Yelp Data Analysis

Yunwen J.,Yueting T.,Yunbei Z.

University of Wisconsin-Madison

STAT 628

March 7, 2019

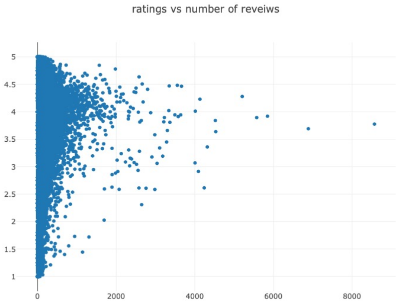
Objectives

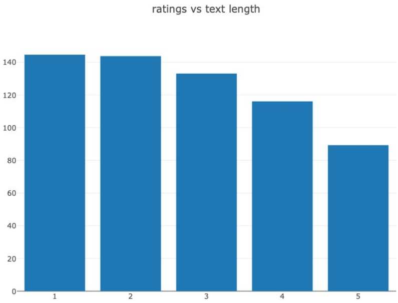
- Provide actionable suggestions to 4366 North American breakfast & brunch businesses
- Predict the ratings of reviews based on a prediction model

Content

1. Data Overview
   - Ratings vs Text
   - Reviews Distribution
2. Data Cleaning
   - Text cleaning
3. Idea based on business attributes
   - Business attributes process
   - Linear function
4. Feature Extraction
   - CountVectorizer
   - TF-IDF
5. Idea based on review texts
   - WordCloud
   - Distribution Plot(TF-IDF)
6. Next Step

Data Overview


ratings vs number of reveiws


ratings vs text length

Reviews Distribution

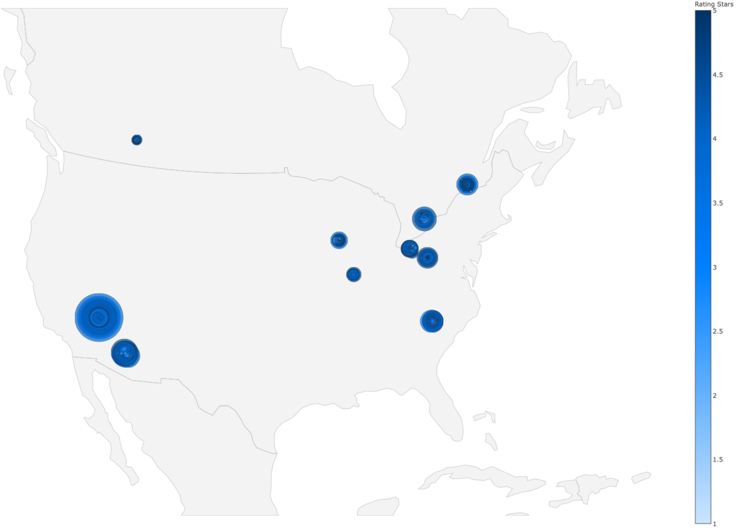


Figure: North American Reviews for All Businesses

Reviews Distribution

- We used keywords "Breakfast" and "Brunch" to select 4456 so-called brunch restaurants
- Then removed businesses with tags like "Asian", "Thai","Japanese" ... in categories. Finally, 4366 businesses with 505,696 reviews left

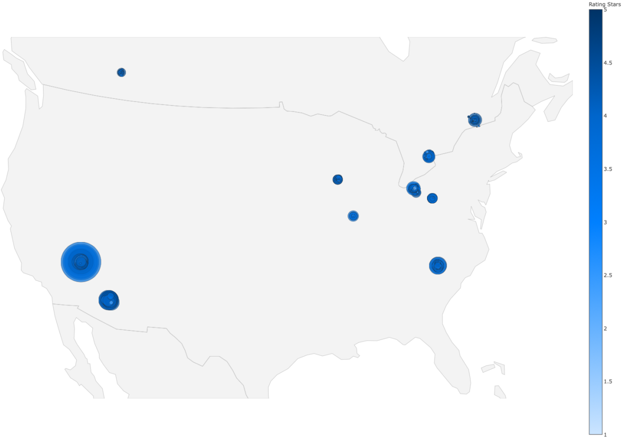
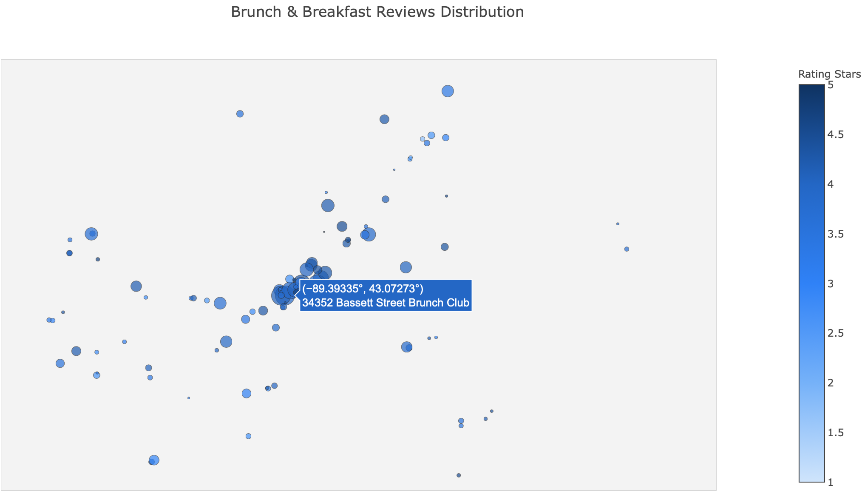


