How to Improve Your Restaurant?

Lize Du, Linquan Ma, Wenjia Xie

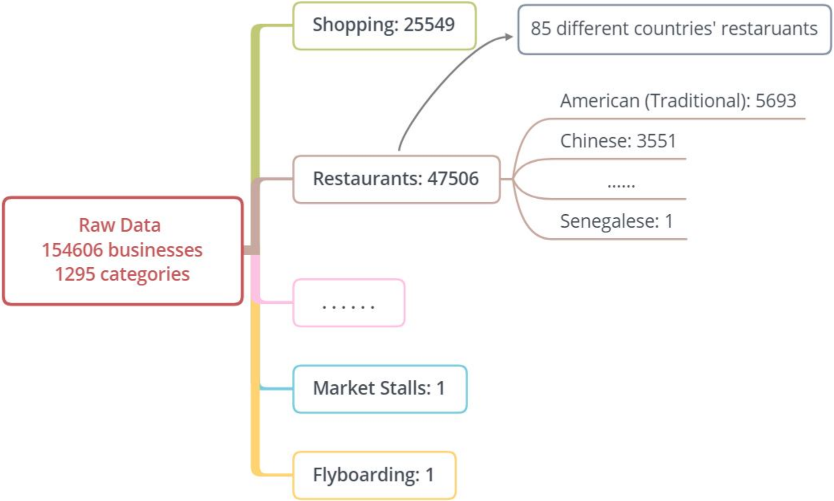
STAT 628

March 4, 2019

Outline

- Data cleaning
- Deal with Review Dataset
- Suggestions for Restaurants
  - ✓ Some Informative Words
  - ✓ Word Clouds
  - ✓ Geography
- Future Work

Data Cleaning: Business Dataset



Data Cleaning: Review

- Stopwords use *nltk.stopwords.words* except:
  1. Third person pronoun: he, she, it, they, their
  2. Adverb of degree: few, most, more...
  3. Negative: don't, didn't, doesn't aren't...
- Pattern matching: words, abbreviation, [a-zA-Z]-[a-zA-Z], ... , ?, !
- Substitute: he's→he is, n'/n't→not, 'd→would...
- Delete: noun's, number+th/st/nd/rd;
- Change to lower case;
- Tokenize using regular expression;
- Add _neg to the words between not/never and the first punctuation;
- Use porter stemmer to do stem extracting, such as amazing→amaz;
- Use wordnet lemmatizer to lemmstize the verb to a normal form, such as loving→love

Deal with Review Dataset

All review: 3.43 GB
American restaurants' review: 503 MB; 845,941 rows
(*grep* command in bash)

Dictionary size: 245,344 words

1. For the first part, just focus on some top words based on frequency, so only contains the most frequent 461 words.
2. Count every word's frequency in every star (even if it appears many times in one review, we just focus on if it appears)
3. Use information gain of each word to rank them.

How to Define Information Gain?

$$\text{Information gain} = H(Y) - H(Y|X)$$

where Y denotes class (star level), X is feature (word).

For example, the proportion of each star in whole dataset is P_i, i = 1, 2, 3, 4, 5.

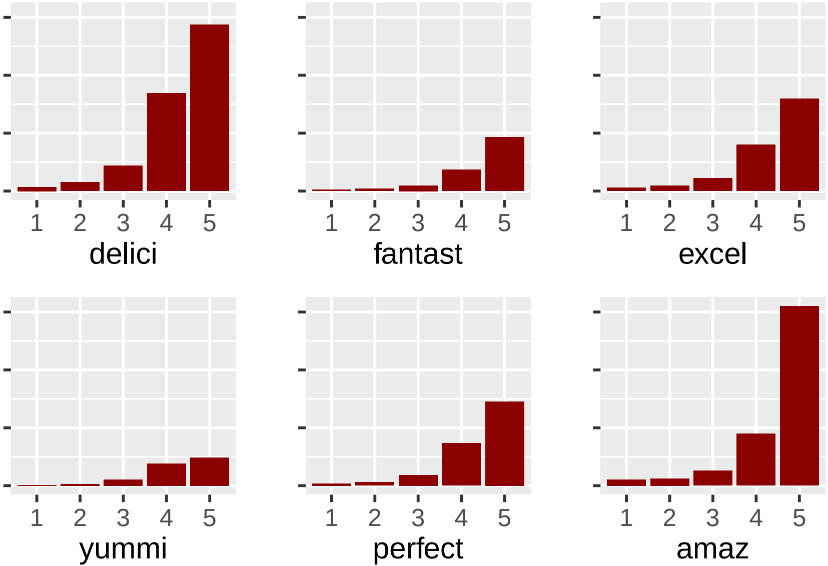
$$H(Y) = -\sum_{i=1}^{5} P_i log_2 P_i$$

