Yelp Data Analysis

Chao Chang, Shuyang Chen, Kunning Wang, Youhui Ye

STAT 628

October 31, 2019

Contents

Data Overview

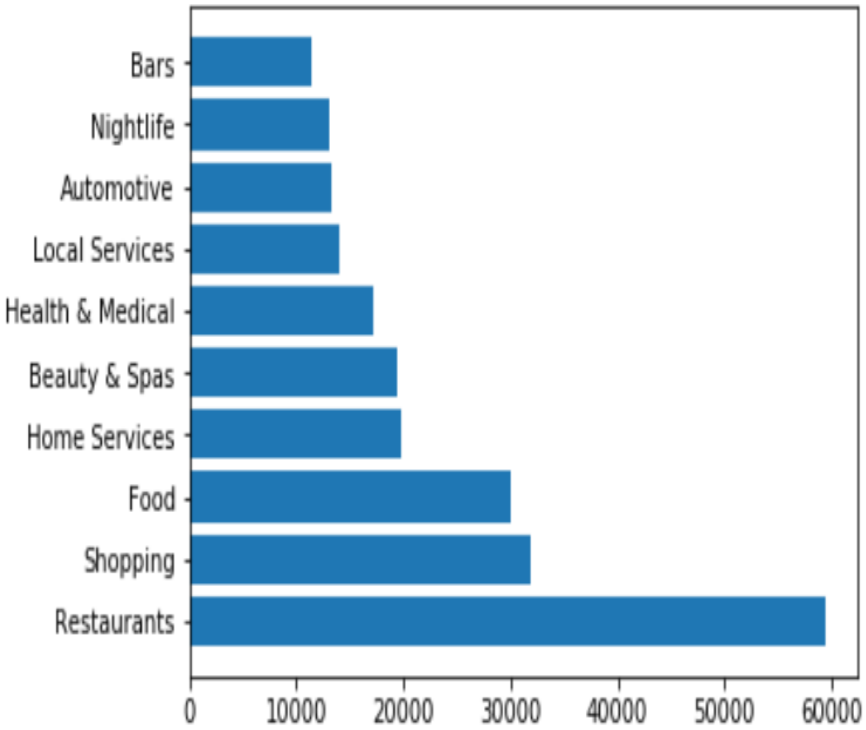


Figure: All Categories

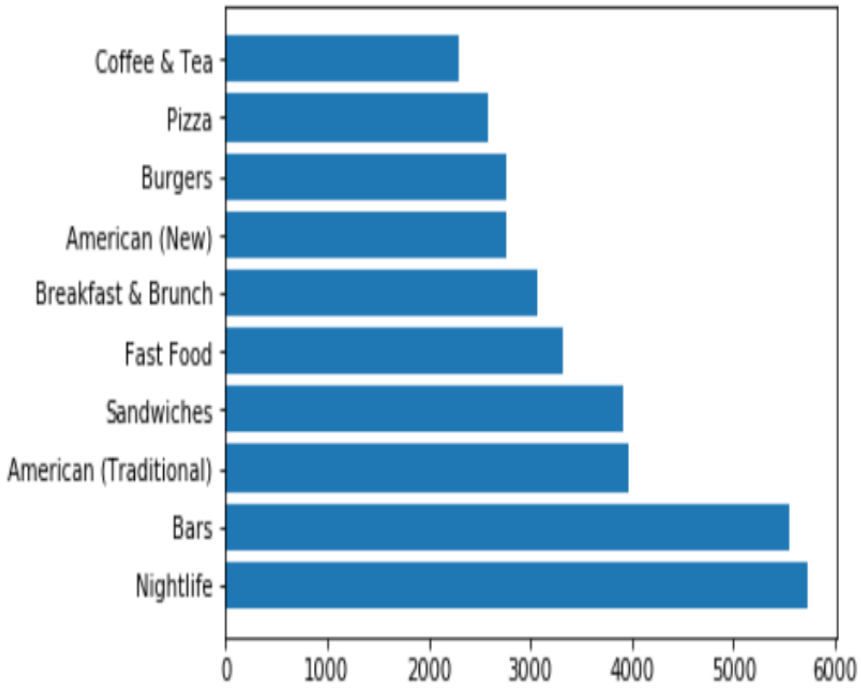
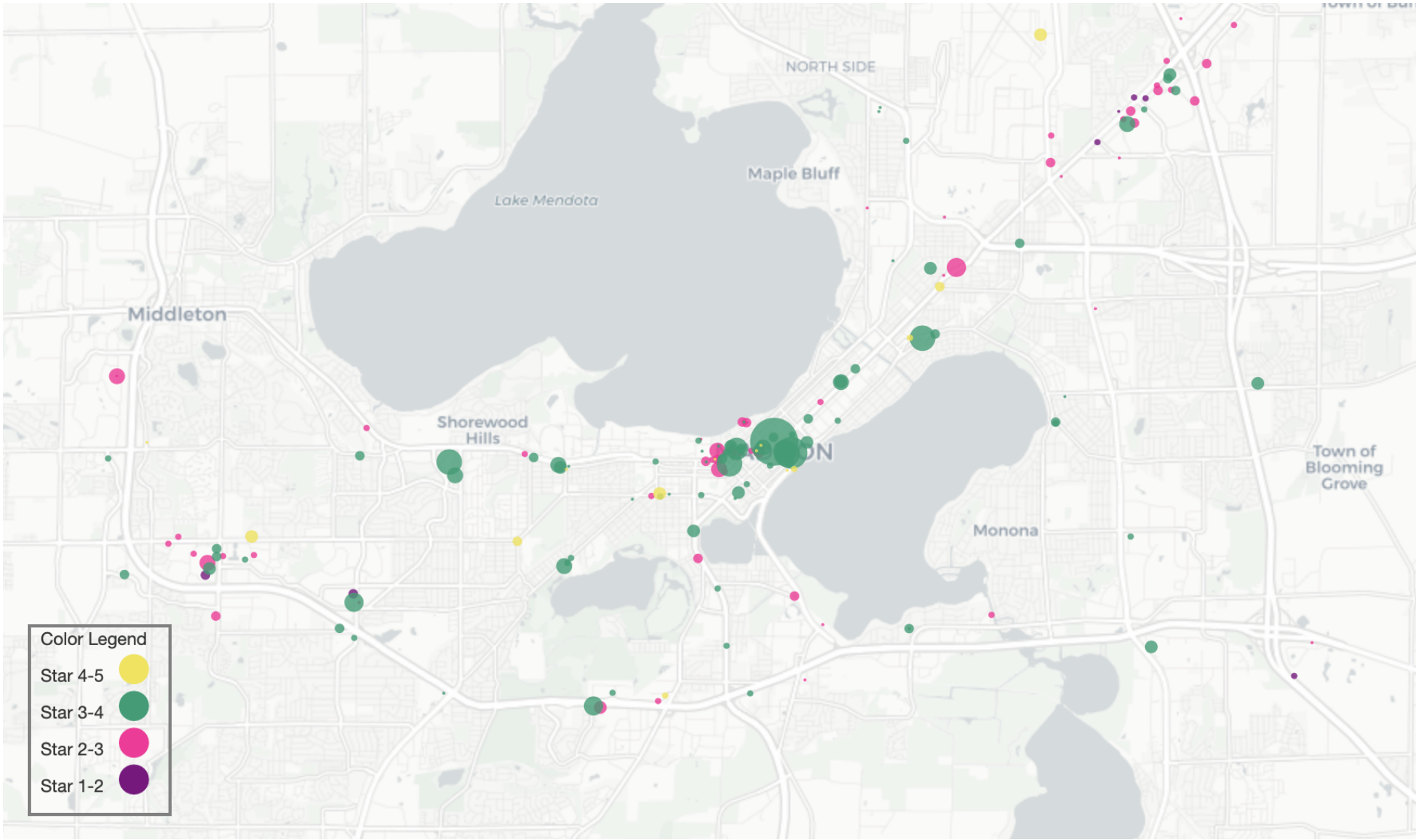


