influence to rating prediction. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 345–354. ACM, 2014.

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

[6] Yelp Inc. Yelp data challenge, 2017.

[7] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[8] Jack Linshi. Personalizing yelp star ratings: A semantic topic modeling approach. *Yale University*, 2014.

[9] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 413–422. IEEE, 2008.

[10] Shiatsu'ssage 365 Sports Massage. Welcome to aya's sports therapeutic massage therapy, 2017.

[11] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. 2012.

[12] Northwood. Northwood, 2017.

[13] Michalis Potamias. The warm-start bias of yelp ratings. *arXiv preprint arXiv:1202.5713*, 2012.

[14] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[15] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.