if we specify x as "good",

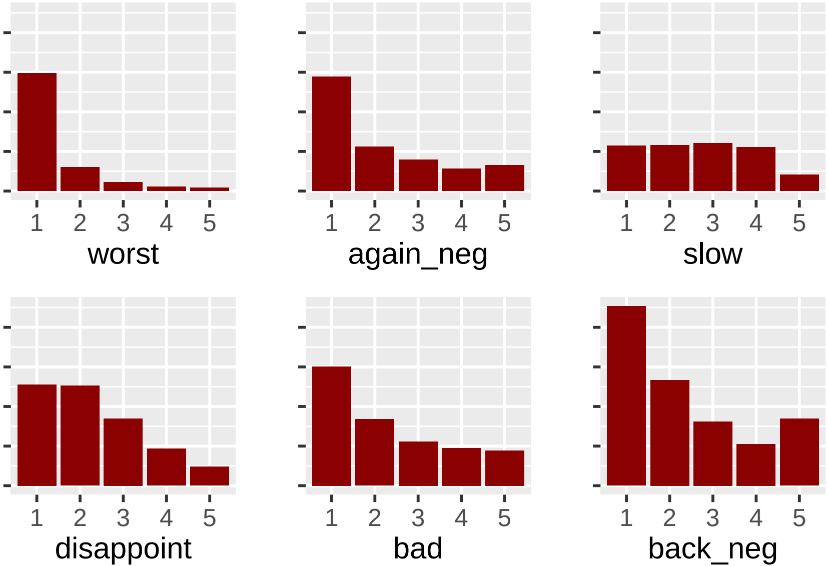
$$H(Y|X) = \sum_{i=1}^{2} P(X = x_i)(-\sum_{j=1}^{5} P_{ij} log_2 P_{ij})$$

where $x_i = 0, 1$ (1 denotes review contains "good"), p_{ij} is proportion of star j when $X = x_i$

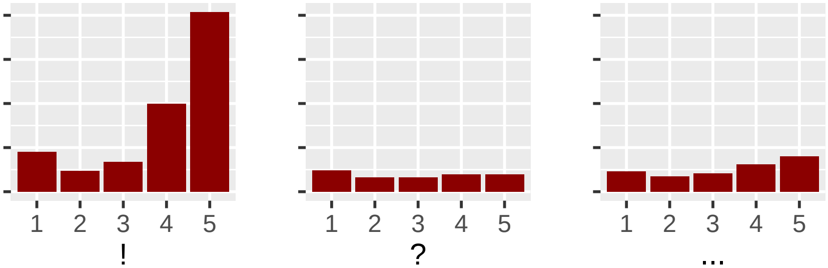
Informative Words (Positive)



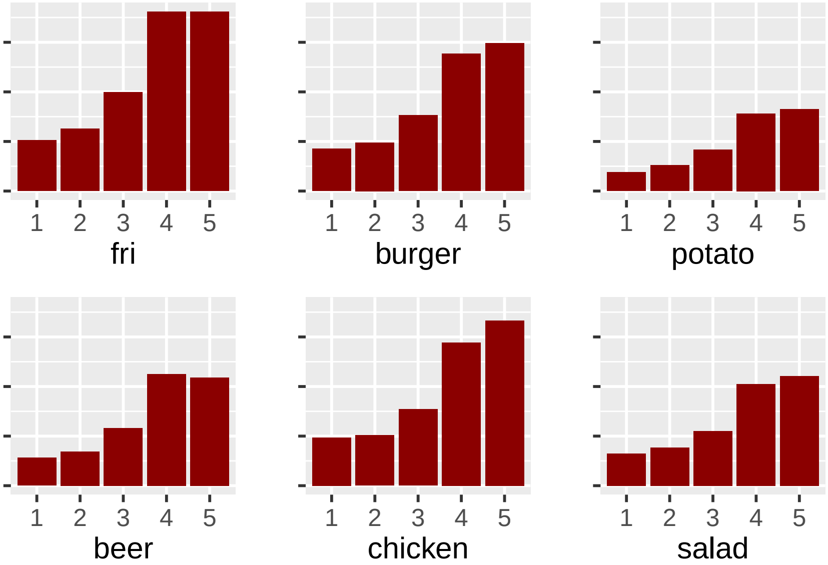
Informative Words (Negative)



