What do good Yelp reviews look like?

Yelp Review Votes Prediction

Yuka Abe 2/1/2018

Details on Github: https://github.com/yukaabe/Data-Science-Projects-Portfolio-Repo/tree/master/Yelp%20User%20Review%20Quality%20Analysis

Project Workflow

#1 Define question & Goal

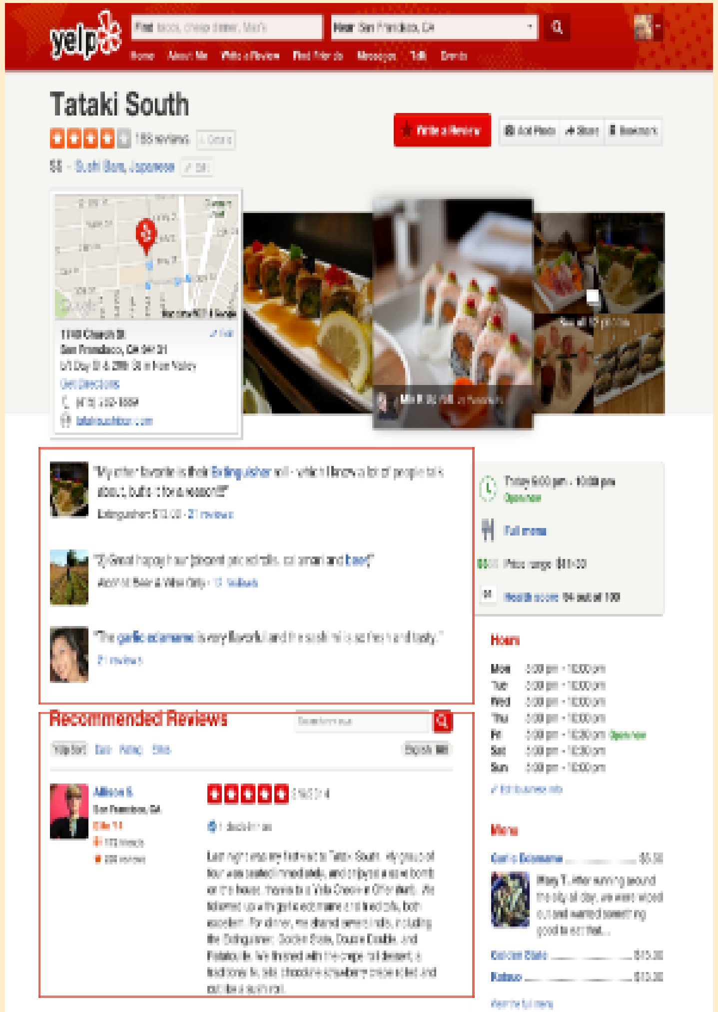
Background & Goal

- Background: 3 community-powered metrics to track the review quality: Useful, Cool, and Funny.
- Goal: Understand what the high-quality Yelp reviews look like and make predictions on the good reviews in the future.

USEFUL

Cool

Funny



Why is this important?

- Always push the most recent and good-quality reviews in the "Review Highlight"
- Get insights for developing better content advertising

#2 Data Acquisition & Transformation

Main Components of the datasets

Business

Review

Check-in

User

Data Tranformation for Modeling

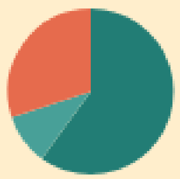
VS

- Mutually exclusive review groupings: Useful vs Not useful, Cool vs Not cool, Funny vs Not funny
- Total Review Votes = Funny + Cool + Useful votes
- Groupings based on vote count: Below average votes Vs. Above average votes

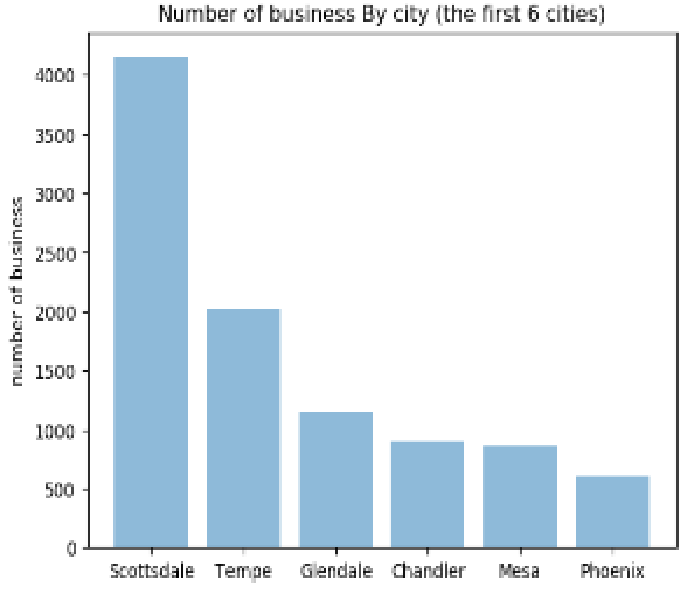
#3 Exploratory Analysis

- 11,537 business
- 99% located in Arizona
- Most of the business (80%) located in Phoenix, Scottsdale, Tempe, Mesa and Chandler.
- 90% Open businesses
- 60% of the businesses are restaurants.
- 200,471 reviews and 43,873 Users
- 94 check-ins on average

Number of business by categories



Restaurant (60%) Shopping (10%) Other (30%)



