Mining Yelp Review Data

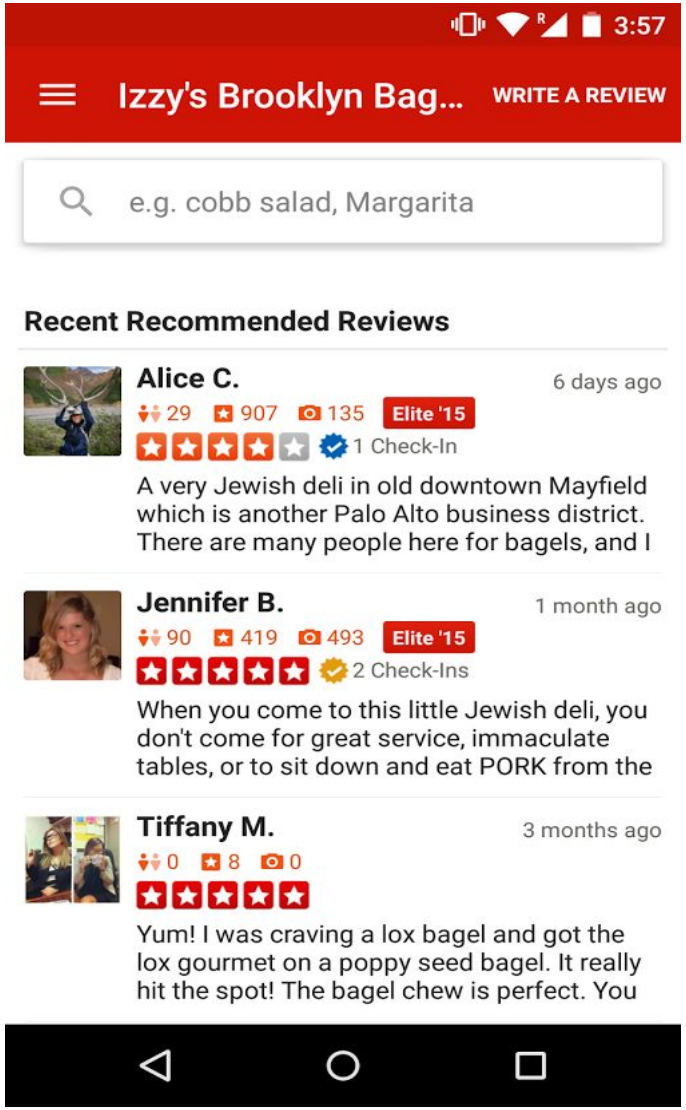
From latent ratings to a recommendation system

