How to run a successful brunch restaurant ?

Yelp Data Analysis

Yunwen Jiang, Yueting Tang, Yunbei Zheng

Objectives

- From two aspects: guests’ reviews and business attributes, provide some actionable suggestions for North American breakfast & brunch businesses

- Predict the ratings of reviews based on a regularized logistic model

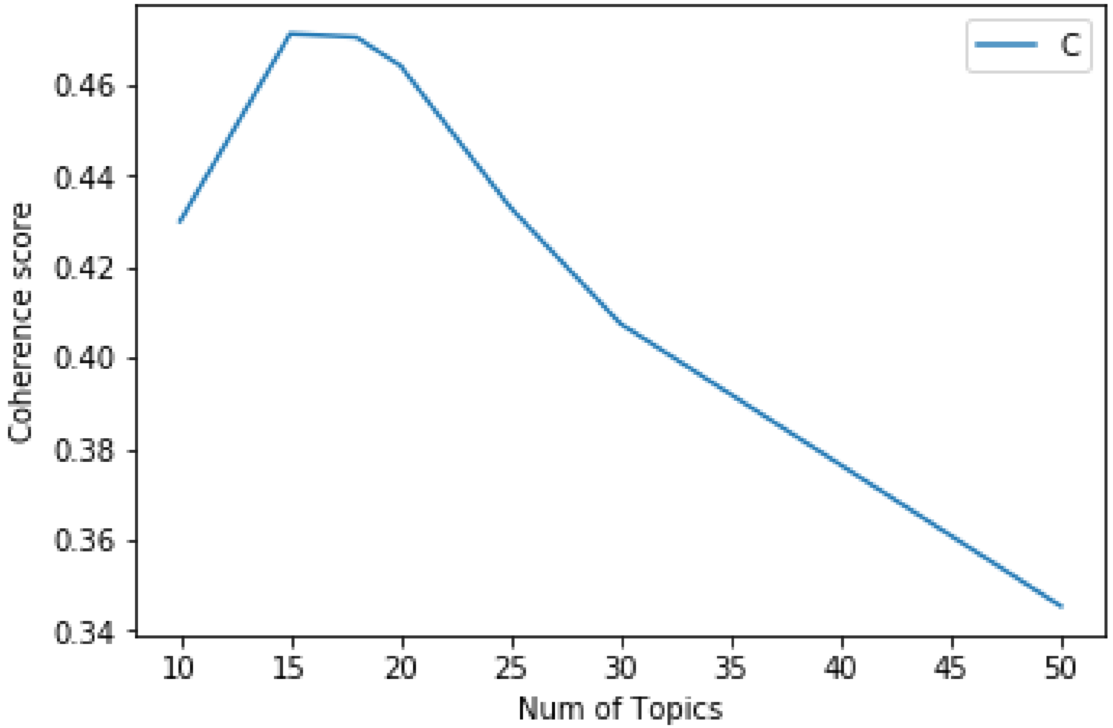
CONTENTS

- 01 Topic Model
- 02 Hypothesis Testing
- 03 Business Attributes
- 04 Suggestions
- 05 Prediction Model

Topic Model

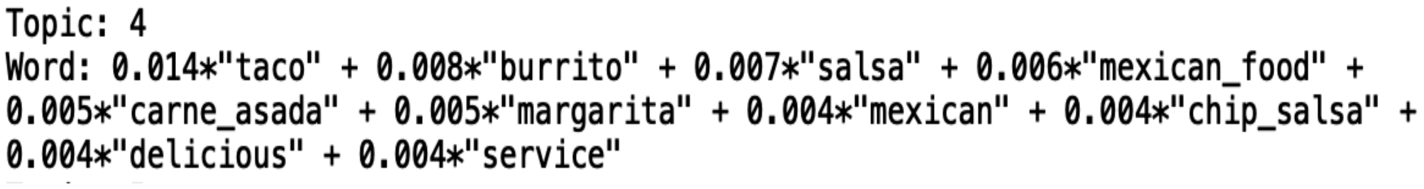
LDA Model

Latent Dirichlet Allocation(LDA) model is a "generative probabilistic model" of a collection of composites made up of parts.
Here, composites refer to reviews and parts refer to words or phrases.

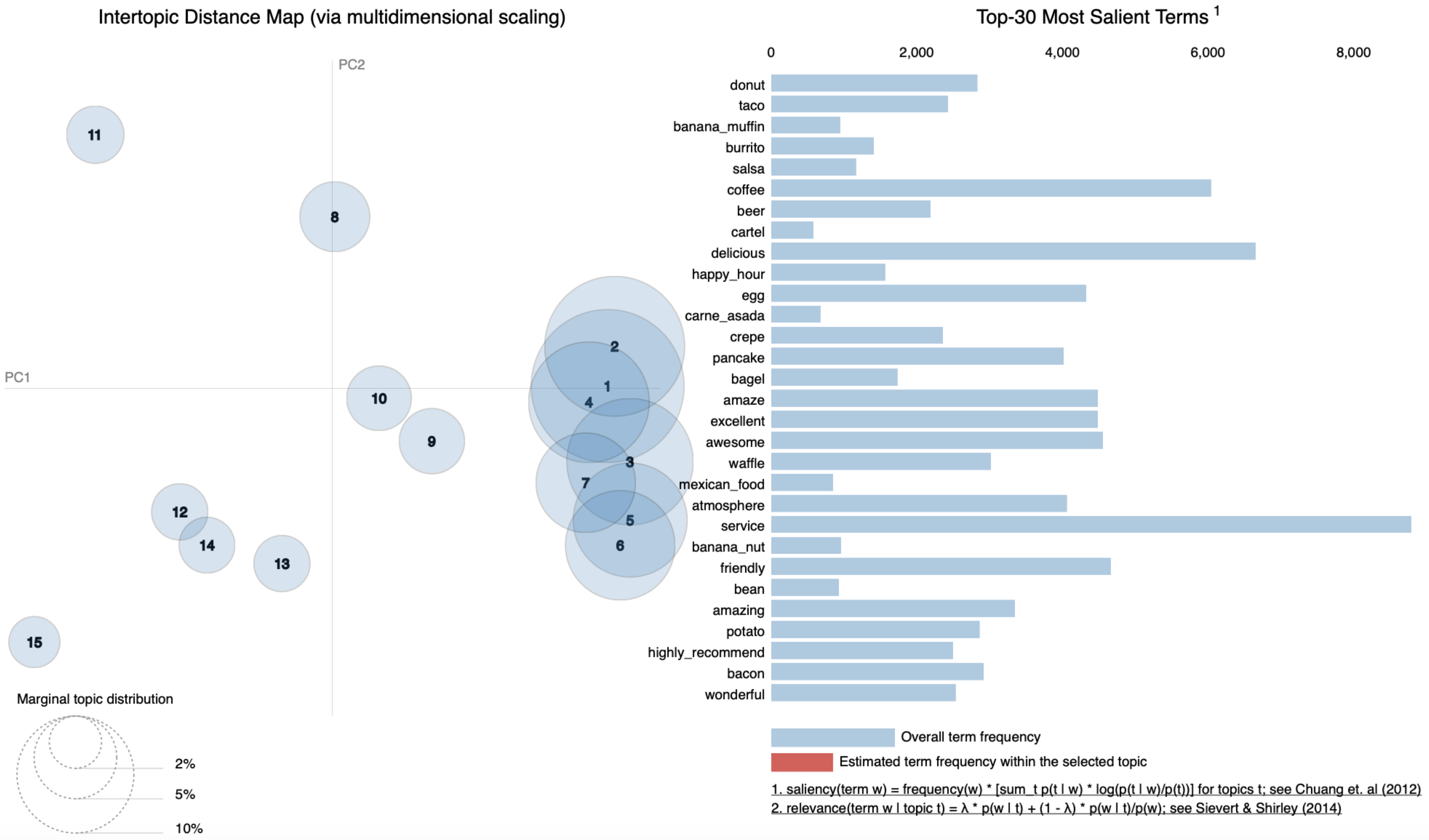


Coherence Score

We keep the LDA model with the highest coherence score, which has 15 topics.

