Yelp Data Analysis

Han Liao, Qiaochu Yu, Yujie Zhang, and Lingfeng Zhu

11/19/2019

Overview

Introduction

- Yelp allows users to give reviews to businesses
- Our project goal
  1. Provide useful, analytical insights to business owners on Yelp.
  2. Build a web based APP to visualize the analysis.

Our Dataset

1. Attributes of 4628 businesses in the business data.
   - Data Cleaning
   - Imputation
   - Analysis
   - Recommendation
2. Review texts of 403941 reviews in the review data.
   - Tokenization/Lowercase/Lemmatization
   - Significant words
   - Analysis
   - Recommendation

Data Cleaning and Imputation

- Filter out the attributes with more than 90% of missing
- Use majority voting from the 5 most similar restaurants to substitue the missing values in the remaining attributes (KNN with k=5)

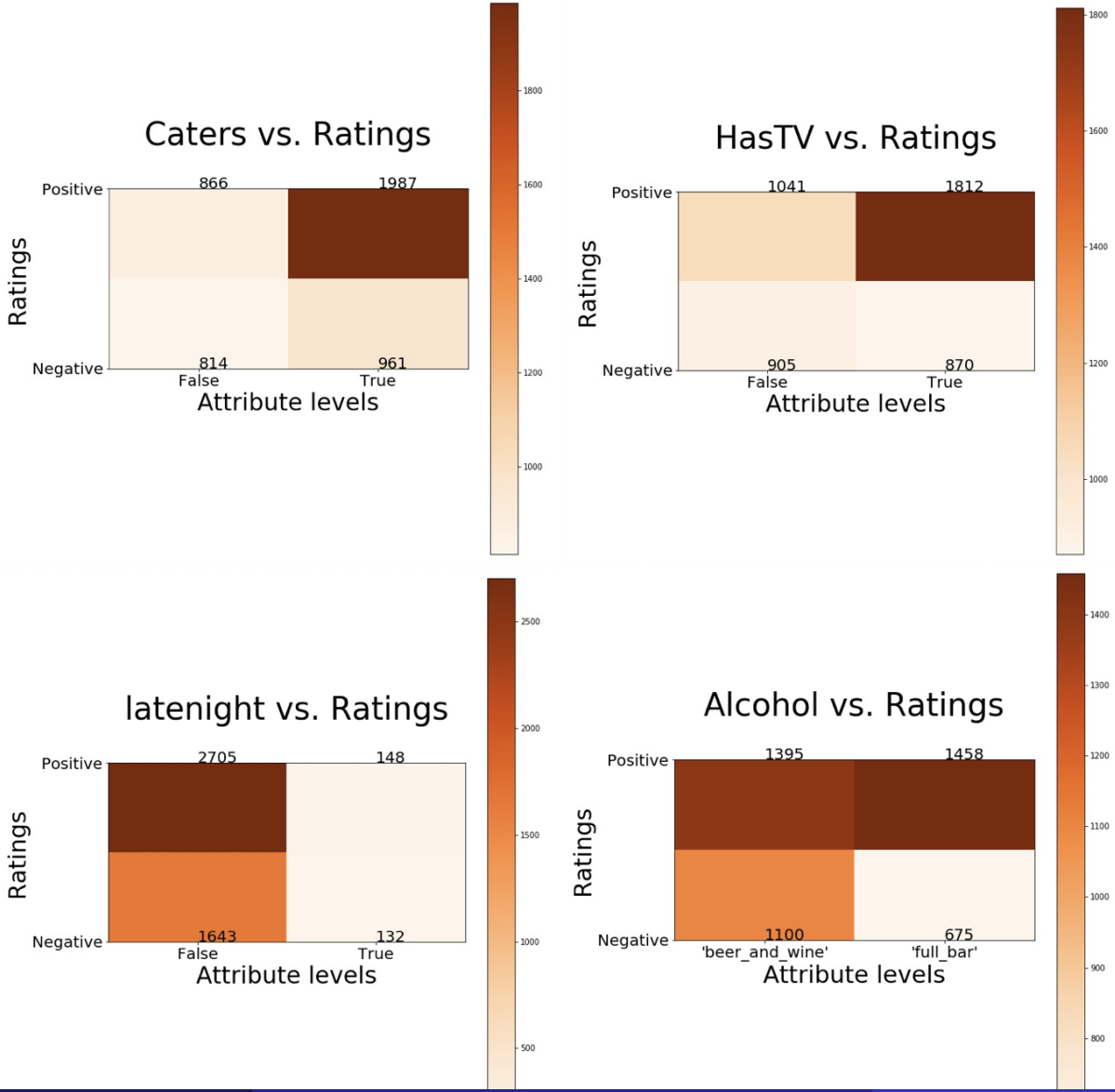
Analyze Attributes

- Fit step-wise linear regression model (using *AIC* as criterion) with the formula: *rating* $\sim$ *attributes*
- Categorize businesses into two groups based on their ratings: 3.5 $\sim$ 5.0, and 1.0 $\sim$ 3.0.
- χ^2 test on attributes cross ratings and dropped attributes that are not significant (OutDoorSeating, WiFi and GoodForKids):

Attribute	Chi_square_p_value	Attribute	Chi_square_p_value
intimate	6.218e-03	NoiseLevel	8.050e-05
hipster	5.350e-06	RestaurantsTableService	4.986e-30
divey	5.737e-04	RestaurantsPriceRange2	2.985e-02
trendy	2.007e-15	BikeParking	7.094e-03
casual	1.104e-14	HasTV	3.505e-22