Rashmi Dwaraka, Ritvika R. Nagula, Ruiyang Xu, Deepak Surana

CS 6220 - Fall ‘17

Northeastern University

Introduction - Yelp Dataset Challenge



Dataset Overview

REVIEW OBJECT

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

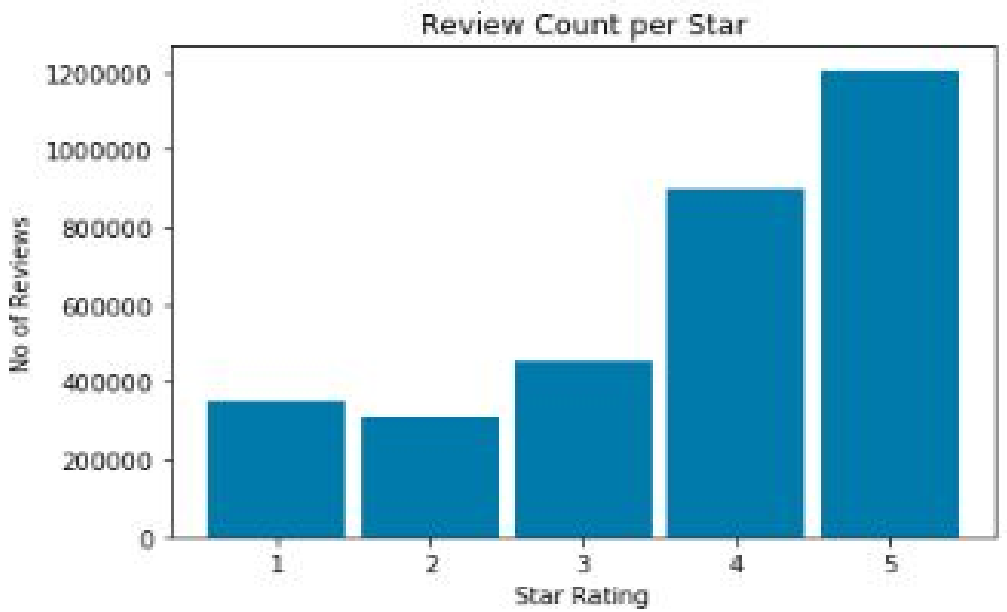
RESTAURANT OBJECT

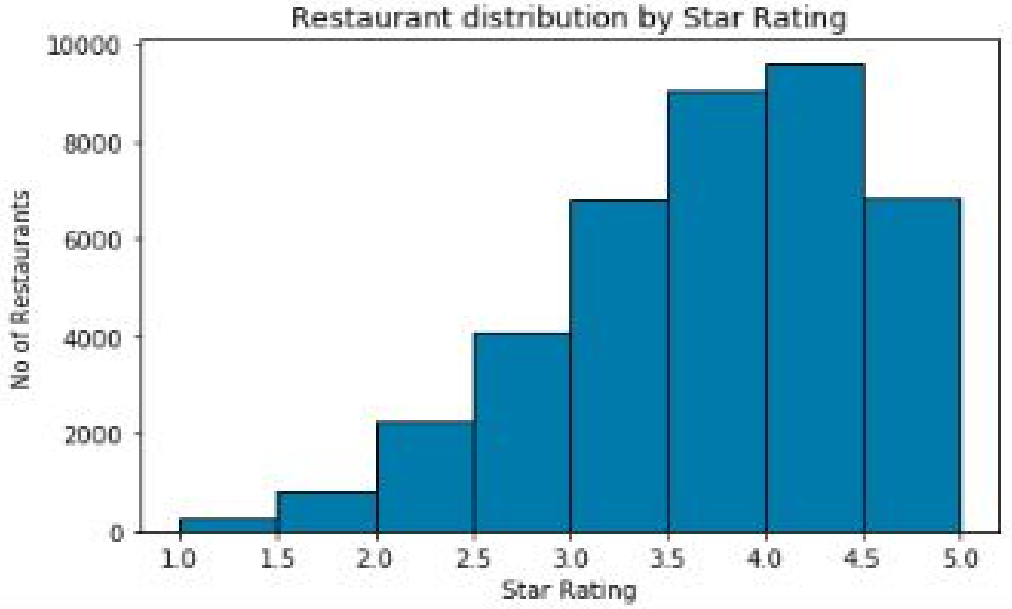
Field	Description
business_id	unique business id
name	business' name
neighborhood	neighborhood's name
address	full address
city	city
state	2 character state code
postal code	postal code
latitude	latitude
longitude	longitude
stars	star-rating
review_count	number of reviews
is_open	0 or 1 for closed or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