Figure: Brunch and Breakfast Reviews Distribution

As we zoom in to Madison...



Brunch & Breakfast Reviews Distribution
(−89.39335°, 43.07273°)
34352 Bassett Street Brunch Club
Rating Stars
5
4.5
4
3.5
3
2.5
2
1.5
1

Figure: Brunch and Breakfast Reviews Distribution

Text cleaning

- Transfer emoticons to English words
- Expand abbreviation
- Convert text to lowercase
- Handle negation
- Remove punctuation
- Lemmatize
- Remove unimportant words
- Extract phrases

Text cleaning

1.Transfer emoticons to English words

$$:) \rightarrow happy, \quad :D \rightarrow laugh$$

2.Expand abbreviation

$$n't \rightarrow not, \quad 'm \rightarrow am, \quad 's \rightarrow is$$
$$'ve \rightarrow have, \quad 'd \rightarrow would$$

3.Convert text to lowercase

$$\text{Lower} \rightarrow lower$$

Text cleaning

4.Handle negation

Add not_ to every word between negation and following punctuation

Before: The food did not taste good.

After: The food did not not_taste not_good.

5.Remove punctuation

remove these punctuation .,:;?!"′

6.Lemmatize

Reduce inflections or variant forms to base form

NN,NNS $\rightarrow$ $n(noun)$, $RB, RBR, RBS \rightarrow r(adverb)$

JJ,JJR,JJS $\rightarrow$ $a(adjective)$, $VB, VBG, VBD, VBN, VBP, VBZ \rightarrow v(verb$

Text cleaning

7.Remove unimportant words

Remove words with the following part of speech:
IN,MD,PRP,PRP$,TO,WDT,WP,WP$,WRB
Eg: the, to, some, i, he, in...

8.Extract phrases

highly recommend → *highly_recommend*
incredibly rude → *incredibly_rude*

Insights to business owners based on business attributes

16 Common attributes	
Alcohol	BikeParking
BusinessAcceptsCreditCards	BusinessParking
DogsAllowed	GoodForKids
HasTV	NoiseLevel
OutdoorSeating	RestaurantsDelivery
RestaurantsGoodForGroups	RestaurantsPriceRange2
RestaurantsReservations	RestaurantsTableService
RestaurantsTakeOut	WiFi

Insights to business owners based on business attributes

Business attributes process

transform 'string label' to 'integer label'

Eg: Alcohol: 'full_bar' to 2, 'beer_and_wine' to 1, 'none' to 0

BikeParking: 'True' to 1, 'none' to 0, 'False' to -1

NoiseLevel: 'quiet' to 2, 'average' to 1, 'none' to 0, 'loud' to -1, 'very_loud' to -2

Insights to business owners based on business attributes

Build a linear function

X: 16 business attributes as variables

y: average stars of restaurants

From the coefficients of 16 variables, we find that a restaurant with **parking place, outdoor seats, delivery and WiFi** will have higher star while a restaurant that **allows dog, good for kids and groups and noisy** will have lower star.

We will do more about this in next days and check if this is reasonable.

Feature Extraction

CountVectorizer:

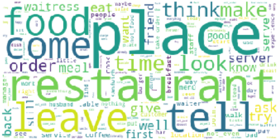
counts only the frequency of words in the text.

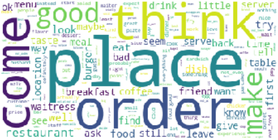
TfidfVectorizer:

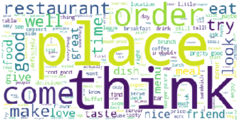
in addition to count the frequency of a word in a text, it also considers the number of all texts containing the word. It can reduce the impact of frequently appearing meaningless words and explore more meaningful features.

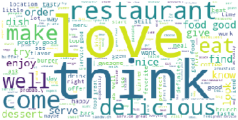
By contrast, the more text items there are, the more significant Tfidf will be.