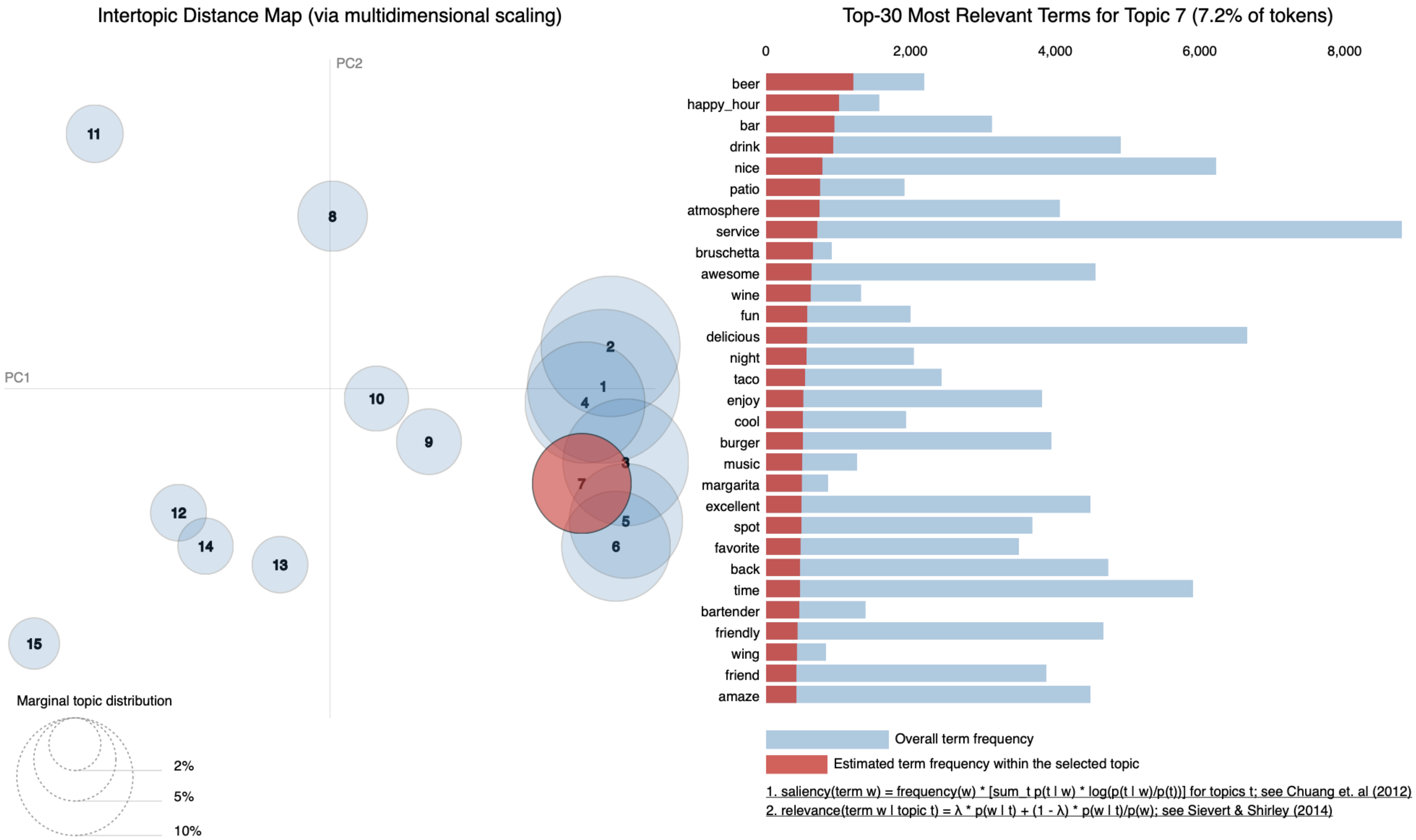


```
Topic: 4
Word: 0.014*"taco" + 0.008*"burrito" + 0.007*"salsa" + 0.006*"mexican_food" +
0.005*"carne_asada" + 0.005*"margarita" + 0.004*"mexican" + 0.004*"chip_salsa" +
0.004*"delicious" + 0.004*"service"
```