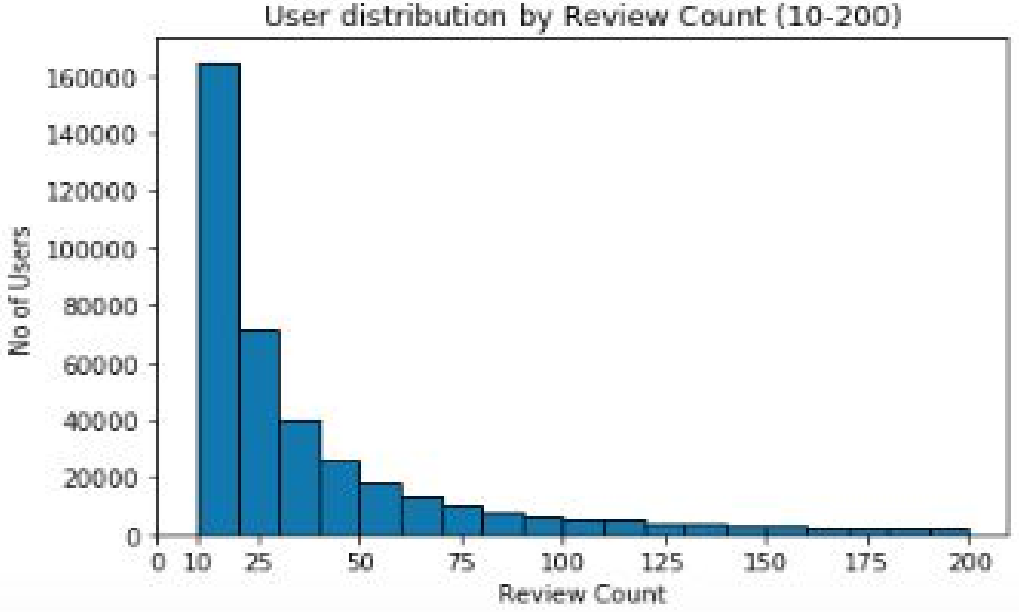
Exploratory Data Analysis

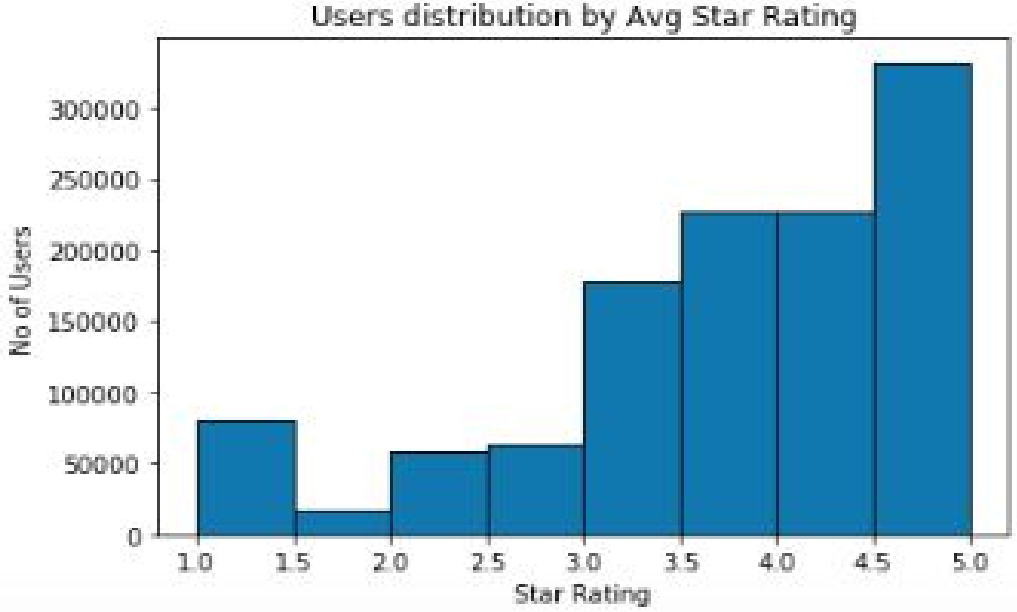
- We are interested only in restaurant data, so we performed filters on the whole dataset to get restaurant reviews only.
  - 65K Restaurants
  - 3.2M Reviews
  - 1.1M Users

Distribution Plots:









Word Clouds:


5-Star Reviews
ice cream
every time
great service
Las Vega
food service
one favorite
really good
definitely back
recommend place
mac cheese
great place
good food
go back
can't wait
even though
food delicious
great food
staff friendly
love place
highly recommend
good service
service good
pretty good
first time
make sure
feel like
happy hour
friendly staff
food great
one best
place great
would definitely
However
coming back
come back
next time
customer service
I've ever
food good
service great


1-Star Reviews
maybe
drive thru
didn't even
guess
customer service
tasted like
going back
took minute
minutes later
looked like
also
go
don't know
that
another minute
come back
ice cream
even though
either
mean
waited minute
take order
actually
go back
told us
well
left
top
get food
last time
since
came back
However
first time
food good
point
never
instead
one star
coming back
I've ever

Problems with the raw ratings

- User gave ratings based on certain latent criteria.
- Raw ratings can’t reflect those criteria.
- BUT user also gave text reviews! Lots of information can be mined from those text reviews.
- Need a way to extract useful information from text.