Figure: Restaurants

American Traditional Restaurants in Madison



Data Cleaning

Purpose

- Simplify and merge all data sets with interested features
- Convert a text to a vector

- Select recent reviews and interested features(name, review text, stars, locations and attributes)
- Deal with the review text

Dealing with the review text

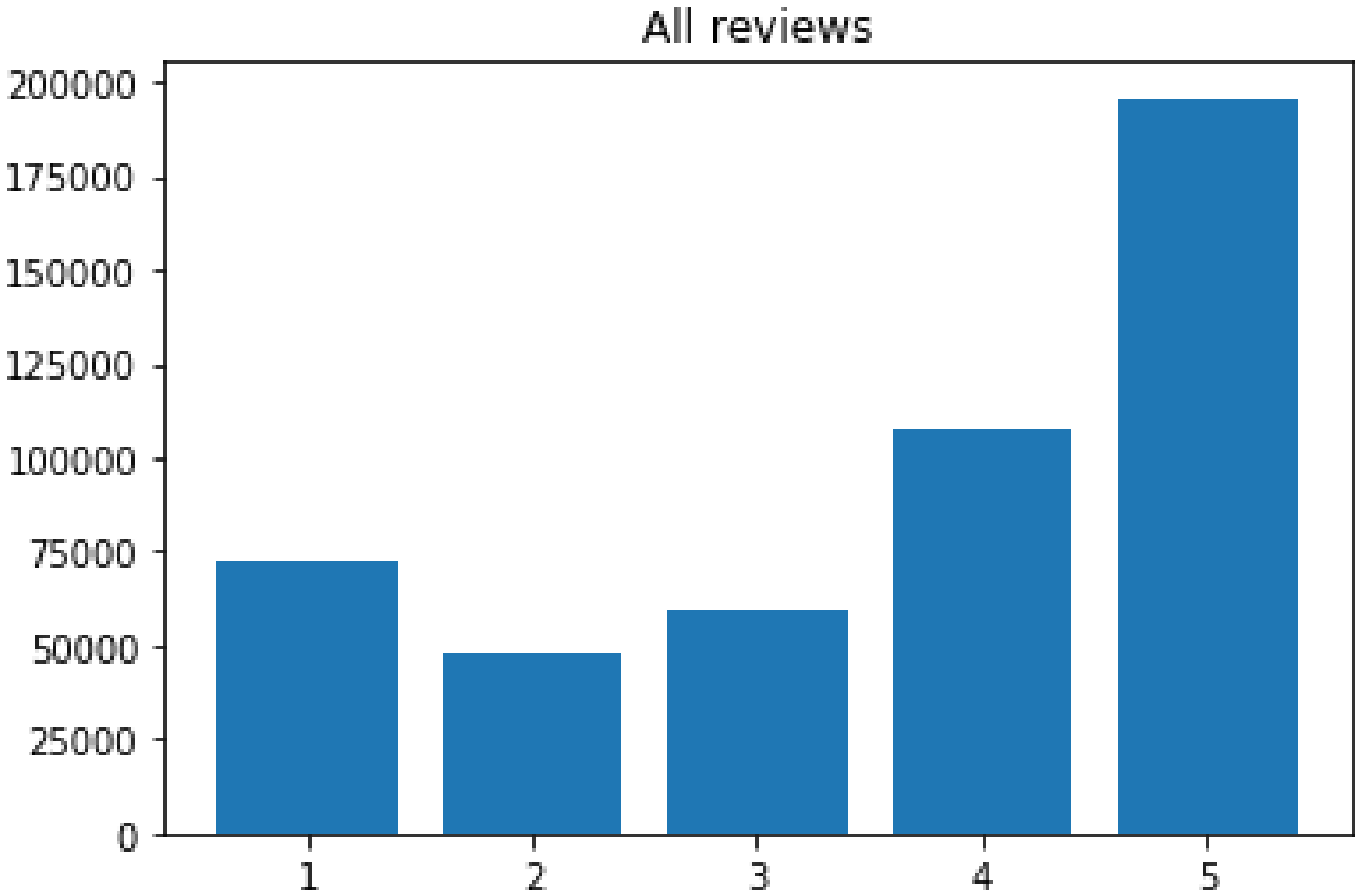
Original	Lowercase, tokenize and expand abbreviation	Remove stop words and add neg tags	Remove all punctuations
I don't like cheeseburger, though my kid does.	['i', 'not', 'like', 'cheeseburger', ',', 'though', 'my', 'kid', 'does', '.']	['not', 'neg_like', 'neg_cheeseburger', ',', 'though', 'kid', '.']	['not', 'neg_like', 'neg_cheeseburger', 'though', 'kid']

