Mysteries of the Yelp Orient

A Proposal for Dish-Based Restaurant Recommendations

Eugene Woo
email: eugene.hw.woo@gmail.com
linkedin: sg.linkedin.com/in/eugenewoo

Context

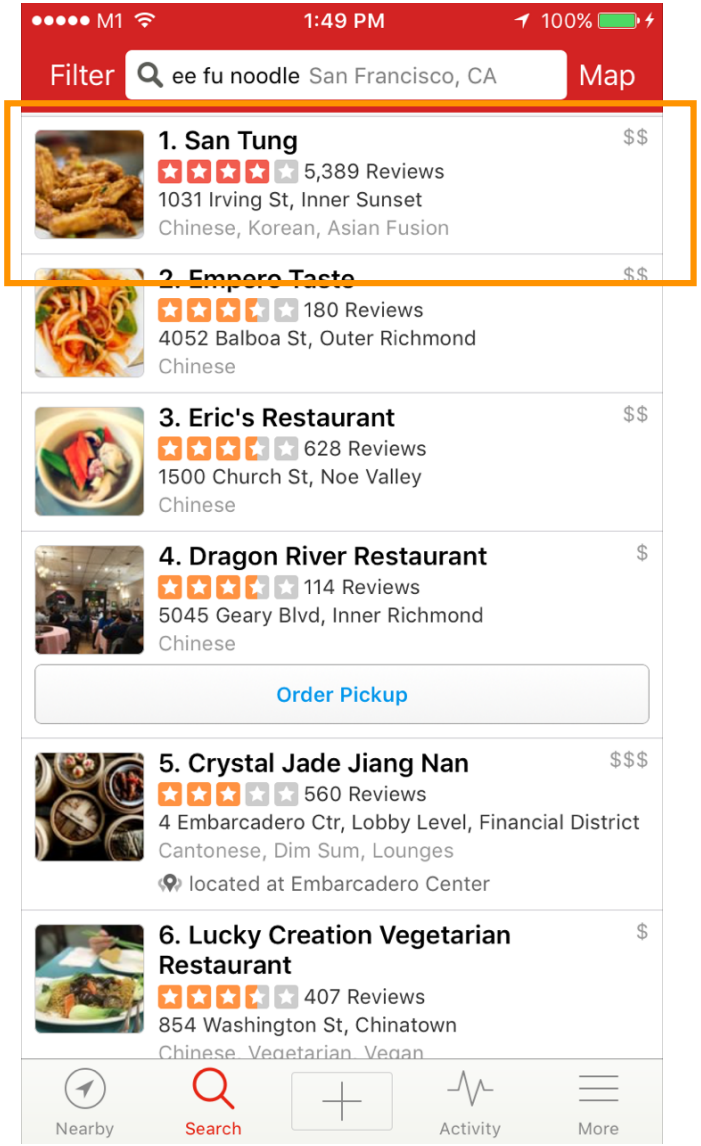
- I got interested in text analysis during Twitter AND also had to complete a capstone project to graduate from the Springboard Data Science workshop. Why not kill 2 birds with 1 stone?

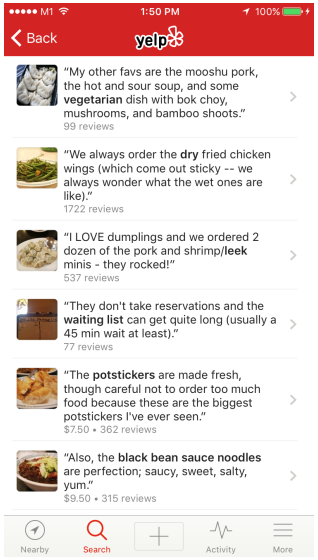
- Yelp does a good job of recommending restaurants from the top-down - proximity, star rating and review count