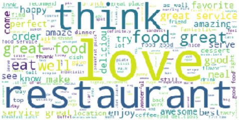
WordCloud





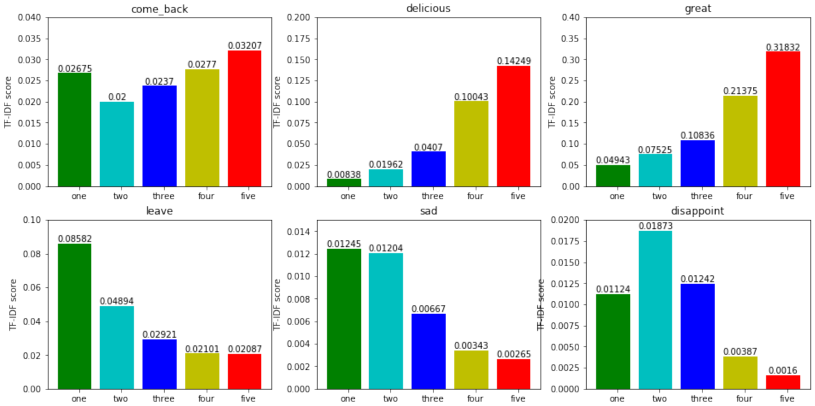






1-leave,tell 2&3-place,order 2&3&4&5-think 5-love

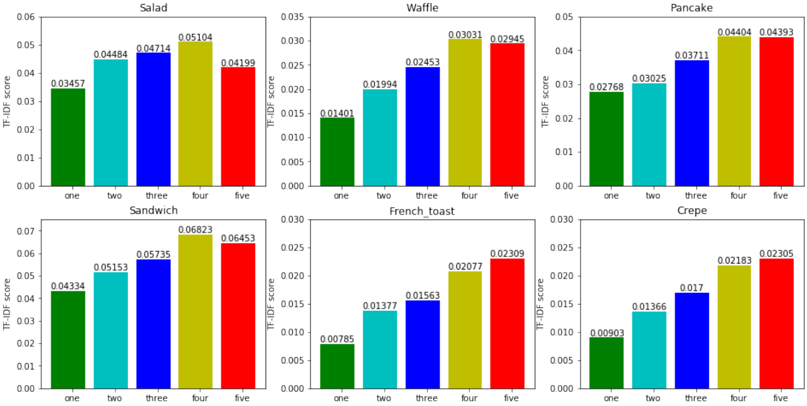
Distribution Plot



never come_back ?

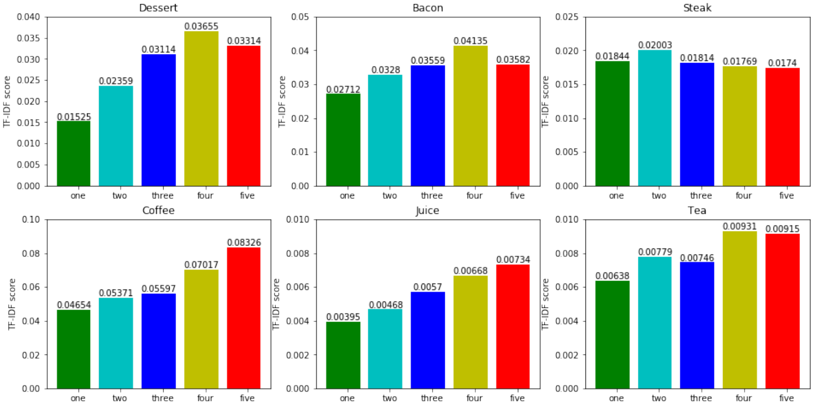
Distribution Plot

salad and sandwich



Distribution Plot

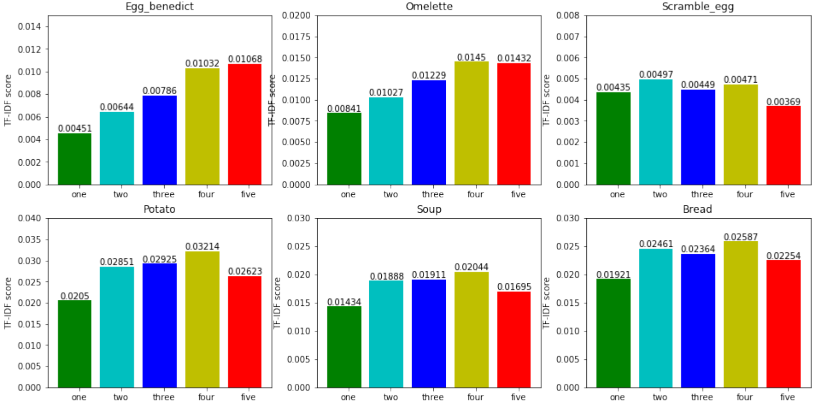
dessert and bacon ?



coffee and juice

Distribution Plot

different types of egg



Next Step

Correlation test

To test the correlation relationship between some food items.

Significance test

To test the significance of one variable's score through different stars' restaurants.

Improve TF-IDF

Score only shows the importance of each word(food or service items) in the text, can't reflect the positive or negative attitude of this item. We need to come up with more ideas of this aspect.

The End