Mysteries of the Yelp Orient

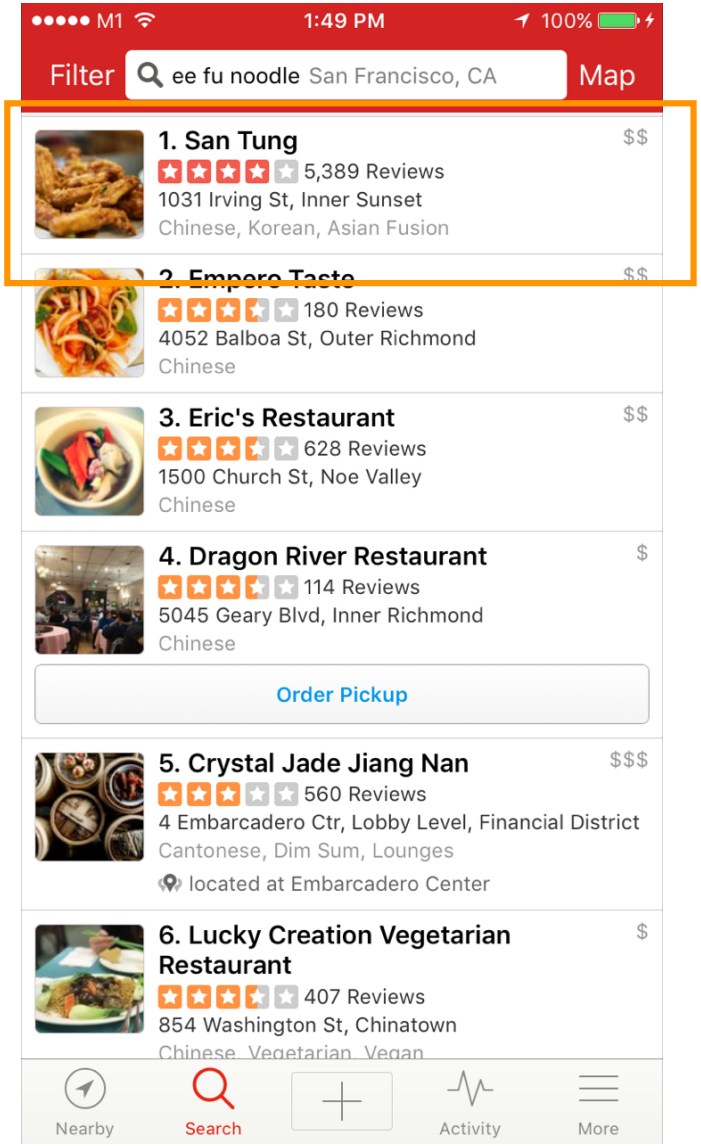
Proposal for Dish-Driven Chinese Restaurant Recommendations

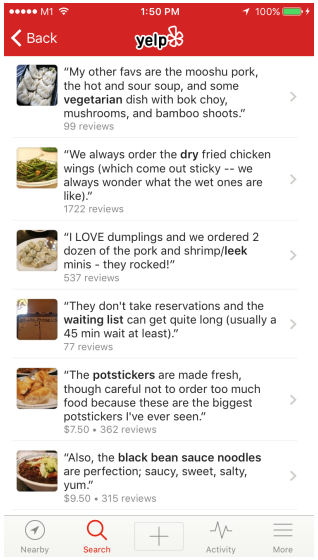
Eugene Woo
m: (65) 9362 1840
url: sg.linkedin.com/in/eugenewoo

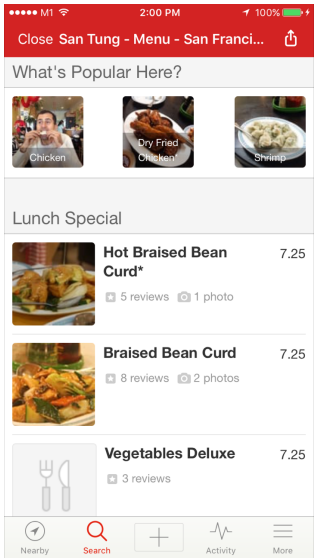
Situation

- Yelp does a reasonably good job of recommending restaurants with a top-down approach - proximity, star rating and review count
- During a trip to San Francisco, however, I had a hard time locating a Chinese restaurant that served a dish called ‘Ee-Fu Noodles’. Currently the results for a dish query favour highly-rated Chinese restaurants whose reviews contain the word ‘noodles’ but do not match the dish.
- For first-time visitors to a city, the problem is selecting the best Chinese restaurant out of a ton of listings with similar ratings. My proposal flips the problem upside down and starts with the dish itself.

