Northeastern University

CS 6220 Data Mining Techniques

Project Update 1

Exploratory Data Analysis

Authors:
Rashmi Dwaraka
Ritvika Nagula
Deepak Surana
Ruiyang Xu

Supervisor:
Nathaniel Derbinsky

November 1, 2017

1 Dataset

In this project we are aiming to analyze Yelp Dataset which is released for academic research purpose. The dataset contains user reviews about various restaurants. The total dataset is of approximately 5.7GB. The Yelp dataset includes details regarding business, reviews, user, checkin time, tip in the form of json. We have decided to limit our analysis to the businesses which are restaurants.

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* and *text* for now.

After filtering out the required businesses and reviews (which will be explained below), 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.

2 Data Processing

Yelp has reviews related to not only restaurants but also other businesses. So we had to filter out the reviews based on the businesses that are restricted to only restaurants. To extract only restaurants from the business.json file, we used the 'categories' field of the business json object. The categories is a list of words of business categories. The Yelp API 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.

3 Topic Modeling

After analyzing the Yelp dataset as mentioned in the previous section, we decided to build our recommendation based on the user reviews. User reviews capture the sentiments of the user group and we feel it is the most important feature for recommending restaurants. Based on this intuition, we did the below data processing and modeling t extract the useful features.

The basic intuition for building a recommendation engine was to identify the user preferences. User preferences guide us with the features based on which we need to build our recommendation engine. We chose to do topic modeling using Latent Dirichlet Allocation tries to infer the hidden topic structure from the reviews to analyze these features.