dessert	1.794e-07	Alcohol	5.284e-18
latenight	2.235e-03	RestaurantsGoodForGroups	1.455e-02
lunch	4.136e-11	RestaurantsDelivery	2.419e-08
dinner	3.136e-25	BusinessAcceptsCreditCards	6.745e-06
brunch	2.081e-09	WheelchairAccessible	2.827e-04
RestaurantsReservations	3.953e-11	park	1.092e-16
Caters	2.056e-26		

Analyze Attributes

Some contingency tables for significant attributes are as following:



Recommendations

Divide the significant predictors into two lists: one for positive predictors (predictors with positive coefficients), the other for negative predictors (predictors with negative coefficients). We will give the business owners suggestions based on such results:

Positve	Negative
intimate	latenight
hipster	NoiseLevel
divey	RestaurantsPriceRange
trendy	BusinessAcceptsCreditCards
dessert	WheelchairAccessible
dinner	
brunch	
Caters	
RestaurantsTableService	
HasTV	
Alcohol	
RestaurantsDelivery	
park	

Example of Recommendations

Taking business **New Mexican Grill** as example, here are two of suggestions we provide:
(business ID: voZnDQs6Hs3YpNcS-9TALg)

- **Ambiance:** Please provide intimate, hipster, divey or trendy ambience. Your estimated rating will increase by 0.49 if you provide intimate ambience, increase by 0.25 if you provide hipster ambience, increase by 0.22 if you provide divey ambience and increase by 0.25 if you provide trendy ambience.
- **Dessert:** Please provide food that are good for desserts. Your estimated rating will increase by 0.16 by doing this.