Situation







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

- **Exploration & Cleaning**
  - Ensure we have sufficient sample of Chinese reviews in Nevada and Arizona from active users
- **Text Mining**
  - Count the single-word, two-word and three word-phrases in corpus of reviews to identify the most reviewed dishes.
- **Create Search Engine from 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

- **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 future loading.

- I focused on Nevada and Arizona data as they contributed a combined 83% of all reviews and also a sizeable number of Chinese restaurant reviews.

Exploration & Cleaning

- **Scrub Inactive Users:** Taking out users who wrote only 1 review removes 13% of observations (Figure 1)
- **Sufficient Overall Chinese Reviews**: 15% of reviews for the Top 5 restaurant categories were for Chinese (Figure 2)

Figure 1

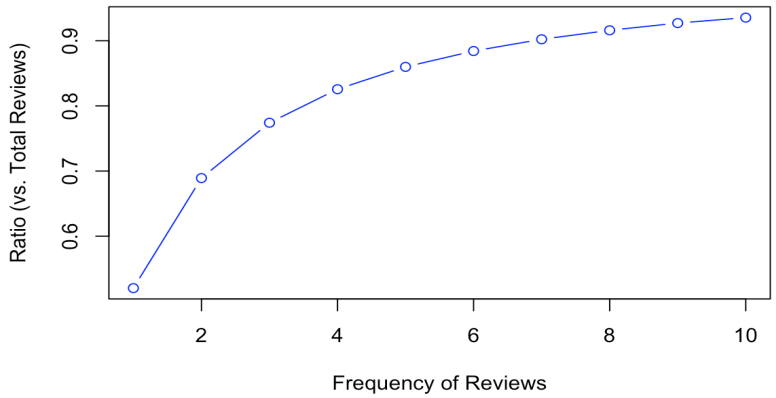


Figure 2

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:** A view by state show Nevada and Arizona dominating with a combined 83% of counts (Figure 3)
- **Sufficient Chinese Reviews in NV and AZ**: 20K to 30K Chinese restaurant reviews per state in Nevada and Arizona (Figure 4)

Figure 3

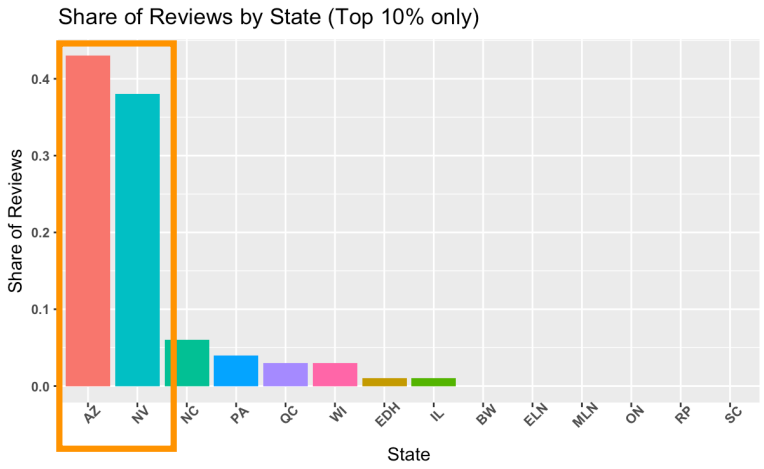
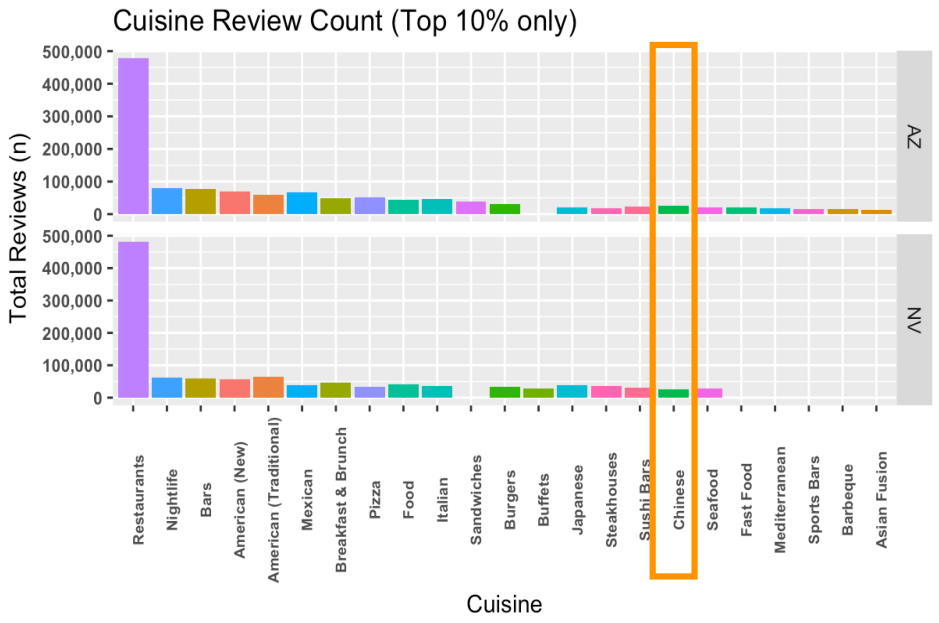