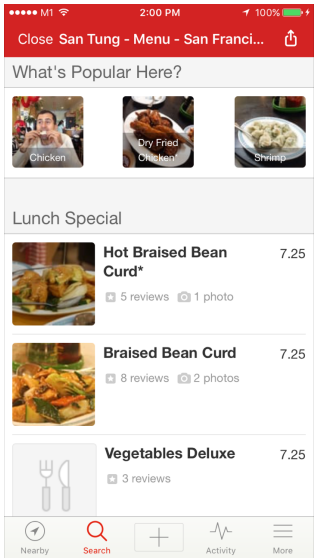
- During a recent trip to San Francisco, however, I had a hard time locating a Chinese restaurant serving a dish ‘Ee-Fu Noodles’. The result of my query was a broad match with ‘noodles’ but not what I wanted!

- For first-time visitors to a city, the problem is deciding which Chinese restaurant is the real deal when faced with pages of listings…

Context







Among the reviews, the closest match is the mention of “Black Bean Sauce Noodles”

No mention of Ee-Fu Noodles in the menu either

Methodology

- **Data Exploration**
  - Ensure we have sufficient quantity of Chinese restaurant reviews in Nevada and Arizona by active users
- **Text Mining & Tokenisation**
  - Count the single-word, two-word and three word-phrases in corpus of reviews to identify the most reviewed dishes.
- **Create Search App with Shiny**
  - Conduct exact match between the review text and the phrase list. This way we avoid boiling the ocean to identify every dish in the universe.
  - Import the match results into a Shiny app that renders the Top 5 Restaurants serving the selected dish in the selected state.

Data Source

- **Yelp Dataset Challenge (https://www.yelp.com/dataset_challenge)**
  - 4.1 million reviews
  - 1 million unique users
  - 144,000 businesses
  - US states: Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison, Cleveland

- The JSON data was read into R with jsonlite::stream_in and converted to .Rds format for efficient loading.

- I focused on the states of Nevada and Arizona given they account for 83% of total reviews and a sizeable quantity of Chinese restaurant reviews.

Exploration & Cleaning

- **Reviews dataset (Fig 1):** Votes, user ids, star ratings and review texts
- **Business dataset (Fig 2)**: Business info like shop names, addresses, opening hours, venue attributes such as ‘Good for Kids’, ‘Wi-Fi’, ’Takes Reservations’.
- I merged the datasets on the ‘Business_ID’ column

Figure 1

```
## 'data.frame':    2225213 obs. of  8 variables:
##  $ votes      :'data.frame': 2225213 obs. of  3 variables:
##  $ user_id    : chr  "PUFPaY9KxDAcGqfsorJp3Q" "Iu6AxdBYGR4A0wspR9BYHA"
3uYrqIBXg" ...
##  $ review_id  : chr  "Ya85v4eqdd6k9Od8HbQjyA" "KPvLNJ21_4wbYNctrOwWdQ"
4kSNW5pgA" ...
##  $ stars      : int  4 5 5 5 3 1 4 5 5 3 ...
##  $ date       : chr  "2012-08-01" "2014-02-13" "2015-10-31" "2013-11-08
##  $ text       : chr  "Mr Hoagie is an institution. Walking in, it does
 old fashioned menu board, booths out of"| __truncated__ "Excellent food.
io machines they used to have, but it's still a great place steeped in t"|
tle out dated and not opened on the weekend. But other than that the staff
cated__ "All the food is great here. But the best thing they have is their
c!!  The \"Wet Cajun\" "| __truncated__ ...
##  $ type       : chr  "review" "review" "review" "review" ...
##  $ business_id: chr  "5UmKMjUEUNdYWqANhGckJw" "5UmKMjUEUNdYWqANhGckJw"
AVUBZMjQQ" ...
```

