Module 3-Group 18

Introduction/Motivation

The main goal of our work is to find the insights of the words and reviews of restaurants containing Chinese food. on YELP and to give some suggestions to the business owners to improve their rating based on the YELP data. And we mainly focus on 3 parts of the restaurants:(1)Taste:like salty,sweet and so on;(2)Food:chicken,cream and so on;(3)Drink:tea, water and so on. Reviews and ratings reflect the quality of product and service of the business, so making full use of reviews can truly improve their ratings. There are lots of reviews on different business on the YELP, it's not a easy work for us to get full useful, information from such number of reviews directly, so we decided to use some statistical methods to do a data based analysis to reach our goal.

Data Pre-Processing/Data Clean/Data View/Model Preparation

Data Pre-Processing/Data Clean

Use R to find out all the business whose categories containing the key word "Chinese" and still open, then based on these business IDs, we extract all the reviews on these business. Then we clean the review texts by the following steps using R:

1.Removing all the numbers.
2.Change all the characters to lower case.
3.Expand abbreviations.Such as "I'm" $\rightarrow$ "I am", "n't" $\rightarrow$ " not".
4.Removing all the punctuations.This step is reasonable because of step 3.
5.Change the words "no, not, never" and the space following them into the prefix "non".
6.Combine the review texts to their corresponding business IDs and ratings.
7.Use the package "tidytext" in R to break all the review texts into pieces of words.
8.Remove all the stopwords based on packages "tidytext".
9.Remove all the words which appear less than 10 times.
10.Construct the review-to-word n*p matrix and n-vectors of star ratings.

Overall View of the Data

Then, we do the following 2 things to get a overall views of the data:

1.Plot the rating VS frequency of some words, mainly adjectives and nouns, or some verbs containing strong sentiment, such as "like, hate" to see if these words affect the star ratings.

2.Calculate the overall rating VS frequency and plot, calculate the Pearson's χ^2 test statistics of all the words and give a sort to them to see which words have strongest influence in the star ratings.

Model Preparation

Consider the results of the steps above, we finally selected 3 groups of words on different 3 parts in our goal, they are:

1.Taste:"sweet","spicy","bitter","salty","bland","sour","crispy","greasy"
2.Food:"bread","burger","cheese","chicken","cream","pizza","sushi","steak" "
3.Drink:"beer","coffee","soup","tea","milk","wine","water"

Then we can use these 3 groups of words as predictors and star ratings as response to fit 3 multiple linear regression models and do further statistical analysis. However, these 3 models can only be used to predict with one certain type of words, like "sweet is good, fish is preferred, water is not welcomed", if we want to see "Whether sweet fish is more welcomed than spicy fish or not?", or "if sweet spicy food is welcomed?", we should fit a global final linear model with interactions to solve the problem above, so we still need to simplify our model, using the correlation coefficients, we finally choose 12 words as our predictors in the global linear model. They are: "sweet","spicy","bitter","salty","bland","sour","crispy","chicken","fish","beef","soup","tea"

EDA

The figure 1 shows the frequency by ratings of reviews containing the key words selected by the steps above:

From this figure we can easily find that most of the key words have actual influence on the rating.

For example, the reviews containing "bland" have more 1 star of 2 stars ratings than other words, which means the bland food may lead a lower rating.