Table: Dealing Process

Word Cloud


food
good
great
place
service
like
time
back
neg_good
neg_but
neg_service
would
one
really
go
get
us
ordered
came
chicken
salad
even
always
nice
sauce
amazing
much
drinks
happy
made
flavor
restaurant
breakfast
asked
first
burger
staff
lunch
night
best
cheese
two
well
also
minutes
want
order
friendly
little
delicious
got
could
table
menu
bar
definitely
drink
eat
times
take
cooked
stars
way
sandwich
waitress
recommend
experience
location
right
nothing
bread
know
give
check
home
try
love
never
excellent
people
friends
side
steak
said
told
meal
bad
since
sweet
bacon
next
looking
tables
though
around
another
took
seated
server
wanted
say
tasty
pretty
still
dinner
small
enjoyed
customer
special
better
went
come
think
fries
ever
visit
beer
half
husband
favorite
atmosphere
something
awesome
served
arrived
tried
price
different
quality
enough
sure
full
big
large
wait
many
kind
first
found
water
potatoes
feel
decided
manager
dessert
day
reviews
selection
looked
coffee
pizza
prices
fresh
family
area
decent
top
Service
disappointed
restaurants
waiter
worth
attentive
hot
cold
fast
bit
far
last
spot
brunch
done
else
places
free
party
fried
servers
waiting
almost
sat
plate
years
wife
new
beers
going
burgers
outside
inside
everything
perfect
super
lot
ask
quite
fish
work
see
Vegas
every
beef
brought
point
wings
pancakes
Wonderful
Great
getting
high
bill
long
clean
used
Also
10
2
5
15
20

Distribution Plots



High frequency words

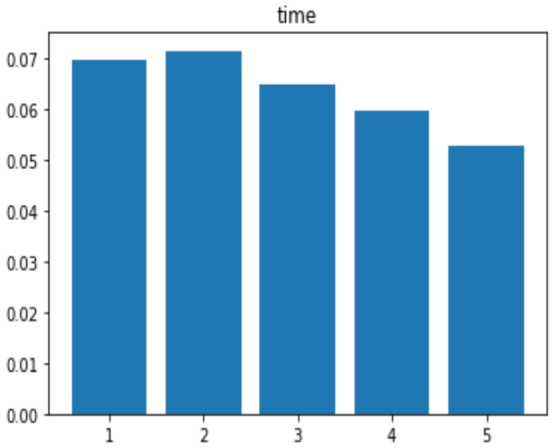


Figure: time

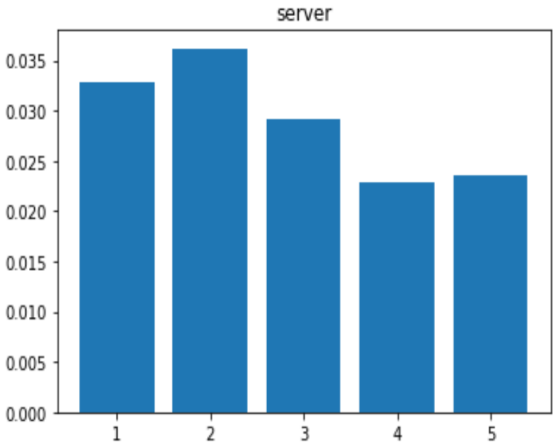


Figure: server

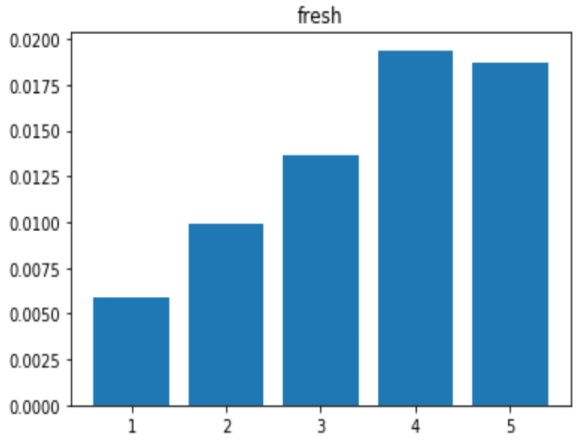


Figure: fresh

Different kinds of foods and drinks

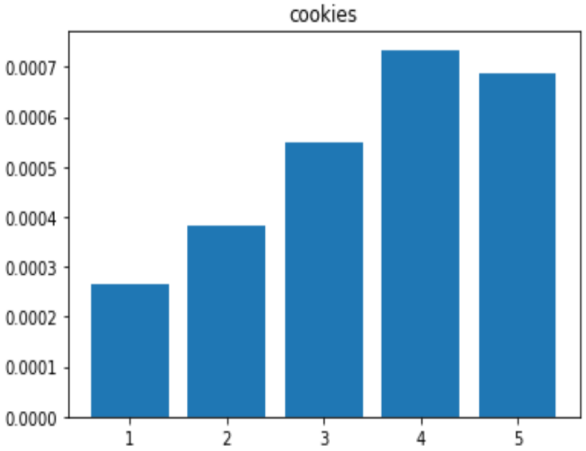


Figure: cookies

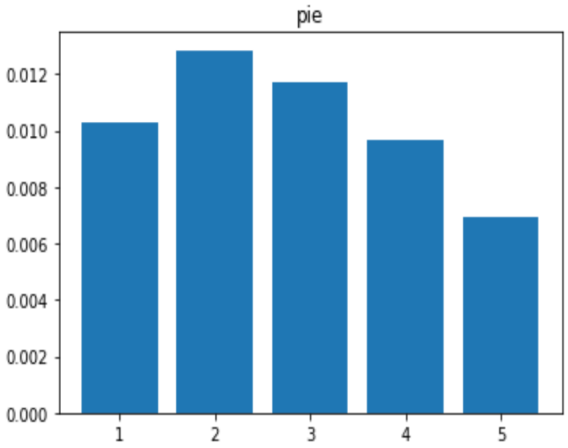


Figure: pies

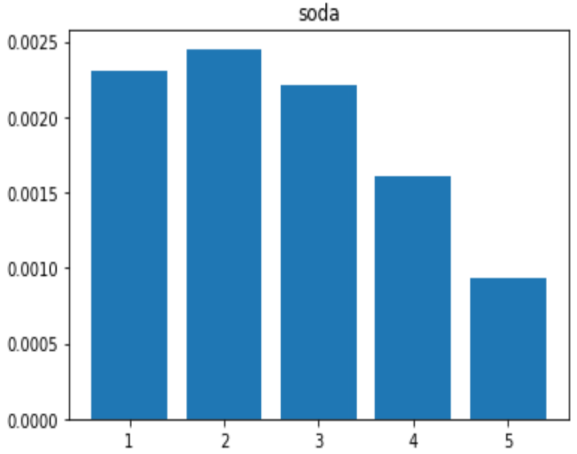


Figure: soda

Insights from the attributes

- **Common Attributes:** 10 most common attributes such as Goodforkids, Wifi, Alcohol.
- **Transform string label to integer label:**
  Eg: Outdoorseating: True to 1, None to 0, False to -1
- **Suggestion based on linear regression:** Having outdoor seats, wifi and delivery contributes to higher stars while allowing dogs and being noisy affect in a negative way.

Future Work

- **Extract noun phrases from the review text**
  "Credit Card" is more informative than "credit" or "card".
- **do stem extracting and lemmatization:**
  "cats"→"cat" and "loving"→"love"
- **Using tf-idf to select useful words**
- **Do linear regression and significance test**
- **Explore attributes and locations**
- **Give suggestions to American traditional restaurants**

Thank you