Topic Modeling and LDA

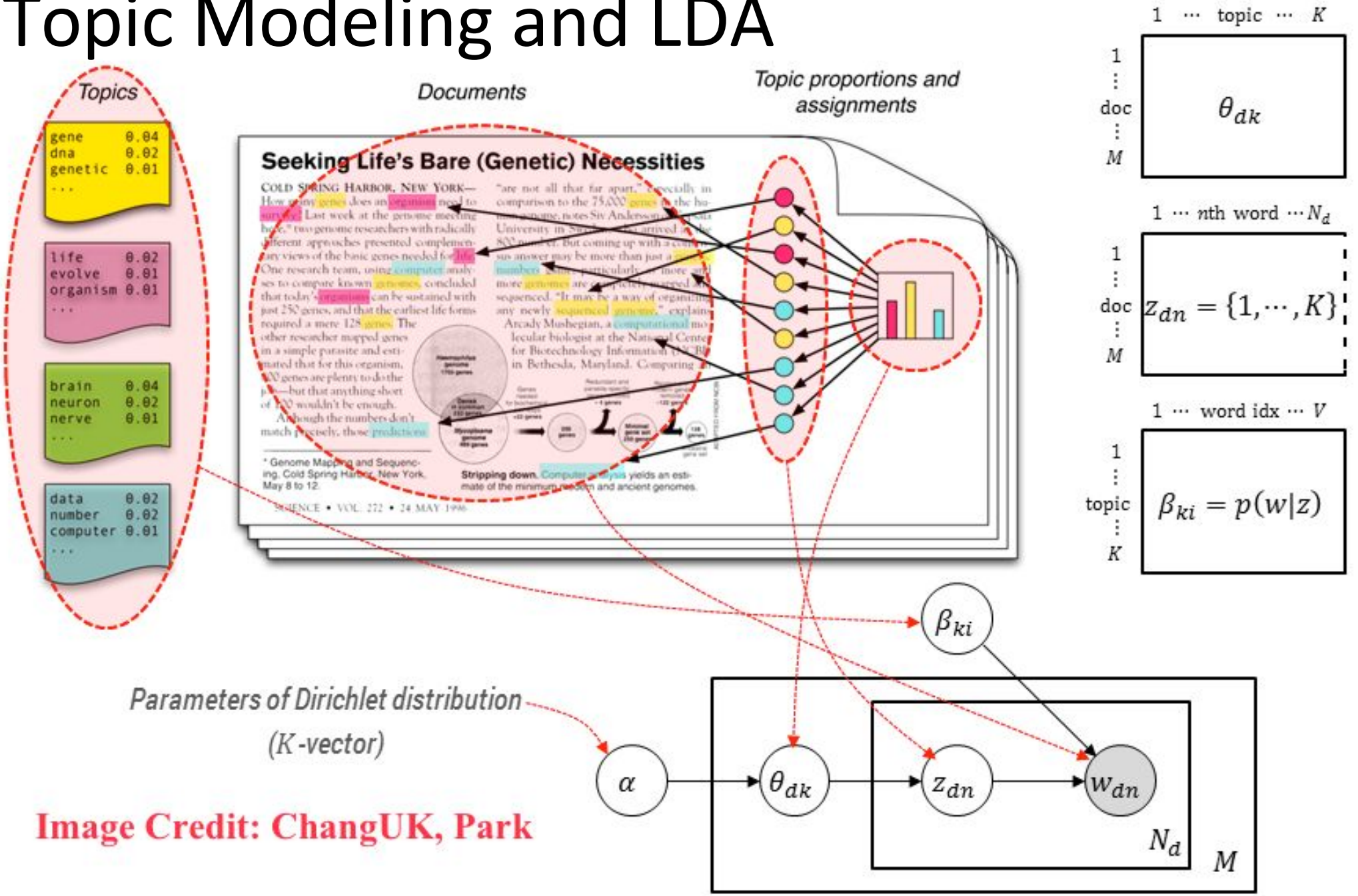


Image Credit: ChangUK, Park

Latent criteria found by LDA

Categories are assigned manually based on top ranked significant words in each topic.

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

Calculating Latent ratings from LDA

- Only first two top ranked topic in each review is significant.
- Average over all relevant reviews and get the rating per latent feature.

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

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 $Sig(r, t_j) = 0$.

Restaurant-Latent Feature matrix

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
lrPNLw0zbfzgWrWlNMjsNw	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
....									

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
....					

Recommendation System

- Ask the user to rank the criteria (features) on a scale of 1-10 for each feature.
- Rank the Restaurant-Latent Feature matrix according to the given criteria.
- Returned the top ranked restaurant back to the user.

Conclusion