Significant Words Related to Ratings

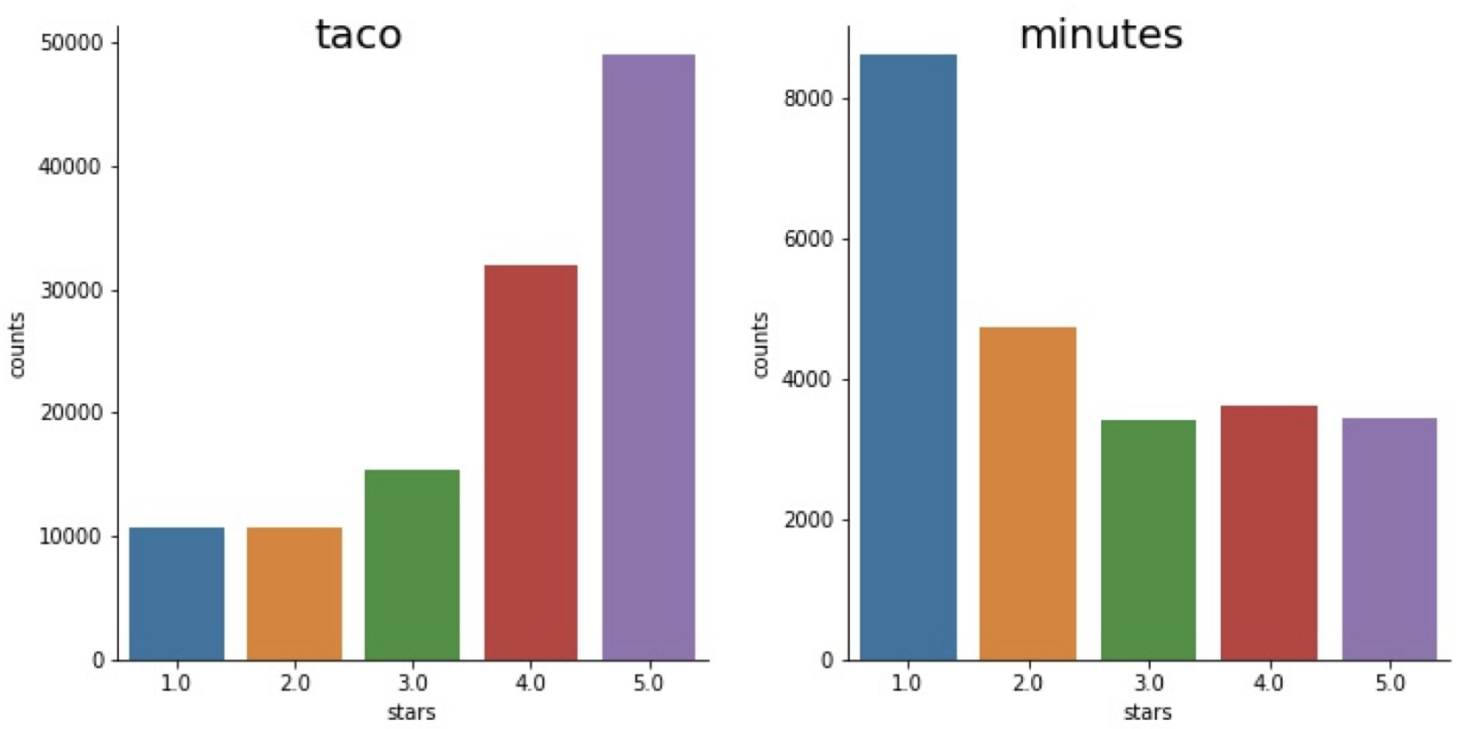
- Tokenize/lowercase/lemmatize the review texts and extract the top 1000 most frequent words.
- Manually pick up 44 words which can be basis of recommendations for Mexican restaurants.

taco	burrito	salsa	chips	rice	guacamole	asada	chipotle	tortilla	salad
nacho	enchilada	meat	time	chicken	fish	shrimp	beef	pork	steak
service	wait	minute	staff	waitress	waiter	atmosphere	price	breakfast	lunch
dinner	line	drink	margarita	table	fries	manager	beer	patio	bacon
vegan	quesadilla	chile	enchilada						

Significant Words Related to Ratings

- χ^2 testing to test the relationship between the sentiment of reviews containing those words and the ratings of reviews. e.g.

	Rating Positive	Rating Negative
Taco Positive	11832	4577
Taco Negative	949	1551



Analyze Reviews for Each Business

- The next step is to classify the reviews of each business into two categories, positive reviews and negative reviews, based on ratings.
- And then, we use **TF-IDF** to extract the 100 most "important" words in these two categories.
- At last, we get positive aspects and negative aspects for each business by getting intersections of those "important" words and significant words we get earlier.
- Below is an example of what recommendations we can give to a business owner based on reviews.

Example of Recommendations Based on Reviews

- We will go back to **New Mexican Grill**. Here are positive parts and negative parts we figured out for it.

Positive parts	Negative parts
burrito	asada
pork	salsa
chile	time
salad	
chicken	

Strengths & Weaknesses

- Our analysis can provide business owners with interpretable and clear recommendations with statistical proof.
- The model itself is too simple and the results are not accurate enough.

Web-Based App

https://1037761185qq.shinyapps.io/Tue-G7/

Reference

The negative/positive wordslist comes from:
`www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar`