Final Presentation

How various features of East Asian Restaurants in CA influence their review stars?

Group 5
12/06/2022

Data preprocessing

Extract

1. Extract **East Asian Restaurants** (Keyword: Chinese, Korean, etc.) in CA
2. Include all original features (e.g. **business name**, **stars** and **attributes** like **TakeOut option**, **Parking availability, etc**).
3. Data size: 14825 * 87

Process

1. Remove permanently closed restaurant
2. Extract object data (e.g. hour, parking, etc)
3. Initialize new features (daily & weekly working hour)

Combine

1. Combine **business data** with **review data** using **'business_id'**;
2. Apply NLP strategies (e.g. remove stop words, punctuations, symbols; lemmatization)
3.Initialize sentiment (based on customer's review star)

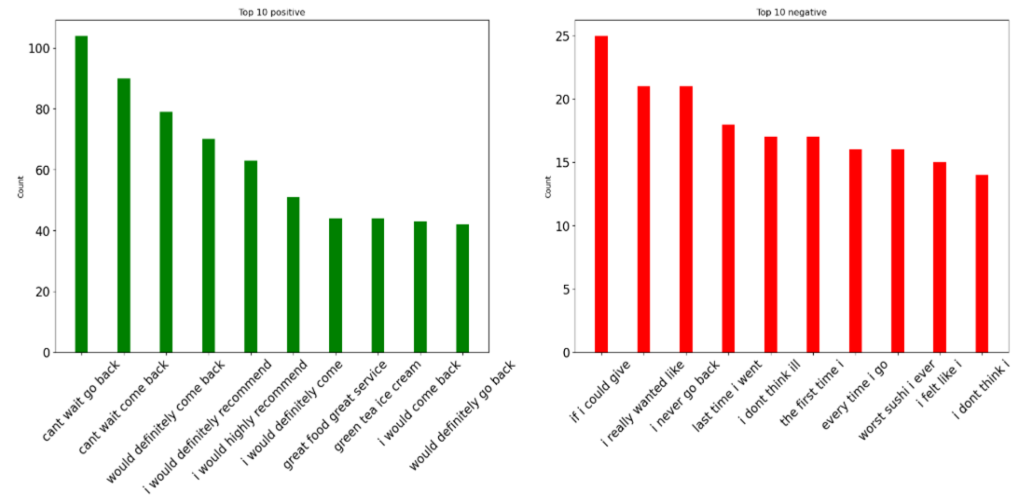
Polish

1. Unify type in mixed-type feature

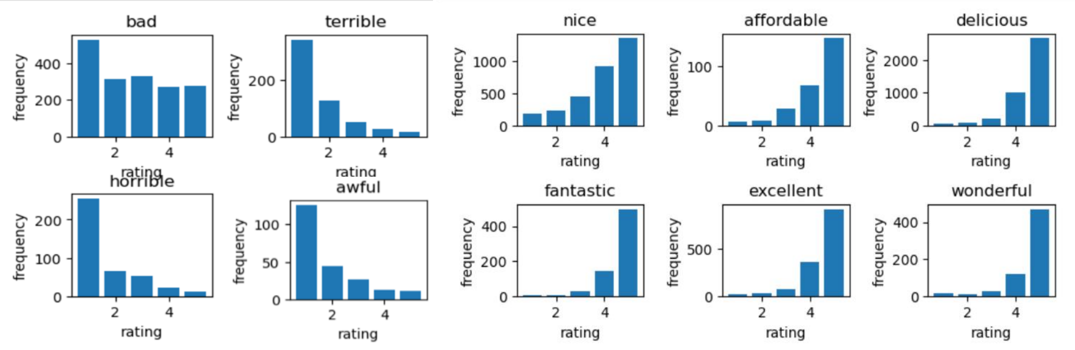
Preliminary Analysis Recap

1. Word cloud

2. N-gram text frequency



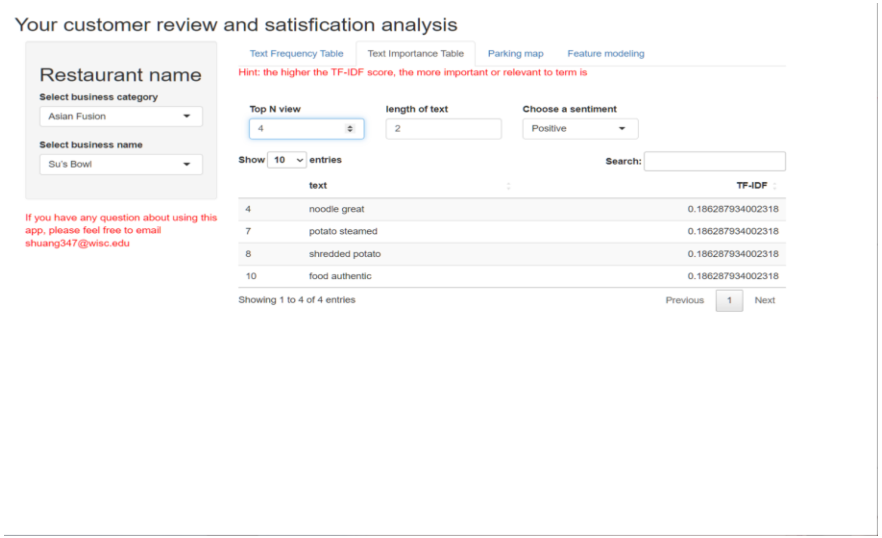
3. Word sanity check



Shiny App: Word Cloud View

Older version

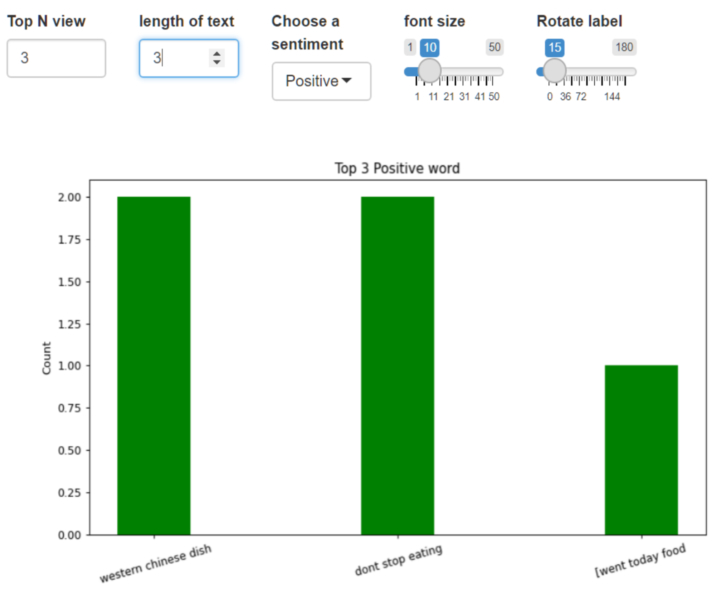
Newer version



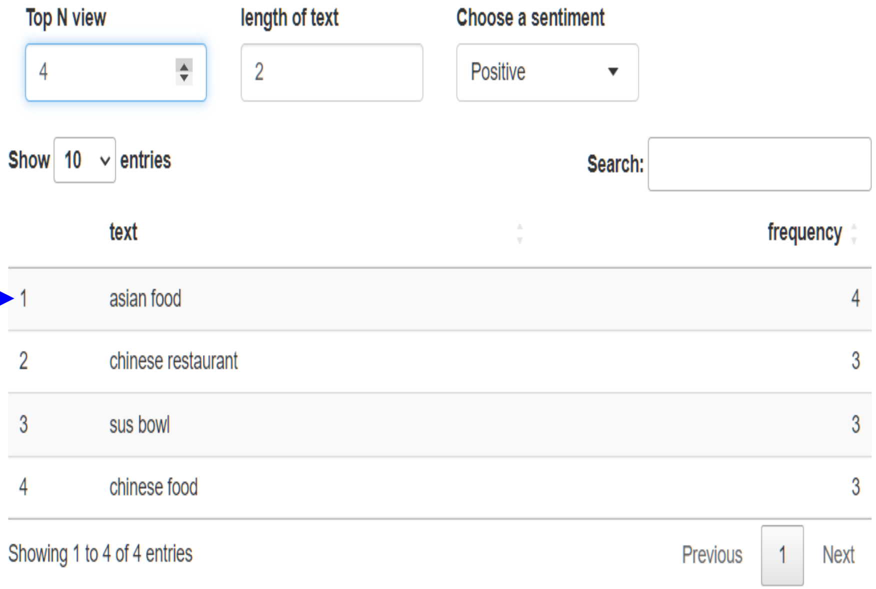
1. Due to python virtual environment on shinyapp.io, python plotting functions can’t be shown properly (word cloud)
2. Word cloud view was removed in newer version of shiny

Shiny App: Text Frequency view

Old version



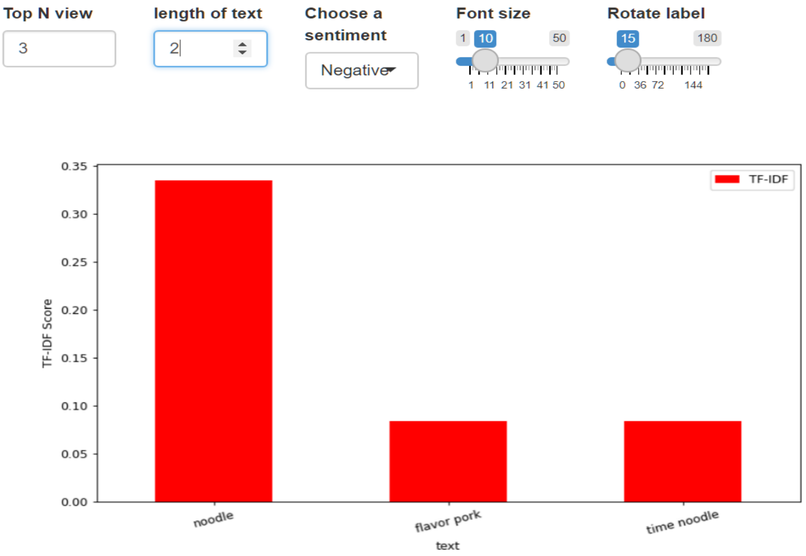
New version



	text	frequency
1	asian food	4
2	chinese restaurant	3
3	sus bowl	3
4	chinese food	3