Figure 2

```
## 'data.frame':    77445 obs. of  15 variables:
##  $ business_id  : chr  "5UmKMjUEUNdYWqANhGckJw" "UsFtqoBl7naz8AVUB
8Qxe4ol6y_g" ...
##  $ full_address : chr  "4734 Lebanon Church Rd\nDravosburg, PA 150
 "1 Ravine St\nDravosburg, PA 15034" "1530 Hamilton Rd\nBethel Park, 
##  $ hours        :'data.frame':   77445 obs. of  7 variables:
##  $ open         : logi  TRUE TRUE TRUE FALSE TRUE TRUE ...
##  $ categories   :List of 77445
##   .. [list output truncated]
##  $ city         : chr  "Dravosburg" "Dravosburg" "Dravosburg" "Bet
##  $ review_count : int  4 4 3 5 5 20 3 21 7 4 ...
##  $ name         : chr  "Mr Hoagie" "Clancy's Pub" "Joe Cislo's Aut
##  $ neighborhoods:List of 77445
##   .. [list output truncated]
##  $ longitude    : num  -79.9 -79.9 -79.9 -80 -80.1 ...
##  $ state        : chr  "PA" "PA" "PA" "PA" ...
##  $ stars        : num  4.5 3.5 5 2.5 2.5 5 2.5 4 2.5 4 ...
##  $ latitude     : num  40.4 40.4 40.4 40.4 40.4 ...
##  $ attributes   :'data.frame':   77445 obs. of  36 variables:
##  $ type         : chr  "business" "business" "business" "business"
```

Exploration & Cleaning

- **Scrub Inactive Users (Fig 3):** Taking out users who wrote only 1 review removes 13% of observations
- **Sufficient Quantity of Chinese Reviews (Fig 4)**: 15% of reviews for the Top 5 restaurant categories were for Chinese

Figure 3

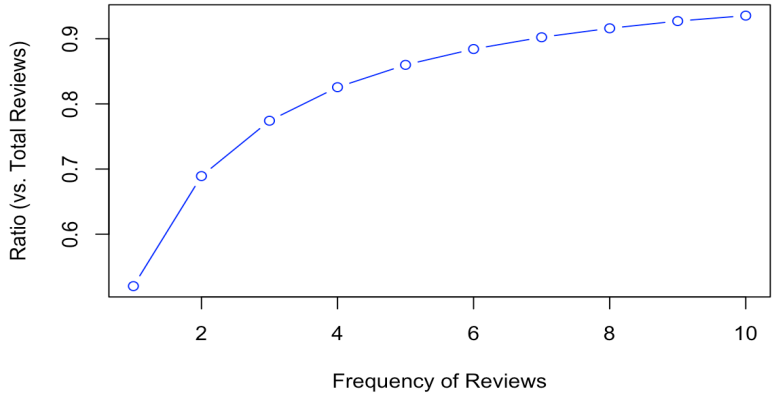


Figure 4

as.character(categories) <chr>	Count <int>
c("Mexican", "Restaurants")	71496
c("Pizza", "Restaurants")	33265
c("Restaurants", "Italian")	28811
c("American (New)", "Restaurants")	28779
c("Chinese", "Restaurants")	28187
c("Sushi Bars", "Japanese", "Restaurants")	27196

Exploration & Cleaning

- **NV and AZ Rule (Fig 5):** A view by state show Nevada and Arizona dominating with a combined 83% of counts
- **Sufficient Chinese Reviews in NV and AZ (Fig 6)**: 20K to 30K Chinese restaurant reviews per state in Nevada and Arizona

