Northeastern University

CS 6220 Data Mining Techniques

Report

Mining Yelp Reviews

From latent ratings to a recommendation system

Authors:
Rashmi Dwaraka
Ritvika Nagula
Deepak Surana
Ruiyang Xu

Instructor:
Nathaniel Derbinsky

December 13, 2017

1 Introduction

Yelp[1] was founded in 2004 to help people find great local businesses like dentists, hair stylists and mechanics. Yelp has become an important website for many businesses. In this era of technology, it is very much clear that Yelp reviews and ratings have an intense effect on the success of businesses. The reviews on Yelp are about shopping,local businesses which offer specific services and restaurants, but we chose to only consider reviews about restaurants. When a person goes to a new restaurant and has a really good experience, he/she would probably make an effort to tell others about it. A good experience may usually imply either good food (in terms of taste), good service, or good ambience. A person may prefer one aspect over the other. Two different user ratings for the same restaurant might be same, but the reason behind the ratings may differ.We aim to build a model to calculate a user preference based rating instead of using the generic ratings provided for each restaurant by Yelp.

1.1 Dataset

The dataset is part of the Yelp Dataset Challenge[2] which is released for academic purposes. The dataset contains user reviews about various businesses. The total dataset is approximately 5.7GB in size. The Yelp dataset includes details regarding business, reviews, user, check-in time, and tip in the form of JSON objects.

Table 1: Attributes of Review

Field	Description
review_id	unique review id
business_id	unique business id
user_id	unique user id
stars	star-rating
date	date
text	the review itself
useful	number of useful votes received
funny	number of funny votes received
cool	number of cool votes received

Table 2: Attributes of Business

Field	Description
business_id	unique business id
name	business' name
neighborhood	neighborhood's name
address	full address
city	city
state	2 character state code
postal code	postal code
latitude	latitude
longitude	longitude
stars	star-rating
review_count	number of reviews
is_open	0 or 1 for closed or open
attributes	business attributes
categories	categories the business belongs to
hours	working hours

Our project involves working with the business.json, review.json files. Each review object mainly contains the text of the review, the unique business id for which the review was given, the user id of the user and the star-rating. Each business object contains many fields out of which the *business_id* and the *categories* fields important. The json attributes are listed in Table 1 and Table 2. In all, the Yelp dataset includes:

- 65K restaurants
- 3.2M reviews for the restaurants
- 1.1M users

1.2 Approach Overview

As is the general case with large datasets, we had to perform preprocessing operations to keep only the data that we need. To identify the user preferences, we mine the latent categories/features of user reviews and extract average ratings for each of these feature for a restaurant. We use Latent Dirichlet Allocation (LDA) to find the main topics in the reviews. LDA is a generative statistical model that allows sets of observations to be explained by unobserved groups. For example, if observations are

[1]Yelp - https://www.yelp.com/

[2]Yelp Dataset Challenge - https://www.yelp.com/dataset/challenge

words collected in review documents, each word in the document attributes to one of the documents topics. Each document is a mixture of small number of topics. Given a review, these topics reflects certain hidden criteria that users used to rate a restaurant. Once we obtain the most significant topics present in the reviews, we manually assign these topics to some particular categories (criteria like 'food', 'service', 'ambiance' etc.). We then calculate the average ratings of each restaurant. Based on the user preference assigned for each feature, we compute the restaurant rating by assigning corresponding weights to those features. This forms the core of our recommendation model.

2 Exploratory Data Analysis

2.1 Data Extraction

To extract only restaurants from the business.json file, we used the 'categories' field of the business JSON object. The category is a list of words of business categories. The Yelp API[6] documentation has details about the different types of categories that occur. We use only those corresponding to food and restaurants to retrieve the required businesses. With the business ids of these businesses, we filter out the reviews which correspond to only these business ids.

The business JSON object contains many fields out of which the 'business_id' and the 'categories' fields are of importance in the exploratory analysis. The review JSON object also contains different fields of which we use the 'business_id', 'review_id', 'user_id' and 'text'.

After filtering out the required businesses and reviews, it can be seen that there are **3216538 reviews** and **65028 businesses**. Out of the 65028 restaurants, there are only 39725 restaurants which are in the US and out of which 29730 are currently open and 9995 have closed down.