Topic	Words														
1	selection	dress	prices	shopping	sale	staff	purchase	clothing	products	stock	dresses	quality	books	customer	section
2	event	staff	highly	photos	professional	flowers	guests	absolutely	photo	extremely	business	beyond	process	quality	customer
3	patio	atmosphere	seating	staff	tables	decor	outdoor	restaurant	chairs	dining	ambiance	vibe	selection	prices	plenty
4	bbq	cheese	brisket	pulled	ordered	sides	potato	order	sandwich	meal	sauces	tasty	slaw	spicy	crispy
5	vehicle	customer	repair	tire	dealership	tires	auto	parts	replaced	business	gave	warranty	manager	purchase	purchased
6	salad	dressing	ordered	greek	pita	hummus	sandwich	salads	lettuce	gyro	order	cheese	meal	tasty	grilled
7	parking	location	mall	located	strip	shopping	building	station	staff	plenty	court	plaza	toronto	places	restaurants
8	options	vegan	vegetarian	staff	gluten	yogurt	veggie	toppings	ingredients	smoothie	selection	places	order	organic	tasty
9	pizza	cheese	ordered	pasta	pizzas	order	thin	slice	toppings	sausage	pepperoni	salad	slices	meatballs	style
10	salon	haircut	stylist	highly	appointment	style	professional	gave	barber	customer	absolutely	cuts	makeup	extremely	highlights
11	staff	appointment	visit	insurance	dental	doctors	hospital	professional	patients	medical	health	glasses	highly	treatment	exam
12	est	pas	des	une	mais	trs	cest	qui	avec	dans	bon	vous	sont	nous	bien
13	seated	restaurant	tables	server	group	hostess	reservation	waited	quickly	ordered	seemed	waitress	order	greeted	saturday
14	buffet	selection	market	quality	foods	grocery	prices	section	station	buffets	produce	meats	asian	shopping	desserts
15	sushi	tuna	salmon	rice	sashimi	spicy	ayce	order	restaurant	quality	ordered	tempura	chef	shrimp	places
16	tacos	taco	chips	burrito	cheese	ordered	guacamole	rice	carne	asada	margaritas	order	nachos	tortilla	burritos
17	location	staff	phoenix	scottsdale	locations	visit	valley	places	business	opened	tempe	north	quality	owned	weve
18	dish	cheese	dessert	meal	restaurant	ordered	salad	pasta	served	cooked	duck	dishes	appetizer	salmon	beef
19	rice	restaurant	order	ordered	tapas	plate	dishes	tasty	hawaiian	portions	style	sangria	meal	dish	beef
20	breakfast	eggs	brunch	bacon	toast	pancakes	ordered	potatoes	hash	waffles	sunday	sausage	benedict	waffle	diner
21	crowd	vegas	group	bartender	lounge	bartenders	packed	saturday	friday	bars	vip	hip	vibe	cocktail	atmosphere
22	montreal	pittsburgh	visit	cleveland	building	toronto	edinburgh	church	places	library	youll	staff	cafe	walls	students
23	restaurant	meal	staff	atmosphere	dining	restaurants	dishes	quality	portions	chef	prices	meals	server	visit	highly
24	groupon	spa	staff	highly	relaxing	treatment	facial	professional	products	services	tan	massages	gave	pressure	appointment
25	order	ordered	server	waitress	bill	waiter	manager	restaurant	waited	meal	tip	gave	seated	tables	servers
26	order	delivery	customer	driver	received	coupon	company	delivered	website	business	rental	gave	manager	waited	extra
27	nail	salon	pedicure	gel	polish	manicure	appointment	customer	eyebrows	pedi	staff	regular	lashes	tech	prices
28	indian	restaurant	dishes	rice	lamb	spicy	ordered	dish	naan	curry	order	buffet	tasty	restaurants	vegetarian
29	cake	dessert	bakery	cupcakes	cheesecake	gelato	crepe	desserts	vanilla	donut	strawberry	cupcake	caramel	crepes	banana
30	ordered	tasted	seemed	cooked	order	pieces	plate	meal	soggy	served	plastic	stars	tasteless	burnt	quality
31	quality	prices	stars	average	places	overall	less	higher	value	priced	extra	staff	portions	paying	portion
32	tea	milk	iced	starbucks	cafe	latte	boba	staff	order	espresso	ordered	teas	bubble	wifi	places
33	vegas	las	strip	casino	hotel	located	bellagio	visit	view	restaurants	places	staying	casinos	planet	caesars
34	beers	selection	atmosphere	tap	pub	bartender	craft	staff	tvs	patio	bartenders	draft	irish	burgers	prices
35	und	der	das	ist	ich	nicht	war	auch	mit	sehr	ein	wir	den	aber	auf
36	company	business	customer	manager	bill	charged	account	bank	insurance	received	management	payment	paying	fees	contract
37	hotel	stayed	desk	bathroom	staff	staying	strip	resort	view	hotels	casino	suite	lobby	beds	vegas
38	gym	airport	classes	staff	yoga	training	machines	equipment	membership	instructors	facility	instructor	students	terminal	machine
39	dining	ingredients	quality	certainly	served	simple	fact	visit	truly	youll	despite	upon	although	slightly	less
40	theyre	review	literally	fact	youll	walking	apparently	stars	bathroom	entire	noticed	realized	stomach	heres	less
41	thai	spicy	ramen	curry	rice	korean	ordered	noodles	dishes	dish	restaurant	beef	order	tofu	broth
42	pho	rice	beef	noodles	restaurant	ordered	dishes	dish	shrimp	noodle	order	dumplings	broth	asian	vietnamese
43	company	professional	cleaning	highly	unit	cleaned	showed	crew	repair	gave	quote	maintenance	business	customer	companies
44	sandwich	cheese	burgers	sandwiches	ordered	bacon	order	onion	beef	bun	grilled	rings	onions	poutine	potato
45	reviews	review	yelp	stars	based	visit	business	rating	owner	previous	gave	update	fact	mentioned	places
46	tour	museum	parking	visit	view	trail	walking	garden	hike	group	areas	views	plenty	exhibit	canyon
47	seats	theater	stage	venue	cirque	vegas	theatre	movies	concert	row	seating	audience	performance	entertainment	comedy
48	shrimp	lobster	crab	ordered	seafood	cooked	meal	filet	oysters	restaurant	rib	steaks	server	salad	rare
49	customer	manager	customers	employees	order	staff	business	owner	employee	attitude	location	management	cashier	gave	treated
50	staff	owner	highly	extremely	owners	tattoo	absolutely	vet	prices	professional	visit	truly	customer	animals	pets

Figure 1: Topics for 4736897 user reviews

We preprocessed the *reviews.json* file which included **4736897 user reviews**. Based on this data, we built the model and performed topic modeling on the user reviews. We trained to model to get top 50 topics based on the size of the user reviews. Above are the top 50 topics we received for all the user reviews.