Figure 5

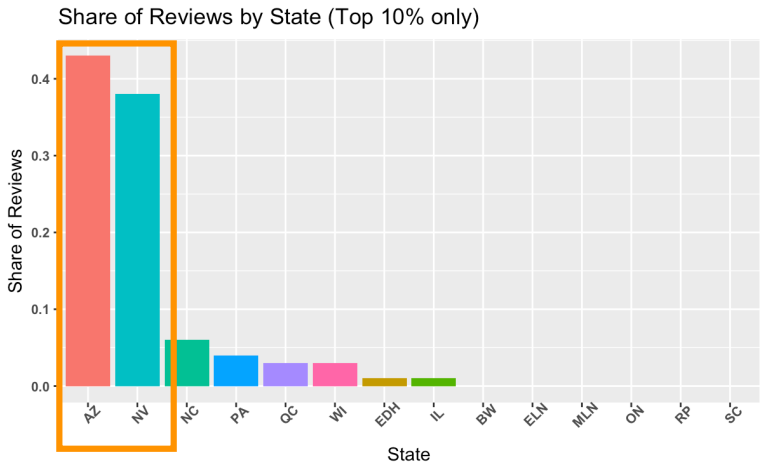
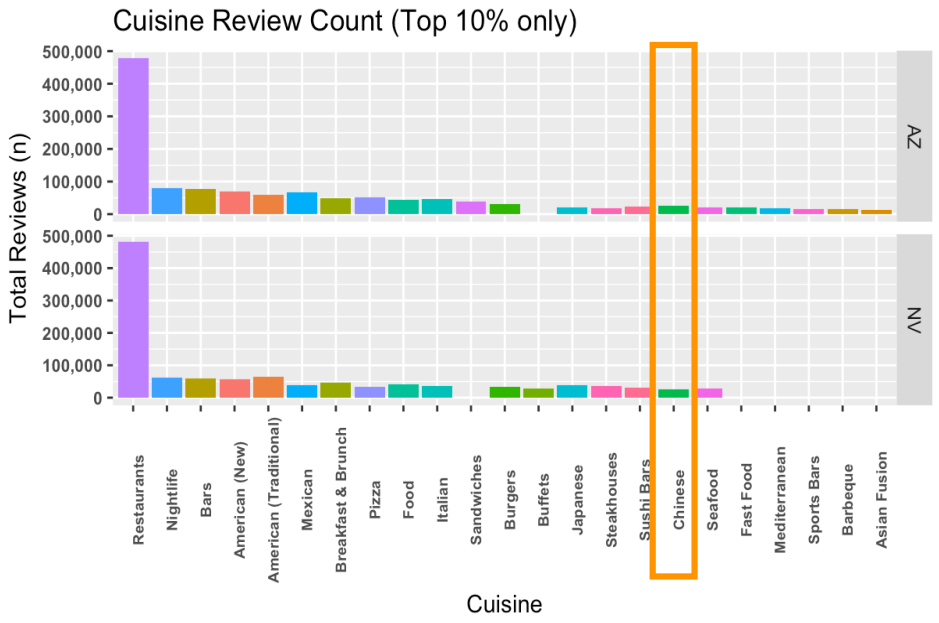


Figure 6



Text Mining in 3 Steps

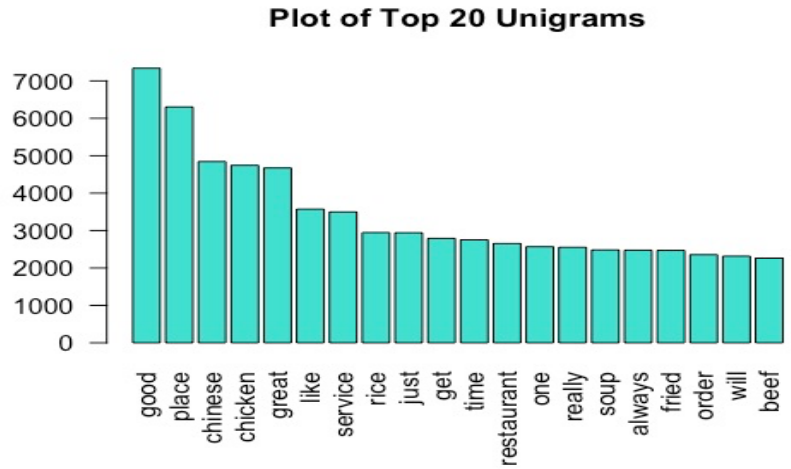
1. **Creating Corpus of Reviews:** Sampled 30% of 4- and 5-star Chinese reviews and converted to a corpus comprising 9,400 documents

2. **Text Preprocessing**: Created cleaning function to remove noise from the text - white space, abbreviation, punctuation, numbers, capitalisation and stopwords (“my”, “your”, “to” etc). This increases the precision of the function when counting the dish names and sentiment words.

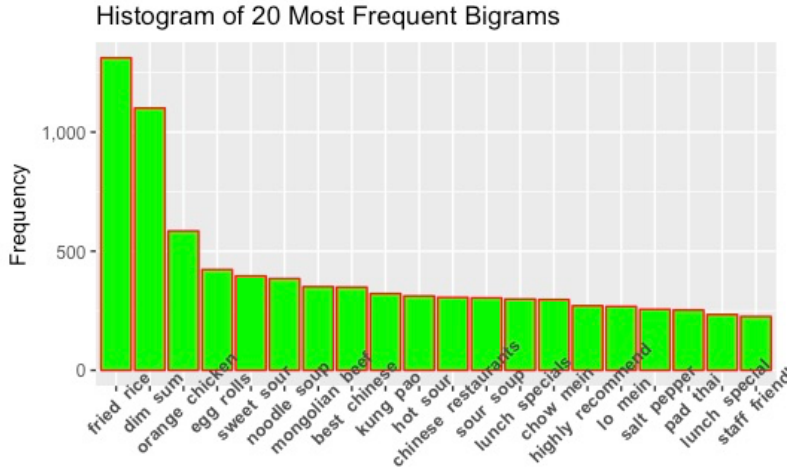
3. **Topic Mining:** Counted the occurrences of phrases comprising one word (unigram), two words (bigram) and three words (trigram) within each review in the corpus