For building the recommendation model, we performed a series of data transformations as below:

- Convert json files into csv files for faster execution time and less memory storage.
- Clean the reviews text by removing stopwords as per the nltk corpus[3].

2.2 Data Analysis

We analyzed the yelp ratings of restaurants by generating basic distribution statistics on the filtered dataset and the results are as follows:

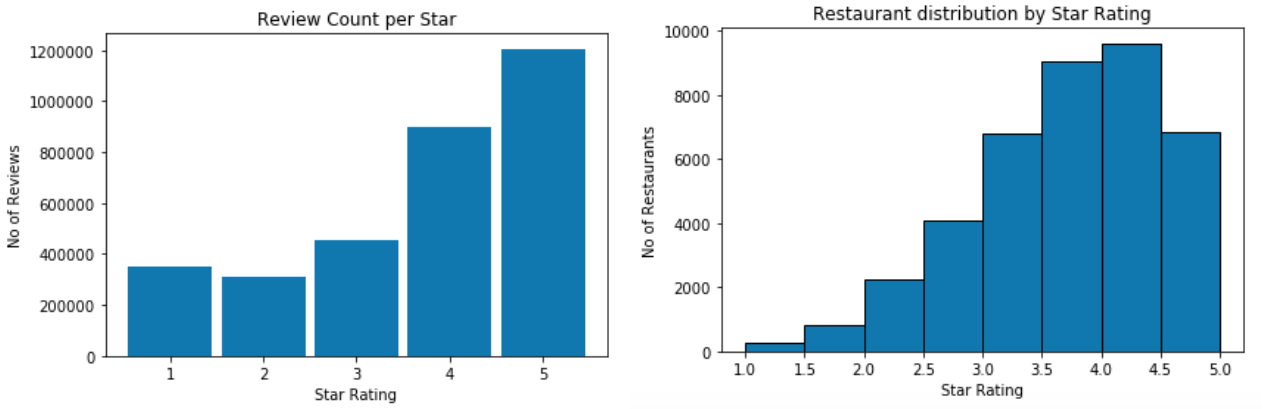

Figure 1: Distribution Plots based on Reviews and Restaurants

[3]NLTK - http://www.nltk.org/book/ch02.html

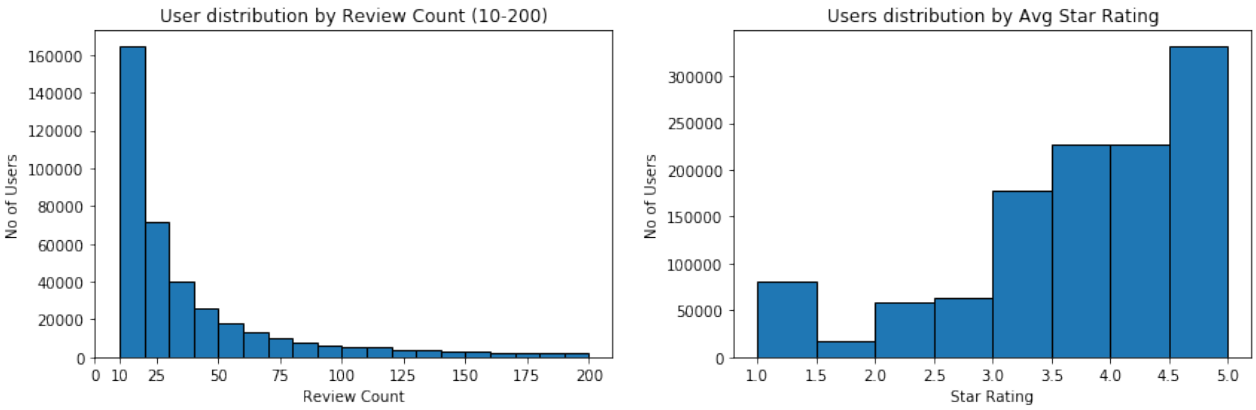

Figure 2: Distribution Plots for User based on Reviews and Ratings

The above distribution graphs verify the trueness of the data. We note that the reviews are distributed well across all possibilities and are sufficient for building a model closer to the real world. Moreover, for text analytics, we look for the main words that comprises the reviews' text. For this, we took 5-Star and 1-Star reviews separately and generated word clouds for both.