Shiny App: Text Importance View

New version



New version

Top N view: 4

length of text: 2

Choose a sentiment: Positive

Show 10 entries

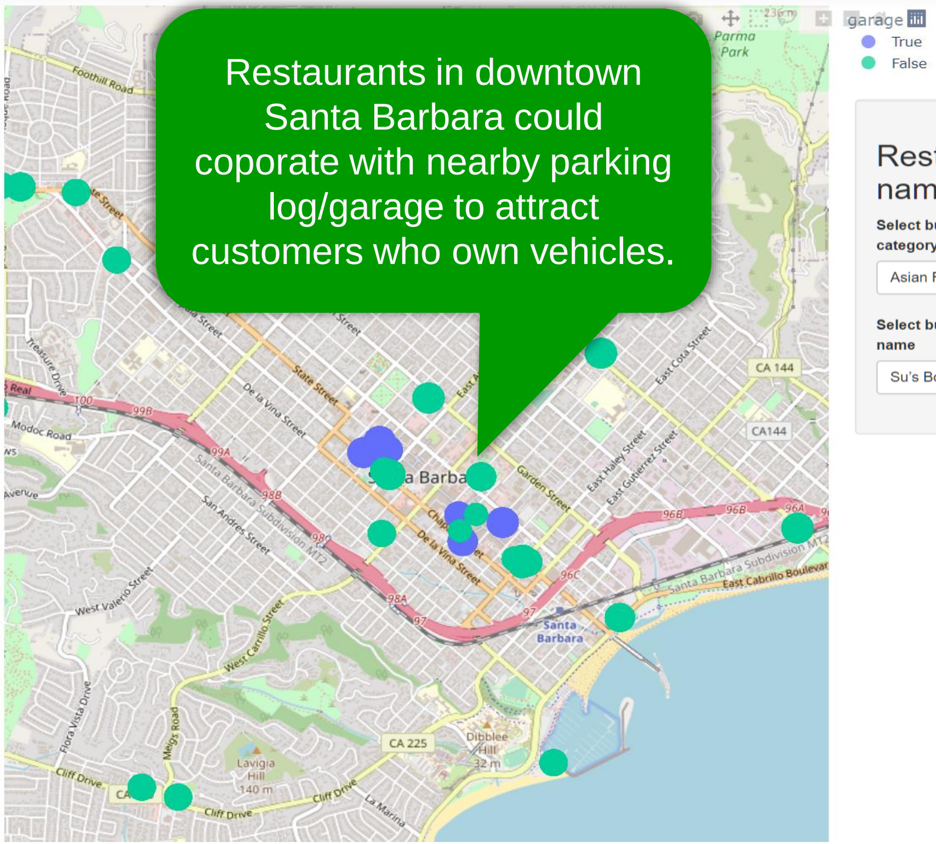
Search:

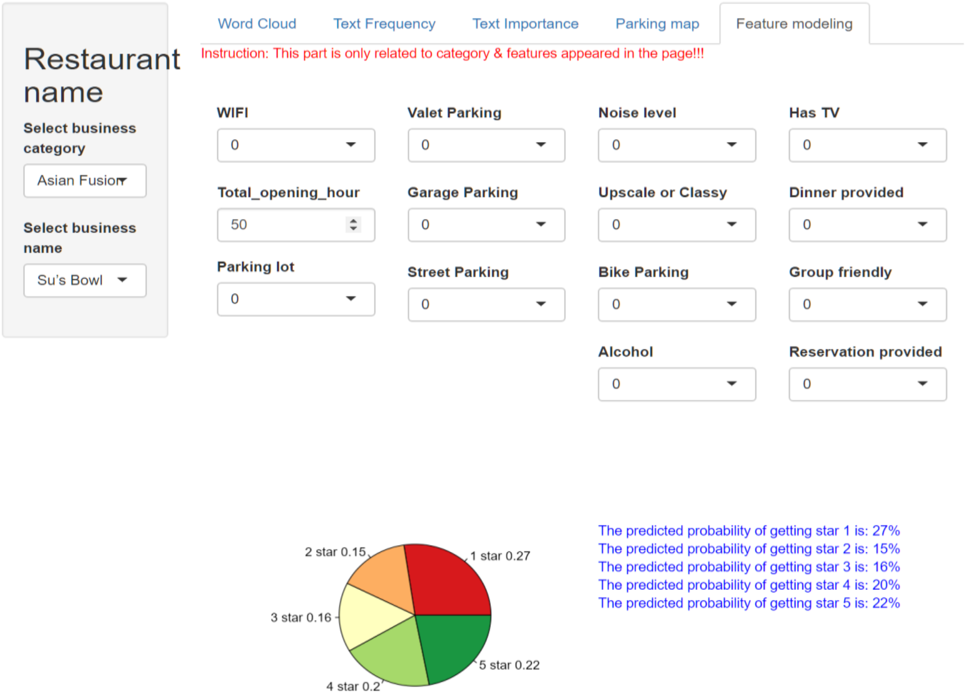
	text	TF-IDF
4	noodle great	0.186287934002318
7	potato steamed	0.186287934002318
8	shredded potato	0.186287934002318
10	food authentic	0.186287934002318

Showing 1 to 4 of 4 entries

Previous 1 Next

Shiny App: Parking Map & Modeling





Modeling data

Y *X*

comment star	Hour per week	HasTV	Alcohol	WiFi	Garage	Dinner	Accept Noise
Star given by customer	Weekly open hours of each restaurant	Is there a TV in the restaurant	Provide Alcohol or not	Is free WiFi available	Any place to park cars	Provide dinner Or not	Is the noise acceptable
Multi-Categorical Ordinal (1,2,3,4,5)	Continuous (16h-102h)	Binary Categorical data (1 indicates yes; 0 indicates no)					

Cumulative Logit Models For Ordinal Responses

Model function:

$$logit[P(Y \le j)] = \log[\frac{P(Y \le j)}{1 - P(Y \le j)}] = \log[\frac{\pi_1 + \dots + \pi_j}{\pi_{j+1} + \dots + \pi_J}] = \alpha_j + \beta x, \ j = 1, 2, \dots, J-1$$

$$P(Y \le j) = \exp(\alpha_j + \beta x)/[1 + \exp(\alpha_j + \beta x)], \ j = 1, 2, \dots, J-1$$

Result with 12 significant features:

$$\frac{P(Y \le j|X = x+1)/P(Y > j|X = x+1)}{P(Y \le j|X = x)/P(Y > j|X = x)} = \exp(\beta) = \exp(0.008) \approx 1$$

variable	coefficient &significance level	variable	coefficient &significance level	variable	coefficient &significance level	variable	coefficient &significance level
Intercept 1	-3.843***	BikeParking1	0.621***	lot1	-0.143**	dinner1	0.268**
Intercept 2	-3.204***	Reservations1	0.133**	valet1	-0.683***	Total_hour	0.008***
Intercept 3	-2.582***	Alcohol1	-0.416***	garage1	-0.877***	acceptable noise1	1.301***
Intercept 4	-1.632***	WiFi1	-0.267***	street1	0.269***	upscale classy1	-0.176**

Significant level: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Cumulative Logit Models under different categories

Difference between Japanese and Chinese restaurants:

Japanese

variable	coefficient & significance level	variable	coefficient & significance level
Intercept 1	-3.563***	BikeParking1	0.948***
Intercept 2	-2.888***	Good For Groups	0.455***
Intercept 3	-2.247***	garage1	-0.801***
Intercept 4	-1.269***	WiFi1	-0.839***
Total_hour	0.014***	street1	0.461***
upscale classy1	-0.360***		

Significant level: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Chinese

variable	coefficient & significance level	variable	coefficient & significance level
Intercept 1	-2.768***	BikeParking1	0.692***
Intercept 2	-2.153***	HasTV	-0.225*
Intercept 3	-1.592**	Alcohol1	-0.687***
Intercept 4	-0.627**	lot1	0.585***
Total_hour	0.021***	dinner1	-0.343*
upscale classy1	0.387***		

Significant level: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Goodness of fit test

H_0: The model fits the data well

H_1: The model doesn't fit the data well

Goodness of Fit test			
	χ^2	residual df	p-value
12features Model	$4.29 * 10^4$	42740	0.238
Chinese	6594.96	6629	0.614
Japanese	$1.802 * 10^4$	18041	0.527

Conclusion

Working longer will **not improve** ratings

Provide more **attentive service** to make customers give more positive reviews

Different categories of restaurants need to focus on different features to make improvements

Limitation

Improve model prediction accuracy

Further **explore features** that are highly correlated with rating stars

Provide **customized suggestion** for each business

Apply **vertical research** on how specific restaurant can be improved based on customers review

THANK YOU