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

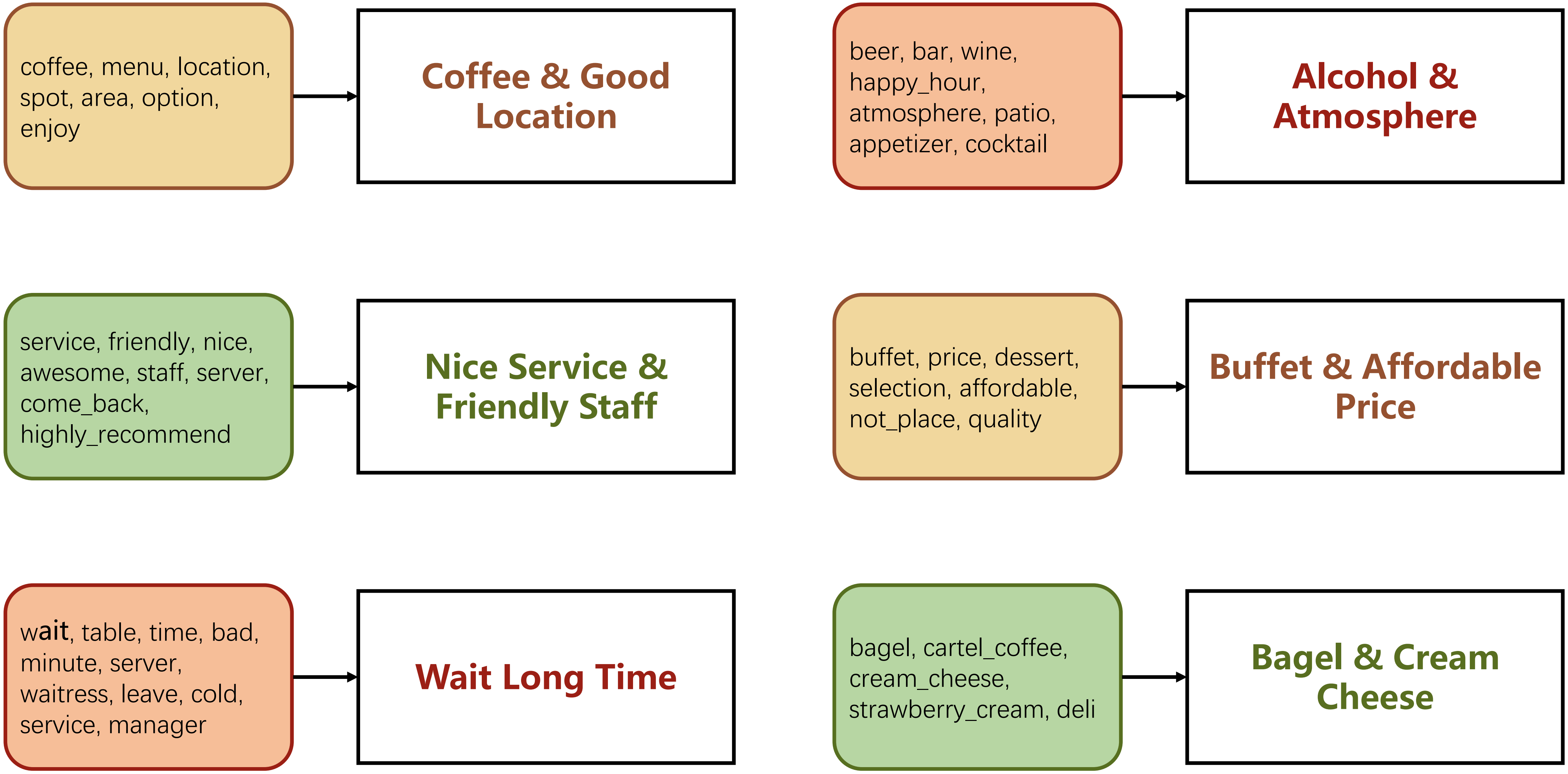


Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 7 (7.2% of tokens)

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)



coffee, menu, location, spot, area, option, enjoy

Coffee & Good Location

beer, bar, wine, happy_hour, atmosphere, patio, appetizer, cocktail

Alcohol & Atmosphere

service, friendly, nice, awesome, staff, server, come_back, highly_recommend

Nice Service & Friendly Staff

buffet, price, dessert, selection, affordable, not_place, quality

Buffet & Affordable Price

wait, table, time, bad, minute, server, waitress, leave, cold, service, manager

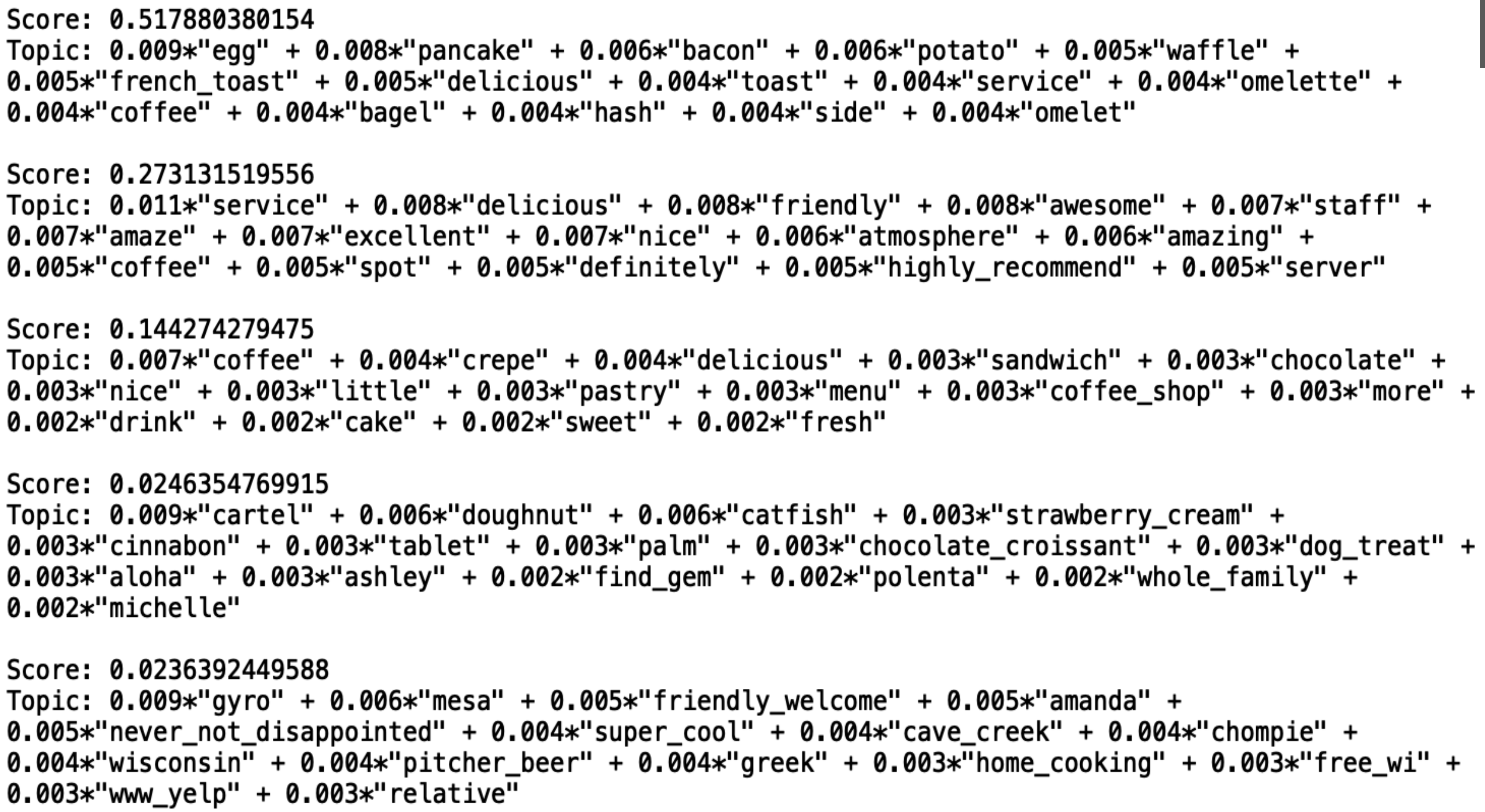
Wait Long Time

bagel, cartel_coffee, cream_cheese, strawberry_cream, deli

Bagel & Cream Cheese

399,991th Review

'long update_review az bread_company fantastic come move_arizona freindliest people_work come hi remember cheerful food general french_toast egg salad fondness green_chili quiche particular come fast unique actually extremely quick bring_out food appear busy wind bring niece_nephew child friendly_staff happy see take time see true_hidden gem glad house green_chile quiche strawberry french_toast egg salad sandwich'