Punctuation



Popular Foods



Word Clouds

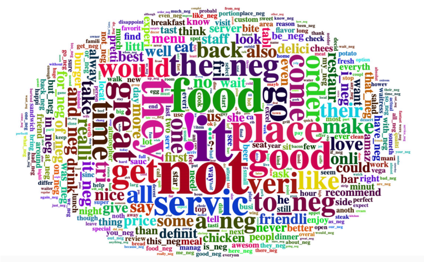


Figure: Rank by Word Frequency

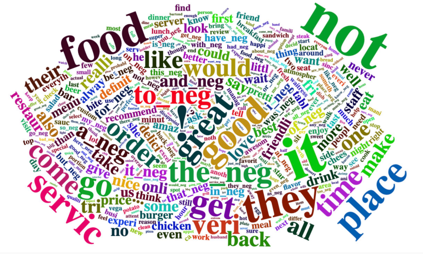
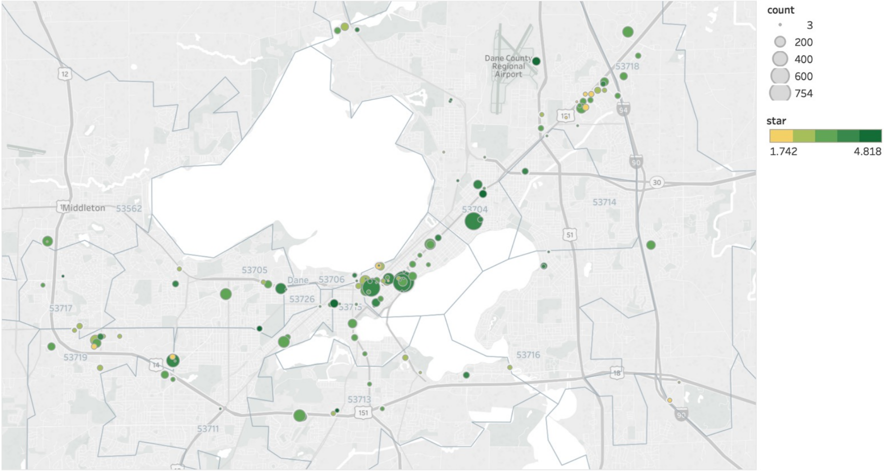


Figure: Rank by Information Gain

Traditional American Restaurants in Madison

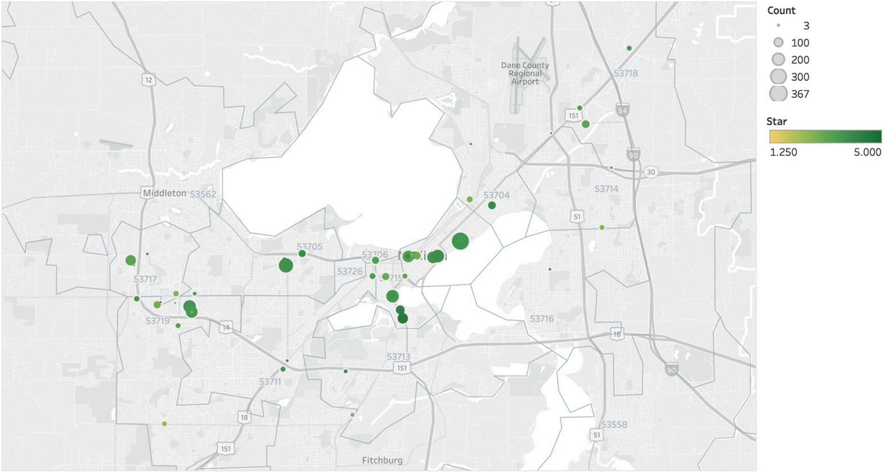
Traditional American food



Map based on longtitude and latitude. Color shows star. Size shows count.

Chinese Restaurants in Madison

Chinese food



Map based on Longtitude and Latitude. Color shows Star. Size shows Count.

Future Work

- Keep more words in final dictionary and give more specific suggestions for traditional American restaurants.
- Analyze tha data of more countries' restaurants and give some generalized suggestions.
- Star prediction: Linear regression, SVM, Bayes net etc.

not_neg
back_neg
today
when_neg
chang
toast
use
eat_neg
sweet
size
an_neg
lot
spot
hope
larg
long
absolut
do_neg
half
walk
order_neg
excel
fill
quit
fresh
fun
select
just_neg
might
front
were_neg
run
instead
again_neg
going_neg
me_neg
did_neg
hear
yummi
star
thank
still
talk
felt
pay
ask
plu
us_neg
ye
fine
slow
pm
disappoint
la
howev
mean
bill
dri
less
averag
els
roll
mayb
guess
leav
manag
up_neg
highli
oh
bad
job
saw
decent
minut
host
ok
okay
seem
word
money
tell
call
let
spici
later
tasti
cold
crispi
noth
final
piec
leave
if_neg
our_neg
fantast
although
perfectli
will_neg
town
he
late
season
someon
month
bite
music
sit
line
complet
least
put
amaz
water
patio
awesom
worth
ice
..._neg
service_neg
arriv
share
cool
favorit
care
notder
last
wine
them_neg
red
may
perfect
almost
their_neg
or_neg
light
total
how_neg
bbq
at_neg
away
do
stay