Figure 3: Word Clouds

The word cloud of 5-Star reviews gives the sense of the positive connotation with words like: 'will definitely', 'really good', 'highly recommend', 'love place', 'great service', 'food great', 'good food', 'come back', etc. On the contrary, the word cloud of 1-Star reviews gives the sense of somewhat negative connotation with words like: 'didn't even', 'never', 'took minutes', 'don't know', 'last time', etc. Moreover, these adjectives correspond to some aspect of the restaurant, like service, food, etc. which can be identified using a statistical model like Latent Dirichlet Allocation, Latent Semantic Indexing, etc.

3 Topic Modeling

As mentioned above, based on the overall average rating of a restaurant, we cannot determine the best or worst aspects of a particular restaurant. Here, we attempt to determine the latent aspects on which a restaurant is reviewed on Yelp. Our intuition is that a user reviews a restaurant based on various factors of his experience. The experience can be regarding the service, taste of the food, cuisine and/or several other aspects of a dining experience. These aspects help us suggest a better restaurant given a user preference. Moreover, it also helps the restaurants to improve on those aspects.

In natural language processing (NLP), Latent Dirichlet Allocation (LDA), is a Bayesian generative model which is used as a topic model to extract the latent topics that are present in a document, and could be used for our purpose. By assuming that there are K hidden topics (each topic is a Dirichlet distribution on the whole vocabulary) covered by a collection of documents, LDA treats each document as a bag of words generated from these K hidden topics. To be more specific, we analyze the below matrices generated by LDA:

- **Topic-Word Distribution**: K-dimensional vector of top 20 most probable words in the topic, ignoring the rest of the words.
- **Document-Topic Distribution matrix,** $\Theta_{D\times K}$: Probability distribution of each topic in the given document-word composition.

A review can also be regarded as being generated by first sampling a topic for each word's position, and then sampling a word from that topic for that specific position. For problems with high-dimensional textual data such as user reviews, it becomes difficult to extract relevant features. We choose the parameter K to be 50 based on the estimate of the large document set that we have.

By applying the LDA model[4] to our set of review documents, we extracted 50 topics. Figure A.1 shows the topics derived for restaurants based on reviews. We then summarized each topic Dirichlet distribution with top ranked words, and assign each topic categories with handcrafted labels.

For example,

1. Consider the words assigned to **Topic 0**

patio	seating	parking	outdoor	tables	atmosphere	staff	plenty	restau-
rant	location	dining	view	options	ambiance	packed	selection	sun-
day	saturday	vibe	building					

All the words for topic 0 indicate that the topic is related to the ambience of the restaurant.

2. Consider the words assigned to **Topic 7**

dessert	crepe	gelato	yogurt	toppings	desserts	strawberry	waffle	
crepes	banana	vanilla	caramel	scoop	serve	coconut	milk	custard
creamy	treat	cone						

All the words for topic 7 indicate that the topic is related to the desserts available in the restaurant.

3. Consider the words assigned to **Topic 14**

ordered	tasted	cooked	meal	seemed	soggy	tasteless	plate	burnt
reviews	overcooked	salty	quality	completely	waitress	stars	pieces	
restaurant	raw	undercooked						

All the words for topic 14 indicate that the topic is related to the taste and quality of the food in the restaurant.

Similarly we assigned the categories and in few cases multiple categories to a topic. We can see from the Figure A.1 that the reviews also depend on cuisine of the food. Hence we have further classified the food category to the type of food extracted from the topic words. Figure 4 shows the categories derived from the topics, as well as the food types.