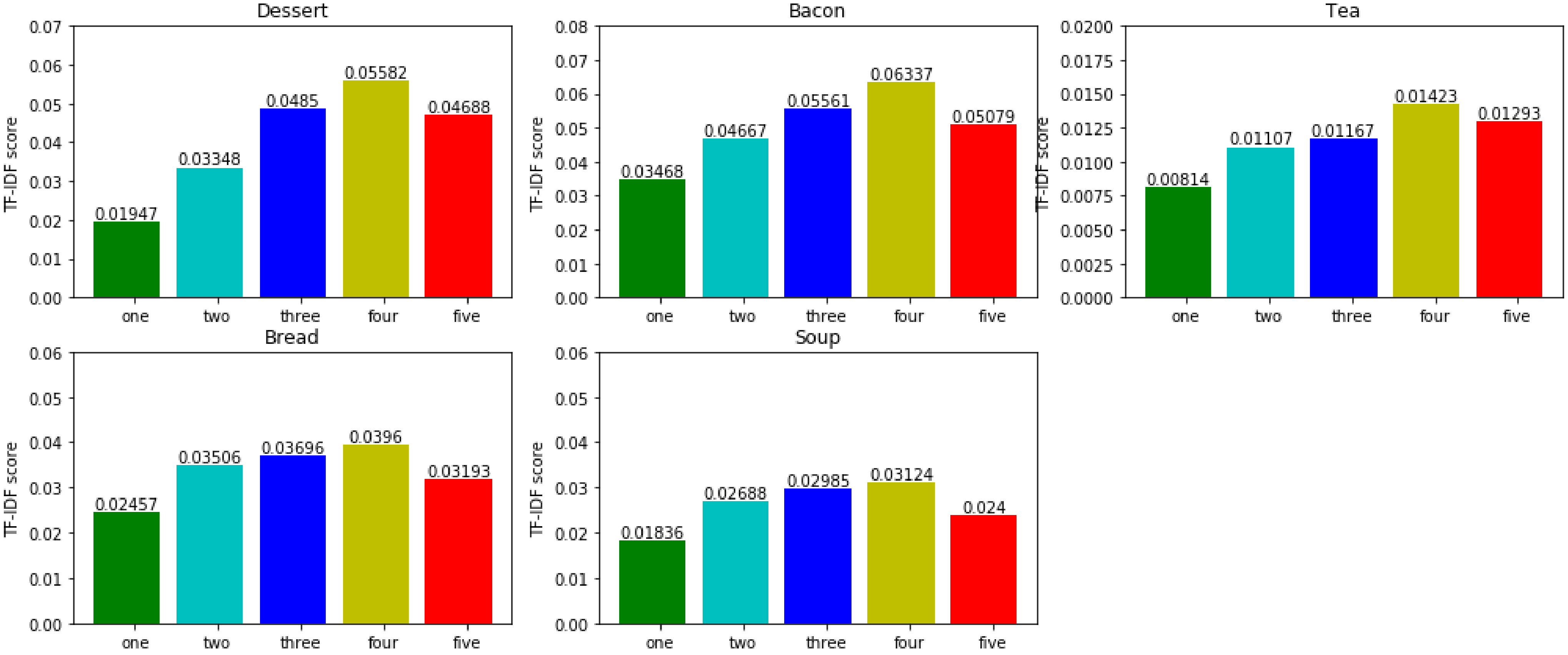
```
Score: 0.517880380154
Topic: 0.009*"egg" + 0.008*"pancake" + 0.006*"bacon" + 0.006*"potato" + 0.005*"waffle" +
0.005*"french_toast" + 0.005*"delicious" + 0.004*"toast" + 0.004*"service" + 0.004*"omelette" +
0.004*"coffee" + 0.004*"bagel" + 0.004*"hash" + 0.004*"side" + 0.004*"omelet"

Score: 0.273131519556
Topic: 0.011*"service" + 0.008*"delicious" + 0.008*"friendly" + 0.008*"awesome" + 0.007*"staff" +
0.007*"amaze" + 0.007*"excellent" + 0.007*"nice" + 0.006*"atmosphere" + 0.006*"amazing" +
0.005*"coffee" + 0.005*"spot" + 0.005*"definitely" + 0.005*"highly_recommend" + 0.005*"server"

Score: 0.144274279475
Topic: 0.007*"coffee" + 0.004*"crepe" + 0.004*"delicious" + 0.003*"sandwich" + 0.003*"chocolate" +
0.003*"nice" + 0.003*"little" + 0.003*"pastry" + 0.003*"menu" + 0.003*"coffee_shop" + 0.003*"more" +
0.002*"drink" + 0.002*"cake" + 0.002*"sweet" + 0.002*"fresh"

Score: 0.0246354769915
Topic: 0.009*"cartel" + 0.006*"doughnut" + 0.006*"catfish" + 0.003*"strawberry_cream" +
0.003*"cinnabon" + 0.003*"tablet" + 0.003*"palm" + 0.003*"chocolate_croissant" + 0.003*"dog_treat" +
0.003*"aloha" + 0.003*"ashley" + 0.002*"find_gem" + 0.002*"polenta" + 0.002*"whole_family" +
0.002*"michelle"

Score: 0.0236392449588
Topic: 0.009*"gyro" + 0.006*"mesa" + 0.005*"friendly_welcome" + 0.005*"amanda" +
0.005*"never_not_disappointed" + 0.004*"super_cool" + 0.004*"cave_creek" + 0.004*"chompie" +
0.004*"wisconsin" + 0.004*"pitcher_beer" + 0.004*"greek" + 0.003*"home_cooking" + 0.003*"free_wi" +
0.003*"www_yelp" + 0.003*"relative"
```

Hypothesis Testing

01 For TF-IDF scores, we applied Spearman correlation test.

02 For topic model by LDA, we applied chi-square test.

Distribution plot (last time)



- Similar trend for some food items.

Correlation test

H0: the value of the association measure = 0, which means the two samples are uncorrelated.

The table shows the TF - IDF score of the food items.

	star1	star2	star3	star4	star5
Bread	0.02457	0.03506	0.03696	0.0396	0.03193
soup	0.01836	0.02688	0.02985	0.03124	0.024

	star1	star2	star3	star4	star5
Bacon	0.03468	0.04667	0.05561	0.06337	0.05079
dessert	0.01947	0.03348	0.0485	0.05582	0.04688

	star1	star2	star3	star4	star5
Tea	0.00814	0.01107	0.01167	0.01423	0.01293
dessert	0.01947	0.03348	0.0485	0.05582	0.04688

Spearman's rank correlation rho:
S = 4.4409e-15, p-value =
0.01667

Spearman's rank correlation rho:
S = 4.4409e-15, p-value =
0.01667

Spearman's rank correlation rho:
S = 2, p-value = 0.08333

All of the three hypothses reject H0 at 90% significance level, so these three pairs of words are correlated with each other respectively. These food items do have a similar trend.