In the above we can see there are topics related to other businesses like automobile, theater , hiking trips, shopping, etc., We further performed few data processing to extract restaurants data as mentioned in the above section. The topic modeling for only the restaurant reviews **3216548 user reviews** are as below:

Topic	Words														
1	beers	selection	tap	bartender	pub	atmosphere	craft	bartenders	draft	cocktails	irish	brewery	staff	tvs	ale
2	pizza	cheese	pizzas	toppings	thin	slice	ordered	order	pepperoni	slices	sausage	delivery	style	oven	dough
3	tacos	taco	chips	burrito	guacamole	carne	asada	cheese	margaritas	tortilla	rice	nachos	burritos	tortillas	ordered
4	beef	dish	rice	dishes	noodles	noodle	dumplings	restaurant	shrimp	duck	spicy	asian	ordered	steamed	crispy
5	location	parking	mall	locations	located	airport	staff	shopping	strip	opened	north	plaza	building	scottsdale	phoenix
6	sushi	tuna	salmon	sashimi	rice	spicy	ayce	tempura	quality	chef	shrimp	miso	order	restaurant	sake
7	reviews	yelp	stars	review	ordered	order	based	gave	closed	ended	seemed	rating	yelpers	fact	upon
8	staff	atmosphere	prices	restaurant	portions	decor	casual	selection	priced	tasty	dining	seating	overall	highly	diner
9	starbucks	cafe	latte	staff	espresso	wifi	location	seating	iced	atmosphere	mocha	barista	milk	shops	tables
10	restaurant	server	group	reservation	staff	reservations	seated	chef	dining	event	waiter	manager	meal	guests	hostess
11	staff	restaurant	owner	atmosphere	customer	highly	absolutely	extremely	visit	server	owners	business	quality	customers	welcoming
12	tea	milk	boba	iced	smoothie	bubble	teas	ordered	smoothies	places	tasted	order	lemonade	matcha	taro
13	thai	pho	curry	spicy	rice	beef	restaurant	noodles	dish	vietnamese	ordered	dishes	broth	shrimp	spice
14	und	der	das	ist	war	nicht	ich	sehr	mit	auch	wir	ein	aber	den	essen
15	vegas	las	strip	restaurant	hotel	casino	located	prices	restaurants	visit	mall	places	staying	view	location
16	theyre	order	youll	extra	meal	places	ingredients	quality	add	less	bang	options	fact	buck	amount
17	restaurant	restaurants	phoenix	places	pittsburgh	cleveland	business	scottsdale	chain	north	chicago	visit	valley	youll	visiting
18	order	rice	ordered	delivery	beef	takeout	shrimp	restaurant	meal	portions	extra	veggies	ordering	spicy	general
19	quality	prices	stars	average	restaurant	overall	places	portions	portion	higher	restaurants	less	value	rating	priced
20	selection	market	prices	grocery	foods	produce	shopping	products	section	organic	quality	staff	meats	joes	etc
21	patio	seating	tables	outdoor	atmosphere	crowd	plenty	decor	tvs	chairs	view	parking	group	vibe	upstairs
22	staff	restaurant	german	tasty	absolutely	places	polish	def	serve	schnitzel	location	potato	style	pretzel	resturant
23	options	vegan	vegetarian	gluten	veggie	staff	ingredients	selection	tasty	restaurant	glutenfree	atmosphere	los	tofu	vegetarians
24	tasted	ordered	cooked	plate	soggy	pieces	seemed	salty	tasteless	burnt	overcooked	chips	meal	served	completely
25	decor	dining	walls	interior	tables	served	edinburgh	upon	certainly	glass	although	fact	lighting	slightly	despite
26	greek	pita	hummus	salad	gyro	lamb	plate	restaurant	falafel	tasty	rice	shawarma	mediterranean	meal	ordered
27	order	customer	location	ordered	employees	manager	customers	waited	cashier	staff	ordering	gave	employee	delivery	attitude
28	visit	review	restaurant	stars	reviews	several	visits	meal	weve	staff	previous	location	recent	based	quality
29	hotel	bathroom	staff	stayed	staying	view	casino	vegas	parking	desk	bathrooms	walking	resort	lounge	restroom
30	lobster	cooked	ordered	filet	potatoes	meal	rib	steaks	crab	shrimp	rare	sides	restaurant	salad	dessert
31	buffet	crab	seafood	selection	vegas	buffets	oysters	station	shrimp	quality	dessert	desserts	rib	dishes	section
32	rice	restaurant	hawaiian	staff	order	tasty	plate	portions	meal	style	prices	ordered	highly	beef	visit
33	dish	restaurant	dessert	meal	duck	dishes	chef	dining	gras	foie	tasting	served	beef	cooked	lamb
34	montreal	poutine	crepe	crepes	pastries	cheese	toronto	croissant	bakery	breakfast	cafe	pastry	sandwiches	staff	visit
35	cake	cupcakes	cheesecake	donut	cupcake	yogurt	bakery	velvet	frosting	moist	dessert	dozen	tasted	toppings	icing
36	burgers	cheese	onion	bacon	bun	ordered	rings	order	chili	patty	beef	potato	cooked	onions	juicy
37	indian	restaurant	dishes	spicy	naan	rice	curry	dish	lamb	ordered	buffet	masala	spice	restaurants	tasty
38	manager	customer	business	owner	customers	coupon	review	tip	groupon	bill	gave	management	charged	restaurant	employees
39	ramen	korean	spicy	noodles	broth	rice	dishes	beef	restaurant	ordered	kimchi	bbq	tofu	dish	miso
40	sandwich	sandwiches	cheese	beef	sub	ordered	deli	order	grilled	philly	chips	pastrami	subway	roast	subs
41	bbq	brisket	pulled	cheese	sides	ordered	potato	spicy	sauces	slaw	meal	shrimp	sandwich	juicy	order
42	dessert	gelato	vanilla	caramel	banana	waffle	desserts	strawberry	pudding	creamy	scoop	custard	brownie	coconut	flavours
43	salad	dressing	ordered	lettuce	salads	meal	caesar	cheese	grilled	order	greens	tasty	potato	salmon	tomatoes
44	ordered	server	order	waitress	waiter	bill	manager	meal	waited	restaurant	seated	gave	tip	received	offered
45	tables	restaurant	server	seated	waitress	hostess	order	waited	servers	staff	seemed	customers	plates	bartender	greeted
46	breakfast	eggs	brunch	toast	bacon	pancakes	hash	potatoes	ordered	waffles	sausage	benedict	sunday	gravy	omelet
47	est	pas	des	une	mais	trs	cest	qui	avec	bon	dans	sont	vous	nous	jai
48	pasta	restaurant	dish	ordered	salad	meal	meatballs	spaghetti	olive	dishes	calamari	ravioli	veal	cheese	appetizer
49	cheese	bacon	grilled	served	onions	tomato	roasted	dish	crispy	topped	goat	peppers	mushrooms	tomatoes	creamy
50	ordered	server	appetizer	restaurant	appetizers	seated	order	waitress	group	meal	shrimp	entrees	overall	tapas	atmosphere