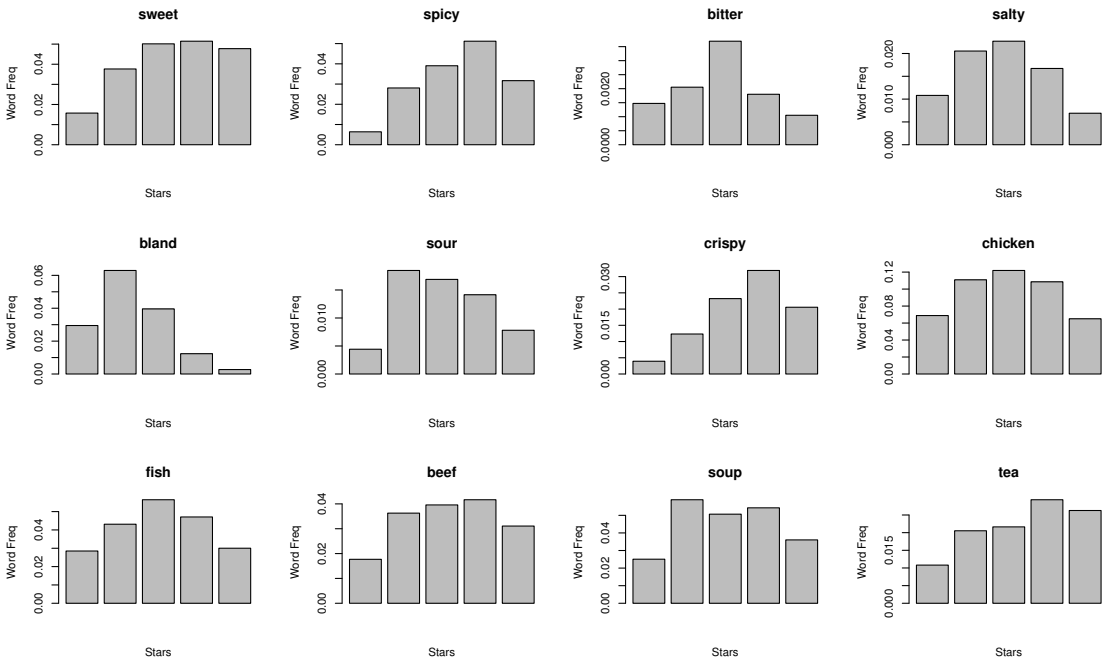

Figure 1: The Frequency by Rating of Views of Some Key Words

And the plot of "beef" is more right skewed than that of "fish", which means beef may be more welcomed than fish in Chinese food restaurants.

The plots of "salty","bitter","fish","soup" seem like highly centralized, so these words may have less influence on rating than other words .

Key Findings about Chinese Food Restaurants on YELP Data

We use 12 words to fit a global final linear model with interactions to see if these words actually have influence on the ratings.After fitting the model, there are $12 + 12 * 12 = 156$ variables in our model, it's not time-wasting to analyse every variable, so we use F-test and t-test statistic to retain several words and their interaction variables.

Of course, the Null Hypothesis of these tests are "the certain variable has no influence on ratings".Here our significant level is not common 0.05 or 0.1, we set it 0.25 for F-test and 0.5 for t-test. So many variables can be shown. This is because in our work here, what we want to know or see is that if some variables have influence on rating, in other words, we don't mind if the prediction is precise. The more variables, the more suggestions we can give, this is also why we don't fit another model using the significant variables.

The Table 1 shows the increasing of mean of rating with exist of some words and words combo, SE, and p-values of the F-test and t-test.

There are some really obvious conclusions in the table:

1.Sweet, spicy or crispy food are welcomed in Chinese food restaurant.

2.Bitter, salty or bland things may lead a bad rating.

3.Using fish seems to have no effect on ratings.

4.Bitter things with much salt may lead better rating.

....

Based on the data, model and table, the next part we will give some recommendations for business.

Recommendations for Businesses/Data-Driven Business Plan

Recommendations

The figure 2 shows contrast of ratings between whether the review includes the certain word or not, what we should know is that the average star ratings for Chinese food restaurant is 3.733, based on this figure and table 1, here are some recommendations for Chinese food restaurants:

words (combo)	increase ratings	SE	P-value of F	P-value of t
sweet	0.3220	0.0680	1.528e-08	2.254e-06
spicy	0.3701	0.0894	9.143e-06	3.496e-05
bitter	-0.7498	0.3876	1.926e-01	5.309e-02
salty	-0.5145	0.1463	2.067e-05	4.387e-04
bland	-1.1846	0.1125	5.769e-49	8.394e-26
sour	-0.1407	0.1619	1.459e-01	3.849e-01
crispy	0.3200	0.1068	3.608e-06	2.748e-03
chicken	-0.1780	0.0497	8.122e-04	3.453e-04
fish	-0.0289	0.0695	3.763e-01	6.771e-01
beef	0.1670	0.0838	2.652e-02	4.632e-02
soup	0.0632	0.0752	6.191e-01	4.002e-01
tea	0.3107	0.0927	1.525e-04	8.048e-04
sweet:spicy	-0.1622	0.1844	3.798e-01	3.789e-01
spicy:fish	-0.1803	0.2274	2.719e-01	4.278e-01
spicy:beef	-0.1884	0.1991	2.269e-01	3.440e-01
bitter:salty	1.4891	2.0486	7.288e-01	4.673e-01
bitter:tea	1.4540	1.6226	3.587e-01	3.701e-01

Table 1: Increasing Rating and Test P-values

Taste

Sweet and Spicy Food are Welcomed: On average, a Chinese food restaurant good at cooking sweet or spicy food will have about 0.3 star higher than a Chinese food restaurant which provide other taste of food(two-sample t-test p-val: 2.08e-11).