Figure 4



Text Mining in 3 Steps

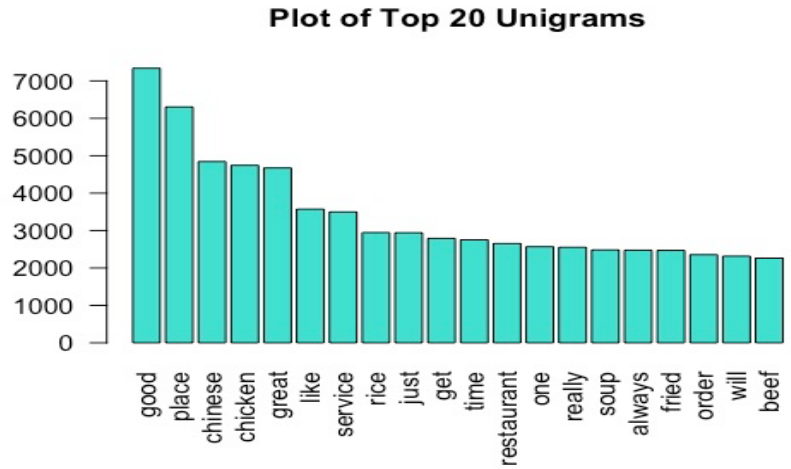
1. **Creating Corpus of Reviews:** Sampled 30% of 4- and 5-star Chinese reviews and converted to a corpus comprising 9,400 documents

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.

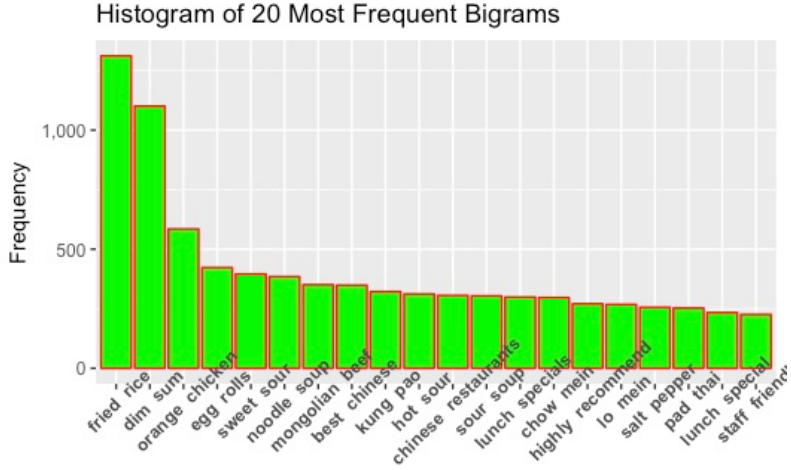
3. **Topic Mining:** Counted the occurrences of phrases comprising one word (unigram), two words (bigram) and three words (trigram) within each review in the corpus