Figure 2: Topics for 3216548 user restaurant reviews

After analyzing the topic words we have identified the aspects on which the users review a restaurant. The users have mainly reviewed based on the type of cuisine. For example, based on the topics we could derive the below cuisines and type of food the users have reviewed. Also, the other aspects of the restaurant that indicate user preferences are as below:

Sl. No.	Food Type
1	Bar/Pub
2	Italian
3	Mexican
4	Asian
5	Japanese
6	Thai
7	Bubble Tea
8	café
9	German
10	French
11	European
12	Indain
13	Vegan
14	Breakfast
15	Dessert
16	Korean
17	Seafood
18	Fast food
19	Greek
20	Mediterranean
21	Italian

Sl. No.	Review Aspects
1	Ambience
2	Service
3	Price
4	Time of Visit
5	Delivery
6	Taste of food

Figure 3: Resulting Cluster topics

3.1 Approach for LDA

We started with Scikit learn LDA package to perform topic modeling. We were able to process the topic models for 1000 reviews and understand the approach. When we tried executing the same code with some optimization, it consumed a lot of time to build the model. So, we tried other libraries which can handle performing LDA for large datasets and finalized on mallet. We chose mallet as it was simple to execute and it was Java based library and hence faster. It took 28 minutes build the entire model and 117 minutes to give us the topics after convergence.
Scikit learn Approach - Attached python notebook - TopicModeling.ipynb
Mallet - https://programminghistorian.org/lessons/topic-modeling-and-mallet

```
cd mallet.2.0.8
./bin/mallet import-dir --input review_docs/
        --output reviewModel --keep-sequence
./bin/mallet train-topics --input reviewModel
        --num-topics 50 --output-topic-keys keys.txt
```

4 Next Steps

4.1 Agglomerative Clustering and K-Means Clustering

On analyzing the topics obtained after performing LDA, we have realized that agglomerative hierarchical clustering might be more suitable to cluster the reviews because of the nature of one of the topics (specifically Food). We would like to go ahead with this approach and compare the clusters formed from both agglomerative clustering and k-means clustering algorithms to be able to make a better decision.

4.2 Ranking Business

To recommend a business to a user, we would have to rank all the businesses present in a cluster so that we can aim to recommend a high rated restaurant. For this, can rank all the business that correspond to the reviews in each cluster using the star ratings that accompany each review.

5 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

6 Appendix

1. Graphs and Analysis - attached python notebook
2. Source code - Github and attached python notebook