One and Two Don’t Make a Party

- Unigrams don’t provide any useful information

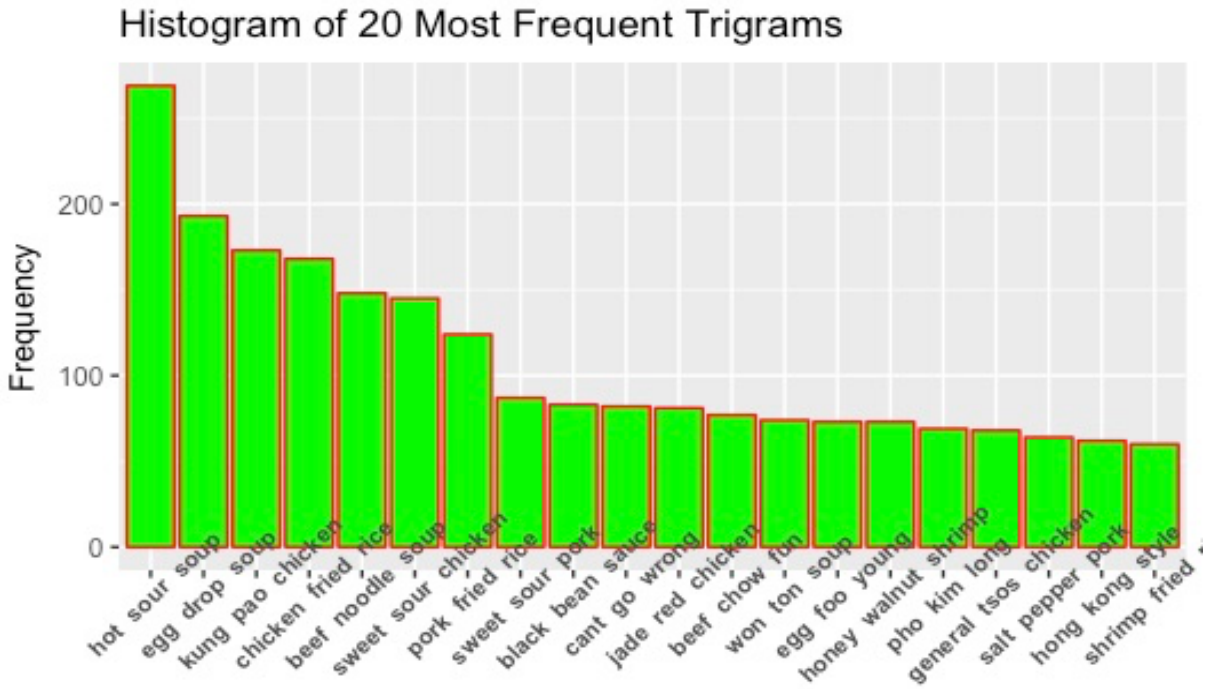


- With Bigrams, dish names start to emerge but they are still not good enough.



Magic Rule of Three

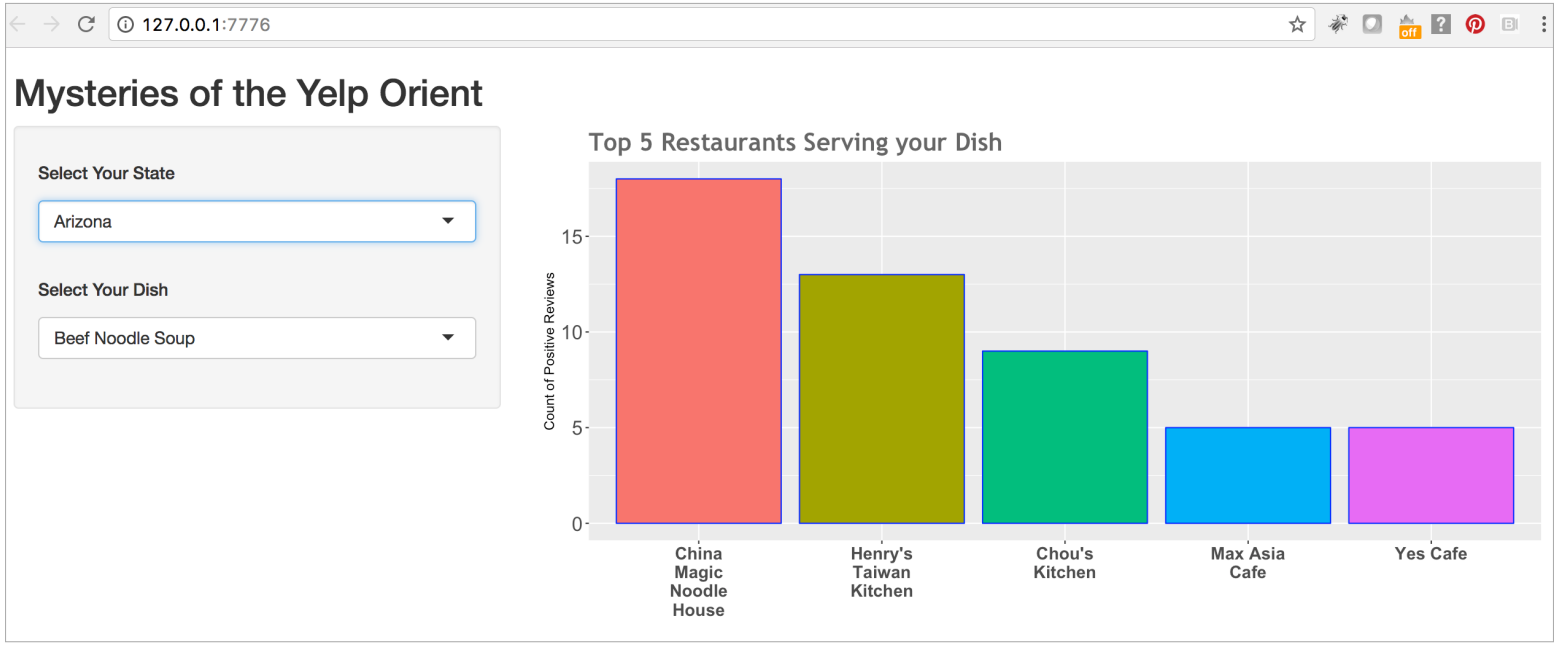
- The Trigrams finally reveal the top dishes across Arizona and Nevada.
- Hot & Sour Soup, the perennial favourite (I have no idea why) topped the list followed by Egg Drop Soup, Kung Pao Chicken, and Beef Noodle Soup.