Chi-square Test

- **H0: The distribution of two data sets are independent. Which means the distribution of reviews with these topic words is different from the distribution of reviews without those.**
- The table shows the topic score of whether or not the reviews of the restaurants have these words among different stars.

	star1	star2	star3	star4	star5
Wait long time	26858	18794	16856	14360	12782
Not contain these	22897	25603	47762	117721	202052

Chi-squared test for given probabilities: X-squared = 160060, df = 4, p-value < 2.2e-16

	star1	star2	star3	star4	star5
Atmosphere&alcohol	708	821	1499	4243	8652
Not contain these	49047	43576	63119	127838	206182

X-squared = 1460.7, df = 4, p-value < 2.2e-16

	star1	star2	star3	star4	star5
Main courses	3417	5850	10788	20898	22782
Not contain these	46338	38547	53830	111183	192052

X-squared = 5137.5, df = 4, p-value < 2.2e-16

Summary of Chi-square test

- All of the 15 hypothses reject H0 at 95% significance level,

  so these top 15 topic words' group are all significant in the reviews among different stars. The distribution of reviews with these topic words is different from the distribution of reviews without those.

Business Attributes

Question: Which business attributes are important to star ratings and how they influence?

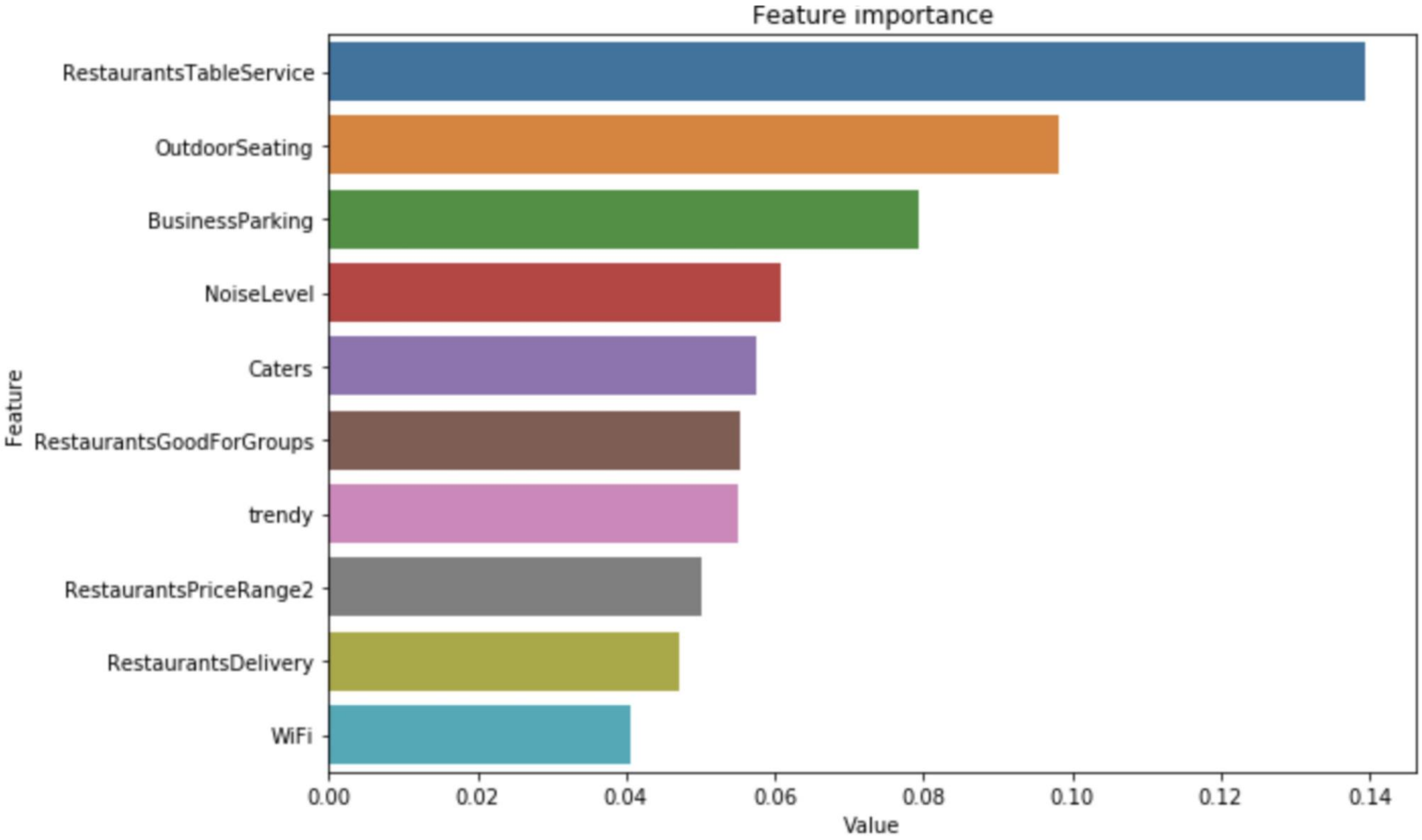
Step 1: Data preprocessing

01 Deleted the business attributes with 80 percent missing valus.

02 Transformed string type into categorical type.

We finally got retained **26** business attributes.

Step 2: GBDT and importance score



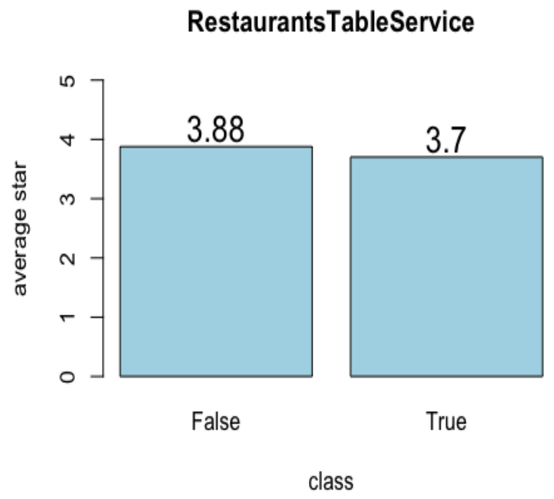
Step 3: Hypothesis testing

01 For business attributes that have only 2 levels, we applied Wilcoxon rank-sum test.

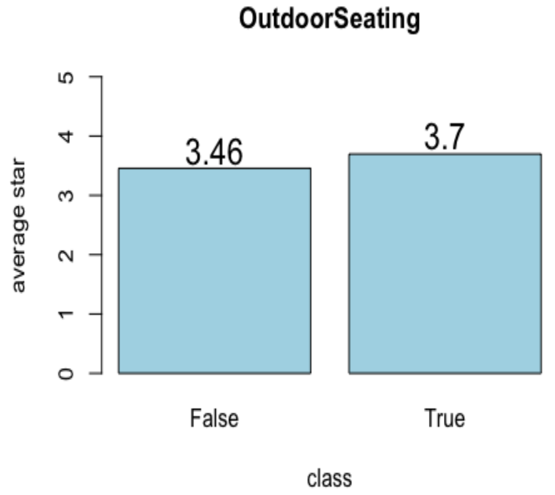
02 For business attributes that have more than 2 levels, we first applied Kruskal–Wallis H test to check if there is something difference among the levels. If p –value < 0.05, we would do pairwise Wilcoxon test with Bonferroni correction to see which pair is different.

Result: The top 5 business attributes are related to star ratings. For the NoiseLevel that has 4 levels, very_loud and loud samples originate from the same distribution.