Three is the Magic Word

- Unigrams don’t provide any useful information

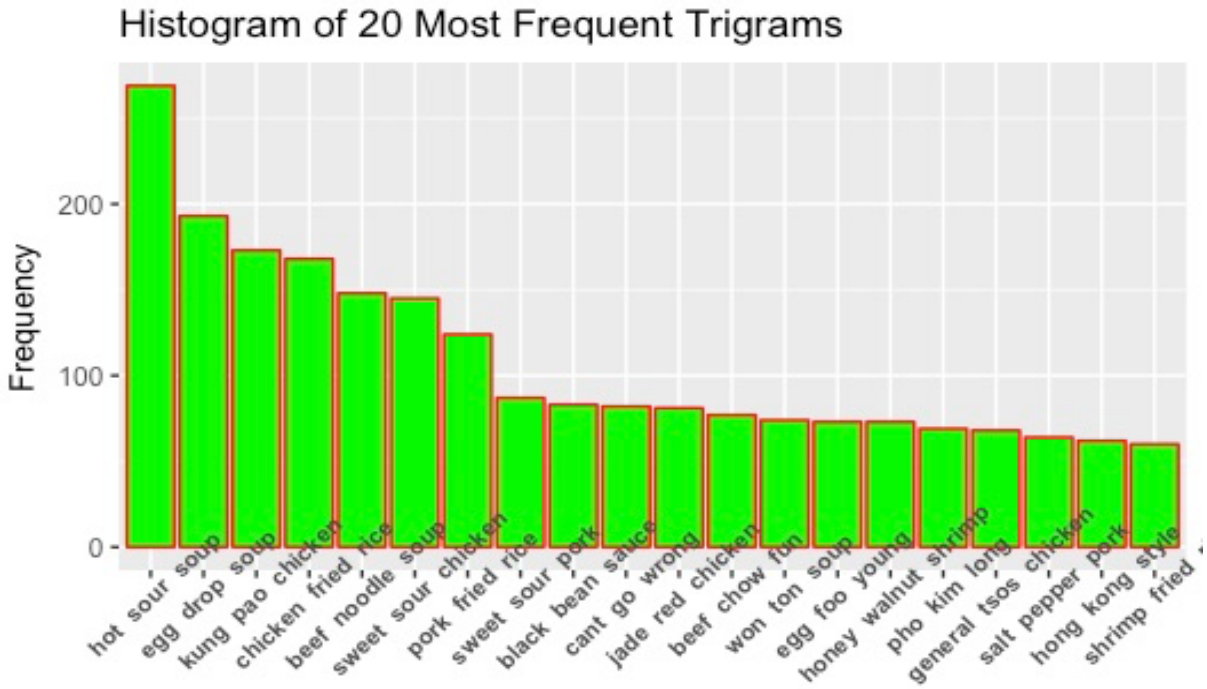


- With Bigrams, dish names start to emerge but they are still not good enough.