Creating the Search App

- The Shiny App allows users to select state and dish from dropdown menus. The app searches for exact matches with positive reviews and finally returns the names of 5 corresponding restaurants with the most reviews.

- Results indicate that **P. F. Chang** is a popular chain for many well-known Chinese dishes in Nevada and Arizona.



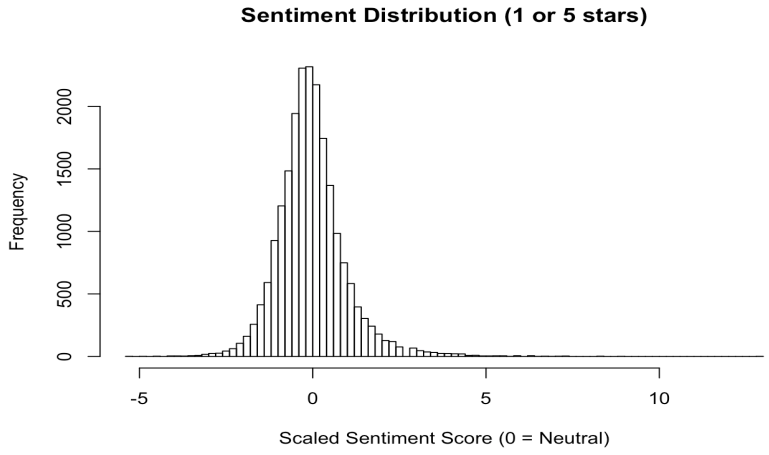
Target Audience at Yelp

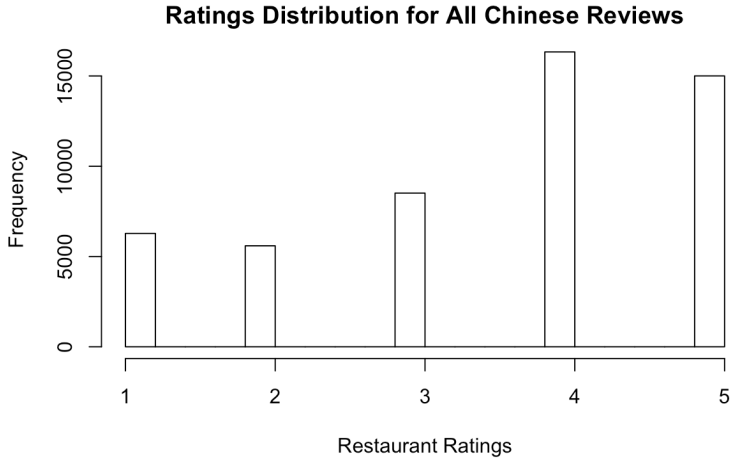
- The Search team inject an expanded version of my project into their search algorithm, allowing users unaccustomed to American-style Chinese food to enjoy the benefits of using Yelp.

- The Ads Product team invites businesses to beta test ‘bidding for dish name appearing in search queries’; doing so will improve the precision of their ad targeting and drive more high-intent traffic to their advertisers' pages.

Future Work

- Many high-quality papers have been written for the Yelp Dataset Challenge around predicting ratings and votes with machine learning.
- In the chart below, the review sentiment scores are normally distributed while correlation with ratings are moderate at best. My hypothesis is that sentiment is probably not the best predictor of ratings…





#ThankYou

Appendix

Github:
github.com/eugenewoo/Springboard_Capstone

Online Report:
rpubs.com/eugenewoo/yelp_sbcapstone