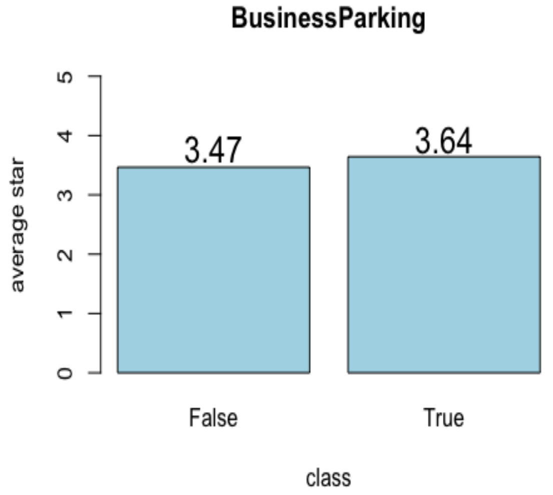
Suggestions



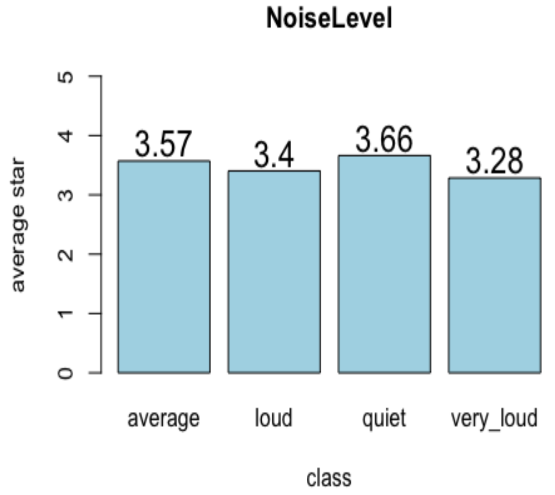
Cancel the table service will increase the average rating by 0.18 stars.



Add some outdoor seats will increase the average rating by 0.24 stars.

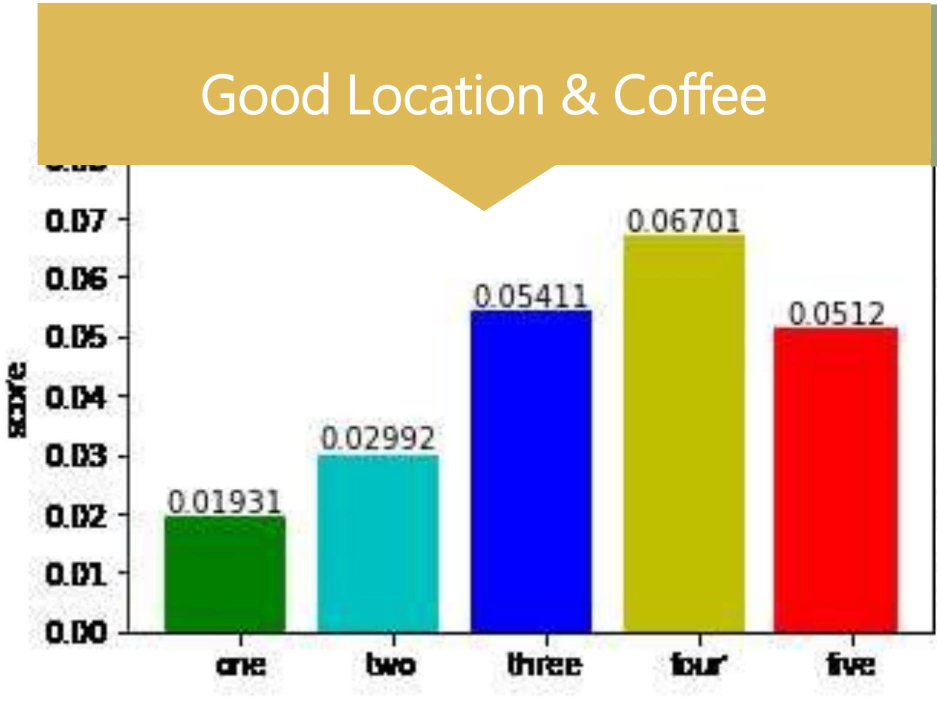


Provide some parking places for customers will increase the average star rating by 0.17 stars.

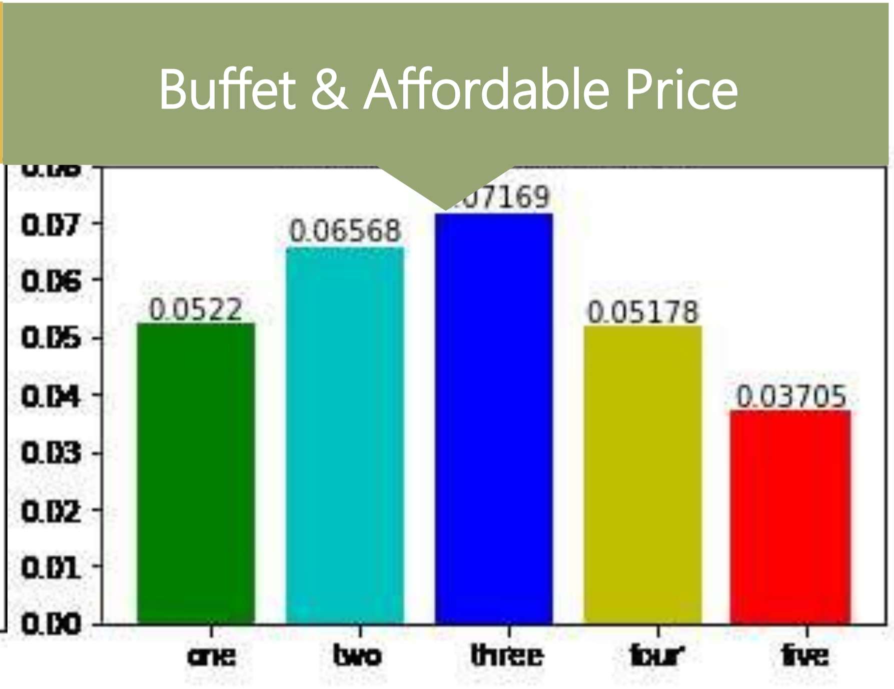


Make the environment quieter will increase the average star by 0.09 stars to 0.26 stars.

Good Location & Coffee

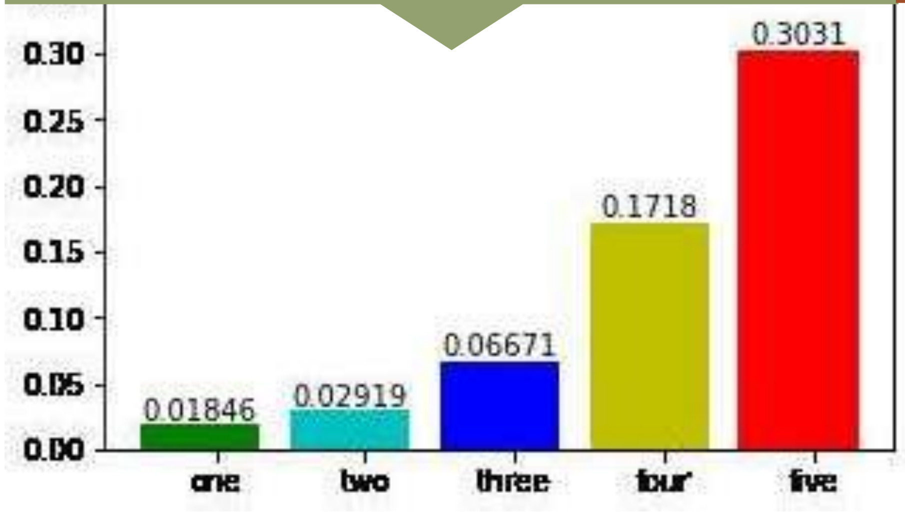


Buffet & Affordable Price

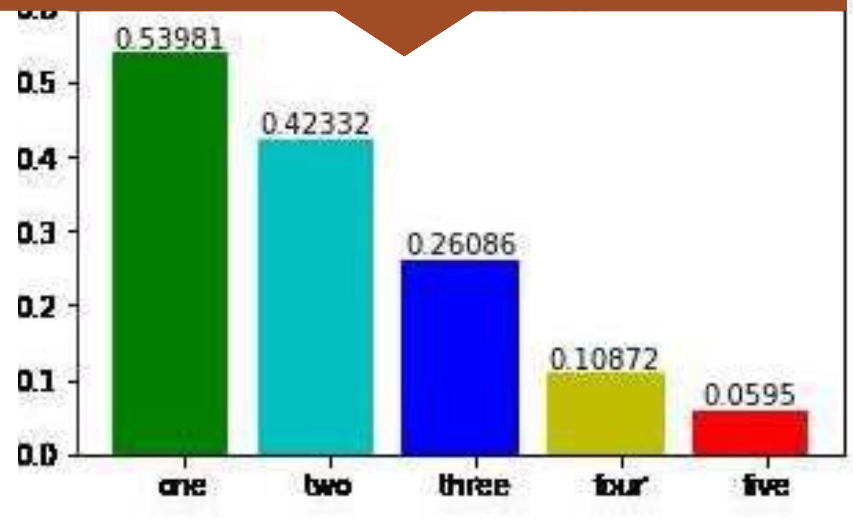


Business owners do not need to pursue low price and good location, which only help achieve average, but not extraordinary.

Nice Service & Friendly Staff

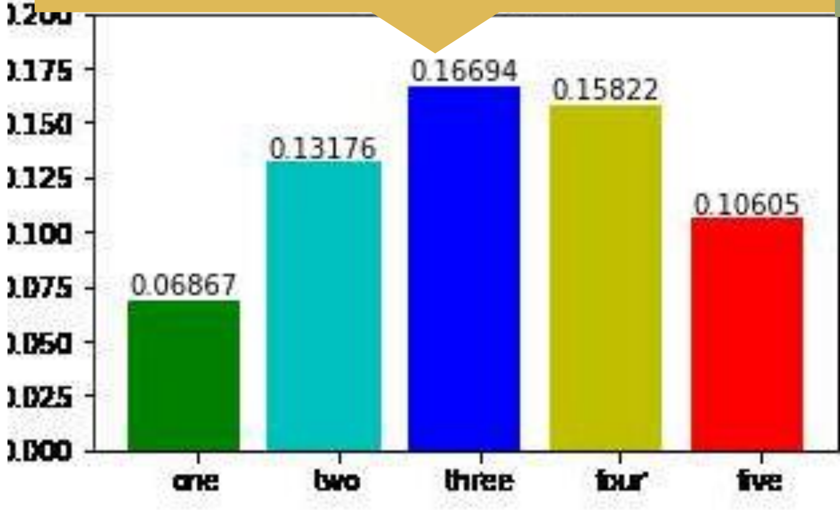


Wait Long Time

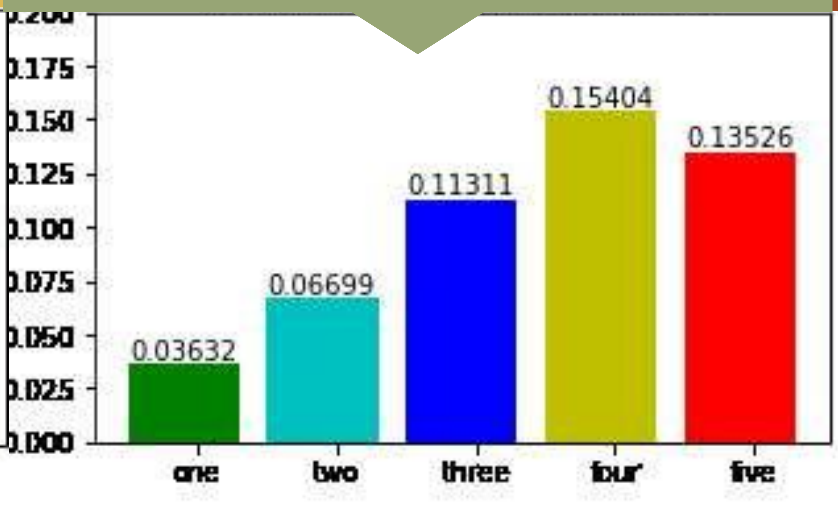


Service is soooooo important!
And try to make the business less crowded and reduce the wait time.

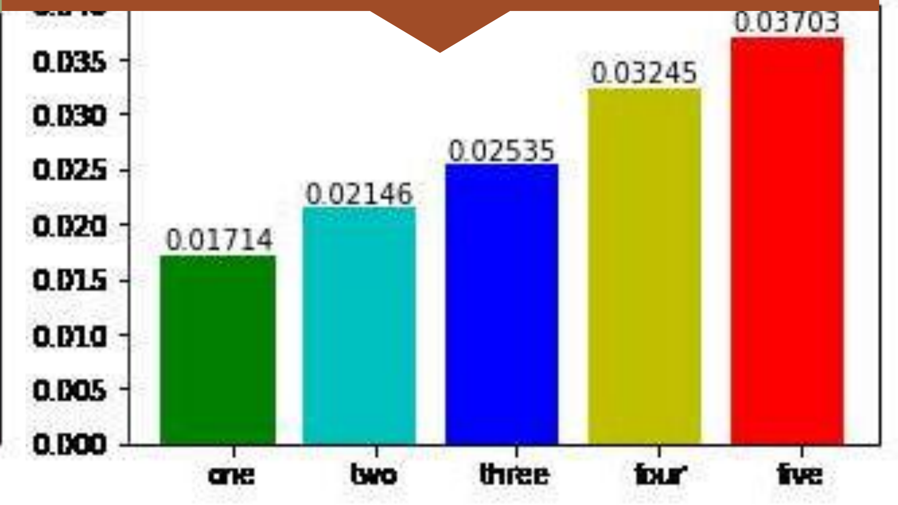
Main Courses



Pancake,Waffle,Toast,Egg,Omelet

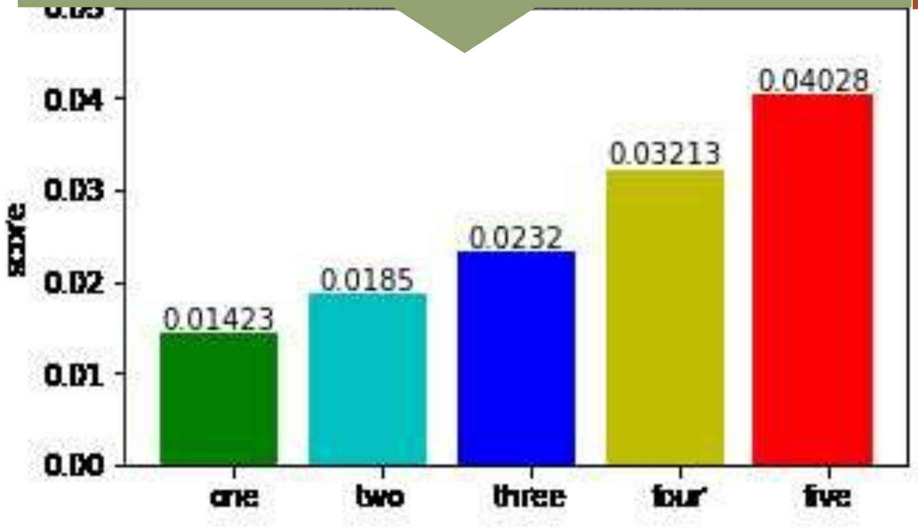


Mexican

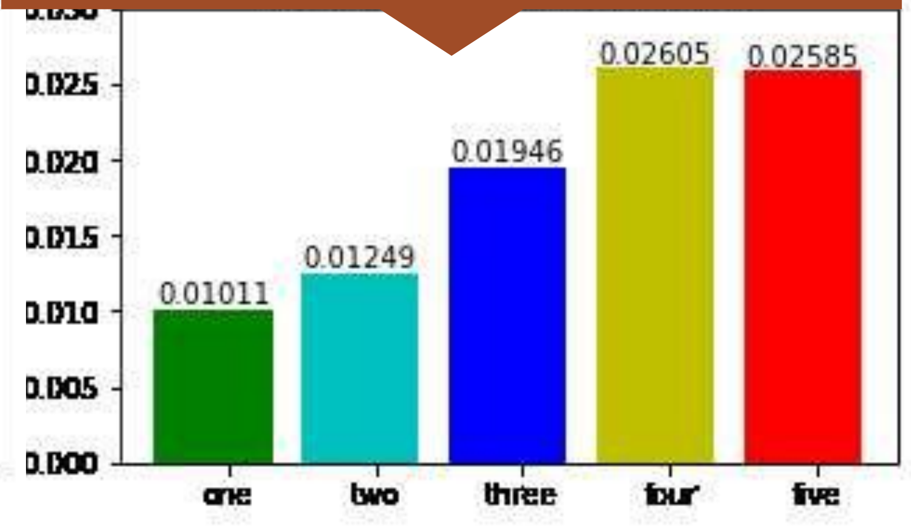


Regular food have no competitiveness for brunch restaurants,
food especially for breakfast are more attractive,
but Mexican style food are highly recommended.

Alcohol & Atmosphere

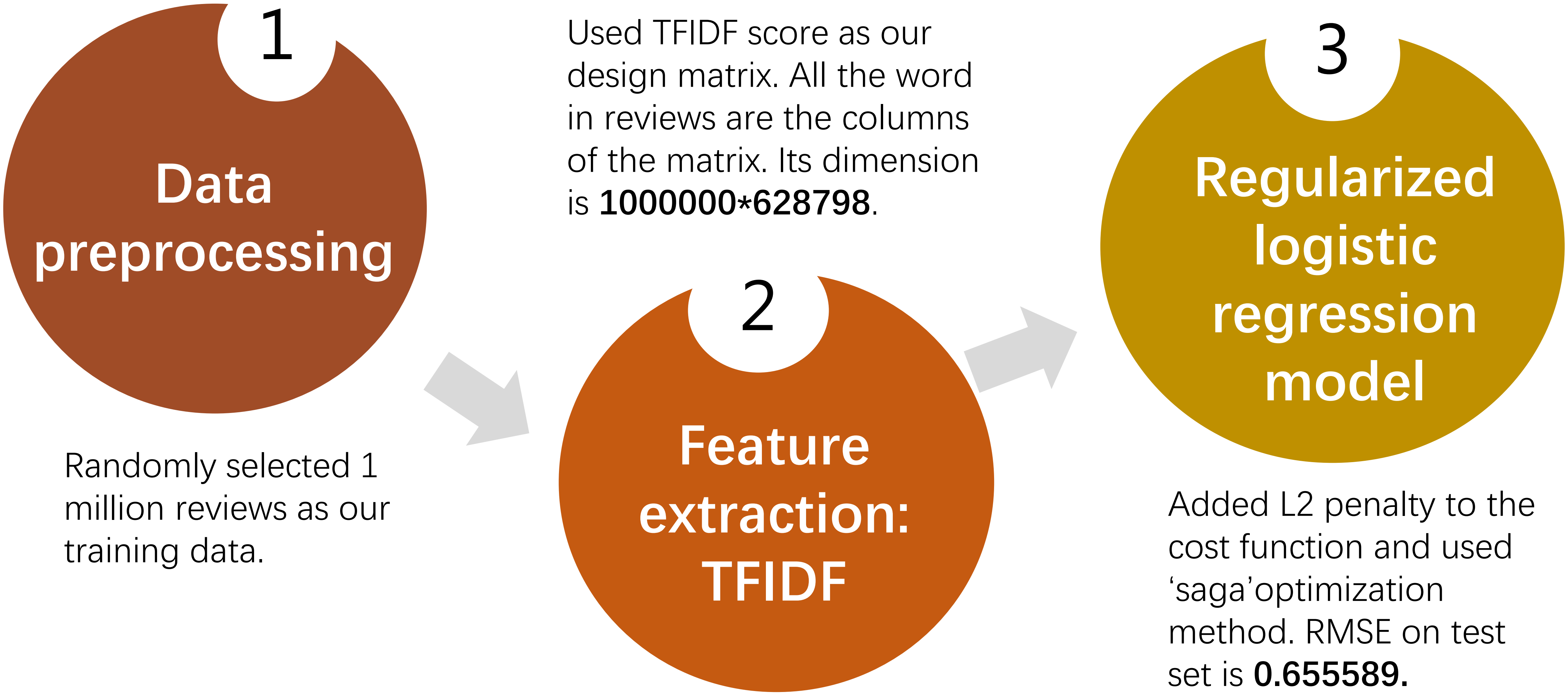


Bar, Bruschetta & Happy Hour



Various Alcohol and music contribute to happier atmosphere.

Prediction Model



1

Data preprocessing

Randomly selected 1 million reviews as our training data.

2

Feature extraction: TFIDF

Used TFIDF score as our design matrix. All the word in reviews are the columns of the matrix. Its dimension is **1000000*628798**.

3

Regularized logistic regression model

Added L2 penalty to the cost function and used ‘saga’optimization method. RMSE on test set is **0.655589.**

THANKS