Three is the Magic Word

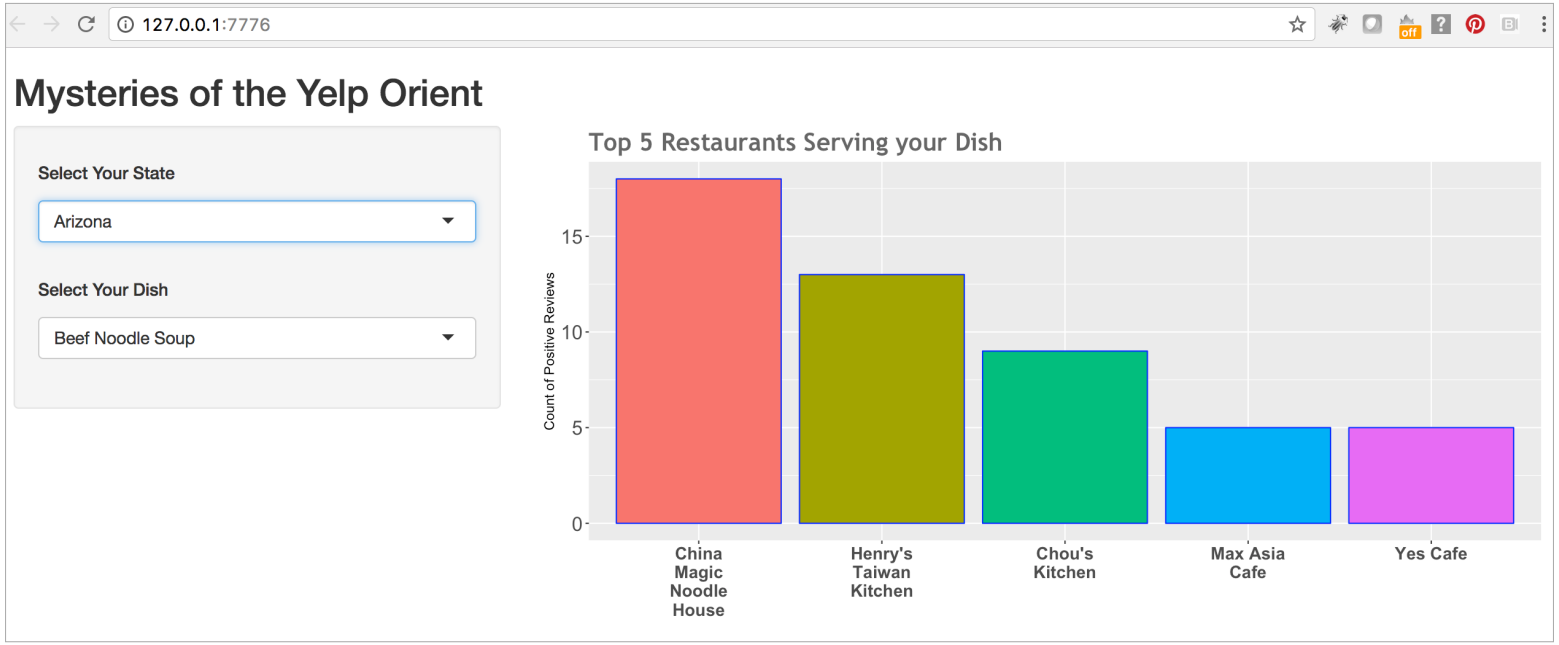
- The Trigrams finally reveal the top dishes across Arizona and Nevada.
- Hot & Sour Soup, the perennial favourite (I have no idea why) topped the list followed by Egg Drop Soup, Kung Pao Chicken, and Beef Noodle Soup.



Creating the Mini Search Engine

- 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

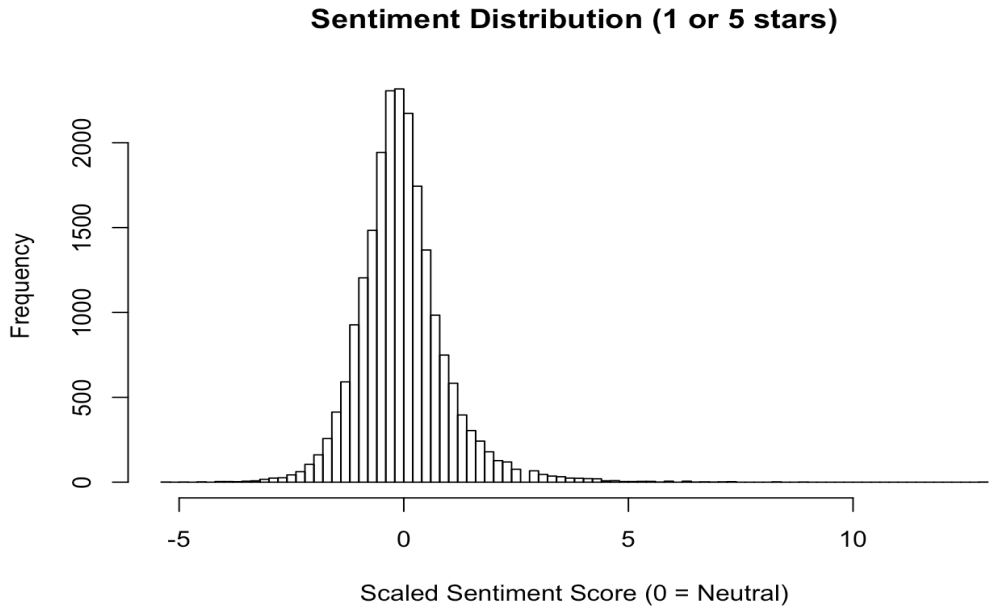
- The Search team inject an expanded version of my project into their search algorithm, allowing users unaccustomed to American-style Chinese food to enjoy the benefits of using Yelp.

- 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 sentiment for extreme ratings of 1- and 5-stars remain normally distributed and I'd argue that review sentiment is probably not a good predictor of ratings. BUT that's the subject of my next project!



#ThankYou

Appendix

Github Repository:
github.com/eugenewoo/Springboard_Capstone

Online Report:
rpubs.com/eugenewoo/yelp_sbcapstone