[4]Mallet - https://programminghistorian.org/lessons/topic-modeling-and-mallet

Category	Topics
ambience	0,1,2,6,16,20,21,24,25,28,33,37,38,48
service	1,16,20,21,24,25,27,28,33,34,35,37,28, 40,45,47,48
price	1,15,24,36,45,48
delivery	22
taste	14,29,35,40,46,49
food	3,4,5,7,8,9,10,11,12,17,18,19,23,26,29, 30,31,32,34,39,40,41,42,43,44,47,49

Food type	Topics
american food	4,11,17,43
greek food	8
asian food	44
thai food	26,
korean food	19,
indian food	29,
vietnamese food	23
mexican food	30,
salad	31,39
breakfast	10,
vegeteranian	42
dessert	7,39,49
spanish food	5
italian food	41,
sea food	12,18,39
café	9
bar	34,40
german food	32

Figure 4: Handcrafted summaries for Categories and Food Type

Also, we analyzed the number of reviews for each topic to understand the distribution in the user reviews. The Figure 5 shows us that the majority of reviews are based on ambience of the restaurant. The distribution of the reviews over the topics can be seen in Figure A.3.

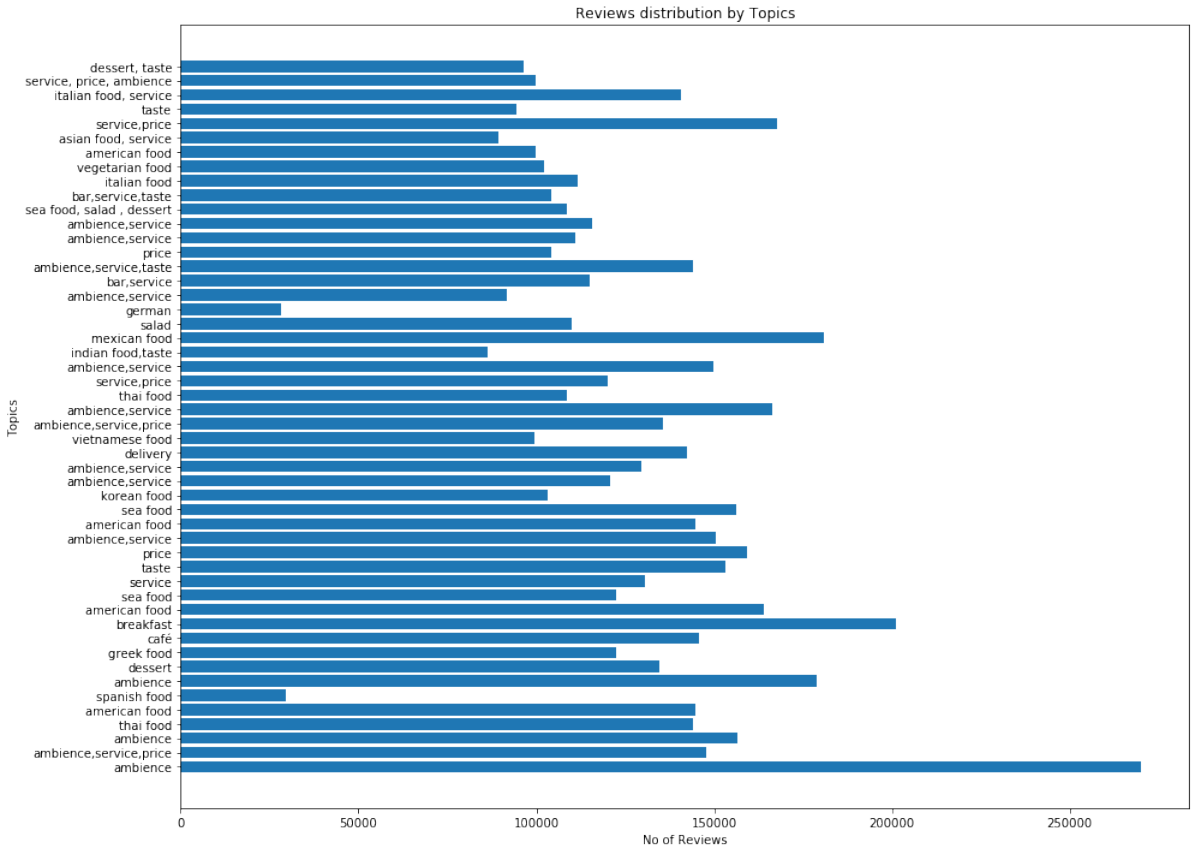

Figure 5: Distribution of Reviews per Topic

From Figure A.2, we note that the document distribution on each topic includes higher probability for two significant topics. Hence, we have restricted to two topics per review, because the other topics' probabilities are negligible in most cases (and using those negligible topics might lead to biases). Hence, these two topics can relate to latent features (categories) of a restaurant.

After extracting the two most significant features of the reviews, we now use these features in the form of a review-topic map in our model to recommend restaurants.

4 Recommendation Model

Using the review-topic mapping and the reviews for each restaurant provided by Yelp, we can calculate the average ratings per category for every restaurant. Due to the high computational requirement of calculating these ratings, we decided to use Apache Spark [5] to build the model. Apache Spark is an open-source cluster-computing framework which helps us in performing operations on large datasets efficiently.

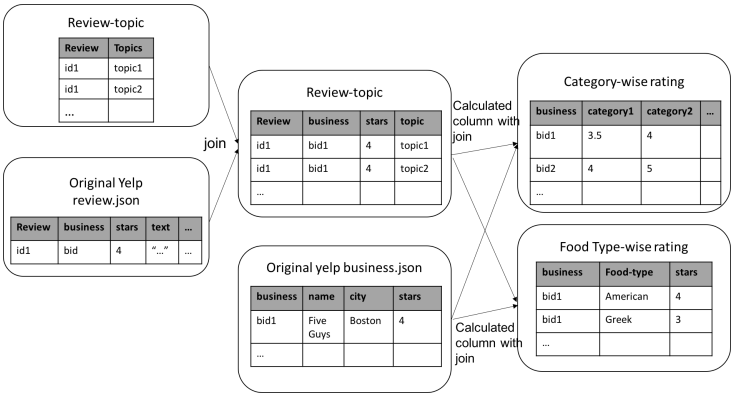

Figure 6: Data Transformation

With reference to Figure 6, we joined the review-topic map and the review.json file to get the business related to reviews. Based on the categories-topic assignment made as per Figure 4, we grouped the topics and considered average rating for those categories per business. By grouping the review ratings per category, we derive the average rating per category for a restaurant as:

$$LR_{b_k}(C_i) = \frac{1}{|R(b_k)|} \sum_{r \in R(b_k) \wedge t_j \in C_i} Sig(r, t_j) Rating(r)$$

where, C_i is the category interested;
$R(b_k)$ is the set of reviews for restaurant b_k;
$Sig(r, t_j) = 1$ iff topic t_j is significant for review r, otherwise $Sig(r, t_j) = 0$.

Similarly, based on the food category which has sub-categories as cuisines, we computed average rating based on food-type also for every restaurant.

business_id	name	overall_stars	city	ambience	service	price	delivery	taste	food
ObelBwE40B5oYm2aA8yTjw	The Harp	4	Cleveland	3.8139534	3.818181753	3.973684311	0	3.833333254	3.861702204
dj2XkboYShGM9WvpkayJWA	D'Rollz	3.5	Toronto	3.666666746	4	4	0	0	3
ydj6cSmcOM_jVYJLkHzdOA	The Dylan Bar	3	Toronto	3.454545498	4.5	4	0	2.5	2.599999905
lrPNLw0zbfzgWrWINMjsNw	Amber Restaurant	4.5	Edinburgh	4.380952358	4.571428776	4.428571224	0	2	4.263157845
38yD8L0Ersu3Z_ZZZsOEXw	International Herbs	4	Toronto	5	0	0	0	0	3
rzByiKaj-bLeLz-zKNBQdg	Dairy Queen	2	Pittsburgh	1.769230723	2	2	0	2.5	2.166666746
....									

Figure 7: A portion of result restaurant-latent feature matrix for a recommendation system.

[5] Apache Spark - https://spark.apache.org/

business_id	overall_stars	city	name	foodType	stars
ObelBwE40B5oYm2aA8yTjw	4	Cleveland	The Harp	american	3.878048897
ObelBwE40B5oYm2aA8yTjw	4	Cleveland	The Harp	greek	3.75
ObelBwE40B5oYm2aA8yTjw	4	Cleveland	The Harp	korean	4
ObelBwE40B5oYm2aA8yTjw	4	Cleveland	The Harp	breakfast	4
ObelBwE40B5oYm2aA8yTjw	4	Cleveland	The Harp	dessert	3.599999905
ObelBwE40B5oYm2aA8yTjw	4	Cleveland	The Harp	seaFood	3.84210515
ObelBwE40B5oYm2aA8yTjw	4	Cleveland	The Harp	cafe	4
....					

Figure 8: Food type subcategories to get ratings.

Once we have calculated all ratings for each restaurant, we can build a matrix where rows are restaurants and columns are latent features (topics). This matrix can tell us what is the rating per each latent feature per each restaurant. On the other hand, we could list those latent features in the user app, and ask the user to select their interested one. We can then rank the restaurant with their latent ratings, and returns the highest ranking ones. In this way, our restaurant-latent feature matrix can be used as an engine of a recommendation system (Figures 7 and 8).

We finally execute our model[6] which given the user preference for each category on a scale of 0-10, we assign weights to each category and calculate the restaurant rating according to the user preference. Sorting the list of restaurants based on this calculated rating, we can recommend restaurants as per user preference.

5 Conclusion

In this project, we tried mining latent rating criteria used by users from Yelp review dataset. We used Latent Dirichlet Allocation (LDA) as our topic modeling tool and extracted relevant information from the LDA results. Inspired by the research paper [8], we designed a new way to calculate latent ratings which only consider the significant topics in each review. The final result is a restaurant-latent feature matrix, which can be used as a core for a recommendation system based on user preference. Given a user preference for each category ranging from 0-10, we take each of this value and consider them as weights to calculate the weighted average rating for a specific user preference. We observed that this calculated rating ended up being more closer to what the user actually needs when compared to the actually existing rating given by Yelp. In all our analyses, we have not taken the location of either the user or the restaurant into consideration. Future work may involve taking this into consideration by recommending only those restaurants to a user which are in his present location.

[6] Jupyter Notebook - restaurant_recommendation.ipynb

References

[1] https://www.yelp.com/dataset/documentation/json

[2] https://datascience.blog.wzb.eu/2016/06/17/creating-a-sparse-document-term-matrix-for-topic-modeling-via-lda

[3] https://docs.scipy.org/doc/scipy/reference/generated/scipy.sparse.coo_matrix.html

[4] https://pythonhosted.org/lda/api.html

[5] https://rpubs.com/Zyrix/yelptask1

[6] https://www.yelp.com/developers/documentation/v2/category_list

[7] https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

[8] Huang, James et al. Improving Restaurants by Extracting Subtopics from Yelp Reviews. (2014).