Bland Food Lead Lower Rating: Although salty or bitter food may decreases the rating star a little(0.3 on average), the ratings of bland food is really low which is about 1.2 star lower than restaurant whose food is not bland(two-sample t-test p-val: 3.056e-47).

Material Choose

Beef May Be Better than Chicken or Fish:Using beef as food material increase about 0.16 stars than average(MLR,p-val:4.632e-02), but using chicken or fish may conduct a decreasing or just a average star rating.

Drink Choose

Providing Tea Will Get Higher Rating:Soup cannot improve the review(two-sample t-test p-val: 9.548e-01), however, a cup of tea may be preferred than other drink, it will increase about 0.22 star rating(two-sample t-test p-val: 2.299e-06).

Words Combos:

Spicy Fish or Beef Will Reduce the Rate:The spicy beef or fish in not in recommendation list although beef and spicy food are liked by customers according to the discussion above, these combos reduce 0.18 star rate on average.

Bitter Tea is Really Good Choice:It seems no one will refuse the bitter tea, this wonderful combo has 1.45 star higher rating than average rating!

Limitation

Our findings is based on multiple linear regression model, however, the sample size is more than 19000 but some words appear only several hundred times, this may lead a sparse design matrix, in other words, there are too many zeros, sometimes the result may not be stable.

Also, our finding is not based on time series methods, so the findings may not have timeliness, it is possible that the popularity of taste varies from year to year.

Conclusions/Discussions

Our main process is: break all the cleaned review texts up to words and construct an n*p WordVSReview matrix; then uses Peason's Chi-square test statistics and words frequency to select words vectors as predictors to fit a multiple

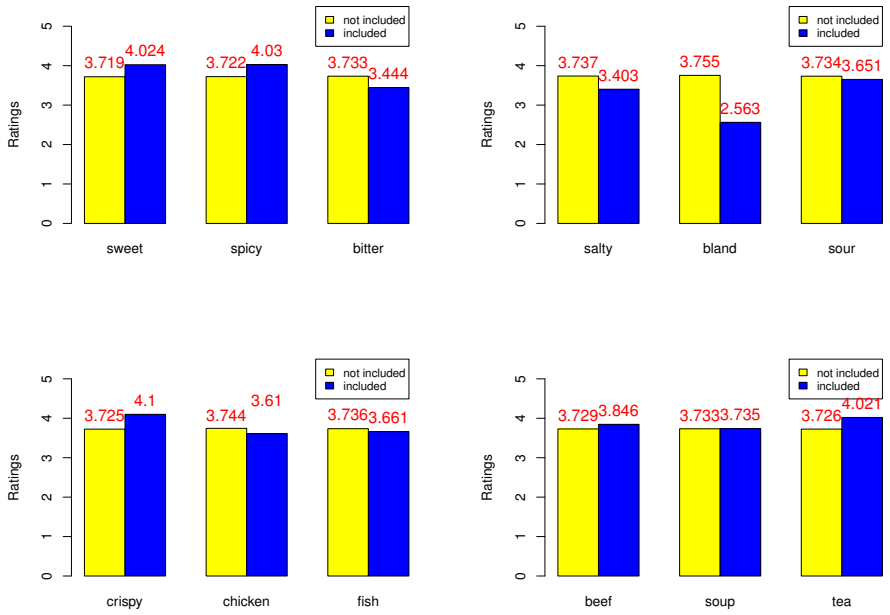

Figure 2: The Comparation Between If th Words Included

regression model; do two sample t-test. By these 2 ways, we find the insight of the review texts based on review and business json file, especially, we focus on the ratings between reviews containing one certain word and reviews not containing that word or reviews containing one certain word and average rating.We also notice that some word combos will lead a very good rating based on analysis of interactions in our MLR. Based on the statistical analysis, we provide some recommendations with plots and tables for Chinese food restaurant business owners.

Contributions

Hongyi Liu:Introduction, data clean, EDA, conclusion part of summary and presentation.Data clean and model selection on code on Github and shiny app.

Tianrun wang:Recommendations of summary and presentation.Data presentation and model diagnostic part on code on Github and shiny app.

Chenyang Jiang:Model build, key findings part on summary and presentation, model build and recommendation part on code on Github and shiny app.