[9] https://github.com/dwaraka-rashmi/Yelp-Recommendation-System

[10] https://youtu.be/WM1LDHFb28M

Appendix A Figures

A.1 Topics for 3.2M Reviews

Topic	Category Assigned	Topic Words													
0	ambience	patio	seating	parking	outdoor	tables	atmosphere	staff	plenty	restaurant	location	dining	view	options	ambiance
1	ambience,service,price	staff	atmosphere	highly	prices	absolutely	restaurant	tasty	los	owner	portions	customer	def	visit	welcoming
2	ambience	dining	visit	quality	ingredients	restaurants	review	fact	certainly	restaurant	less	despite	simple	diners	truly
3	thai food	rice	beef	dishes	noodles	dish	dumplings	shrimp	ordered	restaurant	noodle	order	asian	spicy	duck
4	american food	cooked	ordered	filet	meal	rib	potatoes	rare	steaks	restaurant	sides	dessert	salad	waiter	ribeye
5	spanish food	est	pas	des	une	mais	trs	cest	qui	avec	bon	dans	sont	vous	nous
6	ambience	restaurant	phoenix	staff	restaurants	places	visit	business	airport	scottsdale	cleveland	owned	weve	owner	valley
7	dessert	dessert	crepe	gelato	yogurt	toppings	strawberry	waffle	crepes	banana	vanilla	caramel	scoop	serve	coconut
8	greek food	greek	pita	hummus	salad	gyro	lamb	rice	plate	restaurant	falafel	tasty	ordered	shawarma	mediterranean
9	café	tea	milk	iced	latte	starbucks	boba	cafe	espresso	teas	staff	bubble	wifi	ordered	chai
10	breakfast	breakfast	eggs	brunch	toast	bacon	pancakes	hash	potatoes	ordered	waffles	sausage	benedict	sunday	gravy
11	american food	burgers	cheese	onion	bacon	bun	rings	ordered	poutine	order	chili	patty	potato	onions	beef
12	sea food	shrimp	crab	lobster	seafood	ordered	oysters	cooked	mussels	calamari	scallops	cajun	chowder	restaurant	meal
13	service	reviews	yelp	review	stars	based	gave	closed	visit	ordered	places	rating	fact	seemed	previous
14	taste	ordered	tasted	cooked	meal	seemed	soggy	tasteless	plate	burnt	reviews	overcooked	salty	quality	completely
15	price	quality	stars	prices	average	places	overall	less	rating	higher	review	value	restaurants	restaurant	reviews
16	ambience,service	location	parking	locations	staff	mall	located	opened	shopping	starbucks	chain	visit	employees	plaza	theyre
17	american food	sandwich	sandwiches	cheese	beef	sub	deli	grilled	ordered	order	philly	pastrami	subway	chips	tasty
18	sea food	sushi	tuna	salmon	sashimi	rice	tempura	ayce	spicy	quality	miso	chef	order	restaurant	sake
19	korean food	korean	rice	spicy	dishes	bbq	beef	restaurant	grill	kimchi	hawaiian	ordered	order	poke	tofu
20	ambience,service	restaurant	meal	chef	dining	dish	dessert	dishes	view	gras	tasting	waiter	foie	restaurants	reservation
21	ambience,service	tables	order	restaurant	waitress	plates	server	plate	glass	staff	seemed	plastic	served	napkins	ordered
22	delivery	order	ordered	delivery	ordering	customer	waited	location	delivered	placed	gave	takeout	received	deliver	extra
23	vietnamese food	pho	ramen	noodles	broth	beef	vietnamese	noodle	ordered	restaurant	places	spicy	rice	order	salty
24	ambience,service,price	atmosphere	crowd	tvs	bartender	staff	pub	bartenders	friday	saturday	bars	irish	group	packed	selection
25	ambience,service	customer	manager	customers	staff	business	employees	owner	order	management	attitude	employee	location	review	cashier
26	thai food	thai	curry	spicy	rice	dish	ordered	restaurant	dishes	spice	noodles	beef	order	coconut	shrimp
27	service,price	bill	tip	waitress	ordered	extra	order	manager	charged	server	meal	gave	groupon	coupon	waiter
28	ambience,service	staff	server	owner	visit	customer	manager	restaurant	chef	highly	absolutely	extremely	atmosphere	greeted	event
29	indian food,taste	indian	restaurant	dishes	rice	naan	spicy	curry	ordered	dish	lamb	masala	buffet	tasty	restaurants
30	mexican food	tacos	taco	chips	burrito	guacamole	cheese	carne	asada	rice	ordered	tortilla	margaritas	tortillas	burritos
31	salad	salad	dressing	cheese	ordered	lettuce	salads	grilled	caesar	tomatoes	tomato	meal	potato	spinach	tasty
32	german	und	der	das	ist	war	nicht	ich	sehr	mit	auch	wir	ein	aber	den
33	ambience,service	hotel	staying	casino	vegas	stayed	staff	strip	view	bathroom	parking	desk	group	resort	walking
34	bar,service	beers	selection	tap	craft	atmosphere	staff	draft	brewery	pub	ale	wines	brew	german	brews
35	ambience,service,taste	restaurant	staff	portions	prices	atmosphere	portion	meal	quality	overall	group	tasty	priced	decor	sizes
36	price	youll	theyre	review	fact	literally	walking	stomach	joint	giant	dress	apparently	entire	upon	mention
37	ambience,service	decor	tables	walls	restaurant	atmosphere	interior	chairs	seating	vibe	lighting	staff	seats	dining	ambiance
38	ambience,service	vegas	las	strip	restaurant	visit	prices	located	mall	places	hotel	casino	restaurants	location	visiting
39	sea food, salad , dessert	dish	cheese	dessert	duck	salmon	served	roasted	meal	cooked	salad	plate	ordered	sprouts	dishes
40	bar,service,taste	ordered	cocktails	server	cheese	cocktail	dip	appetizer	tapas	appetizers	atmosphere	bartender	sangria	tasty	order
41	italian food	pasta	restaurant	ordered	dish	salad	meal	meatballs	spaghetti	olive	cheese	calamari	dishes	ravioli	veal
42	vegetarian food	options	vegan	vegetarian	gluten	veggie	staff	smoothie	ingredients	meal	smoothies	tasty	protein	veggies	organic
43	american food	bbq	cheese	brisket	pulled	sides	ordered	potato	sandwich	slaw	sauces	meal	smoked	order	tasty
44	asian food, service	buffet	selection	vegas	buffets	dessert	station	quality	desserts	crab	seafood	rib	brunch	section	dishes
45	service,price	server	seated	waitress	order	ordered	restaurant	waited	waiter	tables	hostess	manager	reservation	seemed	meal
46	taste	crispy	dish	served	texture	cooked	onions	thick	chips	pieces	spicy	amount	topped	crunchy	creamy
47	italian food, service	pizza	cheese	pizzas	toppings	thin	slice	pepperoni	ordered	slices	sausage	style	order	oven	dough
48	service, price, ambience	selection	market	prices	grocery	foods	produce	shopping	products	section	organic	quality	staff	meats	joes
49	dessert, taste	cake	bakery	cupcakes	donut	cheesecake	cupcake	pastries	dessert	desserts	pastry	croissant	frosting	velvet	moist

A.2 Topic Probabilities for Sample Reviews

Review_ID	3VIKxuHcdjr4TXqVoMSY_g	zl6pIK698ZsQJlnV9Ax13A	zGgB9dU9rxfX0dm4MZSsPg	YlT7CgOMpWygckFG4mhfw	bFbiBJ4-la18EEta55-g5Q
Topic 0	0.344444444	0.004545455	0.00625	0.004166667	0.014285714
Topic 1	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 2	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 3	0.011111111	0.004545455	0.38125	0.004166667	0.014285714
Topic 4	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 5	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 6	0.011111111	0.004545455	0.06875	0.004166667	0.014285714
Topic 7	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 8	0.011111111	0.05	0.00625	0.004166667	0.014285714
Topic 9	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 10	0.122222222	0.004545455	0.00625	0.004166667	0.014285714
Topic 11	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 12	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 13	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 14	0.011111111	0.05	0.00625	0.004166667	0.014285714
Topic 15	0.011111111	0.004545455	0.06875	0.004166667	0.014285714
Topic 16	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 17	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 18	0.011111111	0.231818182	0.00625	0.0875	0.014285714
Topic 19	0.011111111	0.004545455	0.06875	0.004166667	0.014285714
Topic 20	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 21	0.011111111	0.004545455	0.00625	0.045833333	0.014285714
Topic 22	0.011111111	0.459090909	0.00625	0.004166667	0.157142857
Topic 23	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 24	0.011111111	0.004545455	0.00625	0.045833333	0.014285714
Topic 25	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 26	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 27	0.011111111	0.004545455	0.06875	0.004166667	0.014285714
Topic 28	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 29	0.011111111	0.004545455	0.00625	0.170833333	0.014285714
Topic 30	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 31	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 32	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 33	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 34	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 35	0.011111111	0.004545455	0.00625	0.3375	0.014285714
Topic 36	0.011111111	0.004545455	0.00625	0.045833333	0.157142857
Topic 37	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 38	0.011111111	0.004545455	0.06875	0.004166667	0.014285714
Topic 39	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 40	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 41	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 42	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 43	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 44	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 45	0.011111111	0.004545455	0.00625	0.0875	0.014285714
Topic 46	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 47	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 48	0.011111111	0.004545455	0.00625	0.004166667	0.014285714
Topic 49	0.011111111	0.004545455	0.00625	0.004166667	0.014285714

A.3 Distribution of Reviews over the Categories