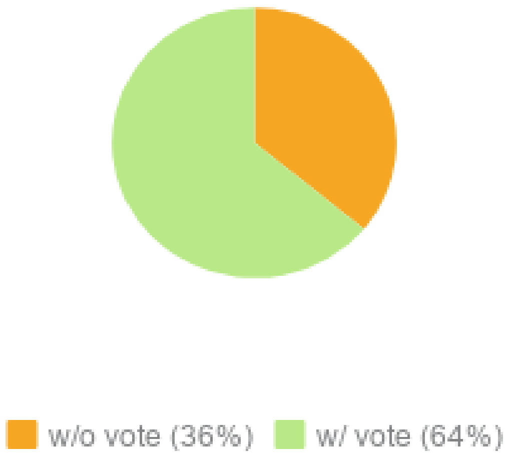
Screenshot of the Data Dictionary

Variable	Description	Type of Variable	Comment
business_id	unique identifier for the business.	text string	Overall 8281 business in the dataset
business categories	Categories of the business	categorical	Overall 1067 business categories. Example value ['Delis', 'Restaurants']
business city	The city where the business is located	categorical	68 unique cities. Example value: Youngtown
latitude	latitude of the business	continuous	
longitude	longitude of the business	continuous	
business name	Name of the business	categorical	Overall 5497 unique business names, business name to business id is one to many relationship
open	whether the business is open or closed	categorical	two unique values: True/False
business review_count	# of reviews a business has got so far	continuous	
business stars	# stars a business has got	categorical	1 - 5
review date	the date when the review was posted	date	
review_id	the id of the reviews	text string	Overall 200471 unique reviews
text	text of the reviews	text string	Overall 200297 unique review texts
user_id	User id	text string	Overall 41005 users
user average stars	the Average Star user has got	categorical	0 - 5
user review_count	# of reviews user has posted on Yelp	continuous	

Yuka Abe 2/1/2018

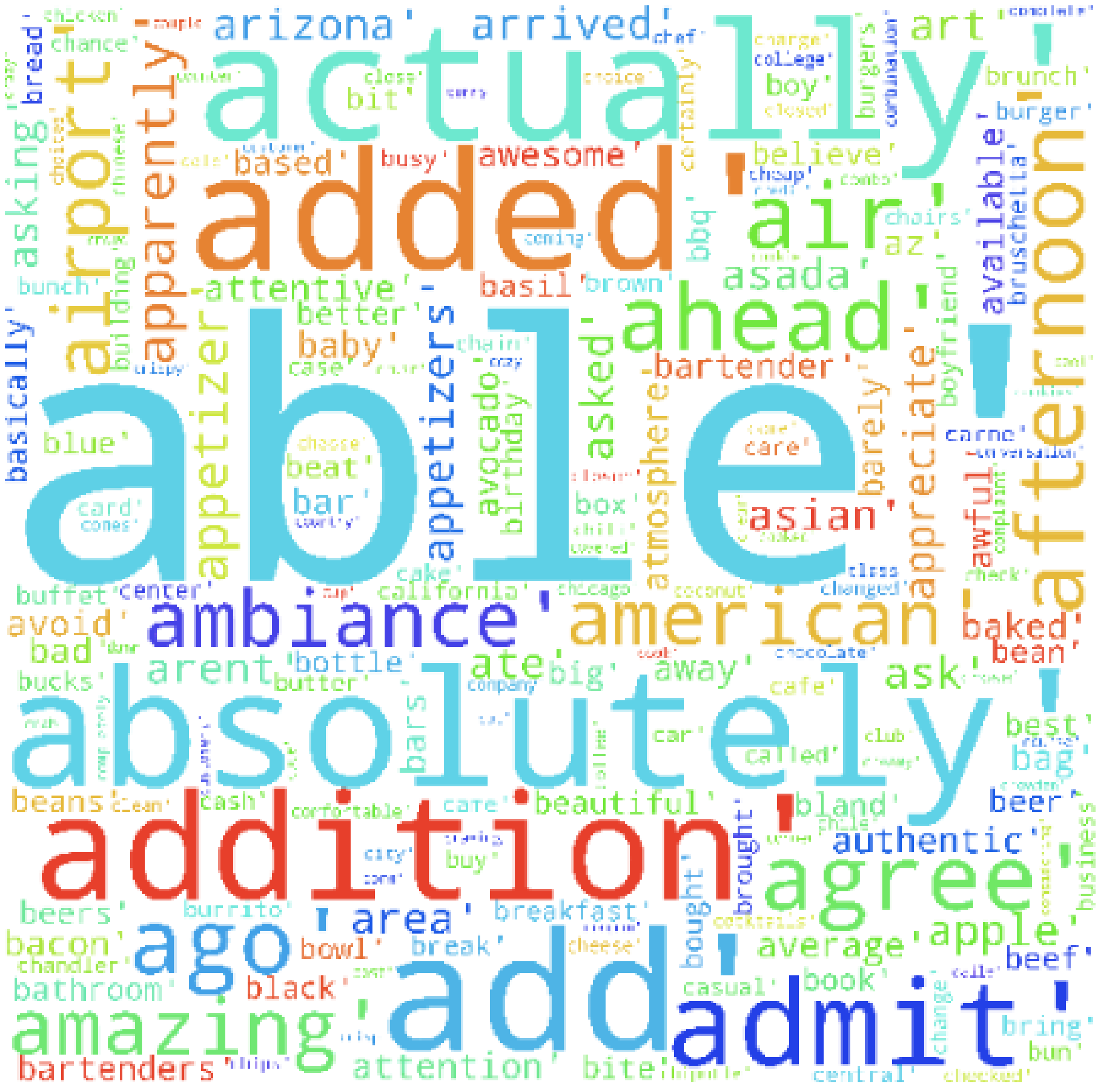
#3 Exploratory Analysis

36% of Reviews has no vote.



- 200,471 votes in total
- 30% useful vs 70% not useful
- 30% funny votes vs 70% not funny
- 37% cool votes vs 63% not cool
- Most reviews have up to 5 votes

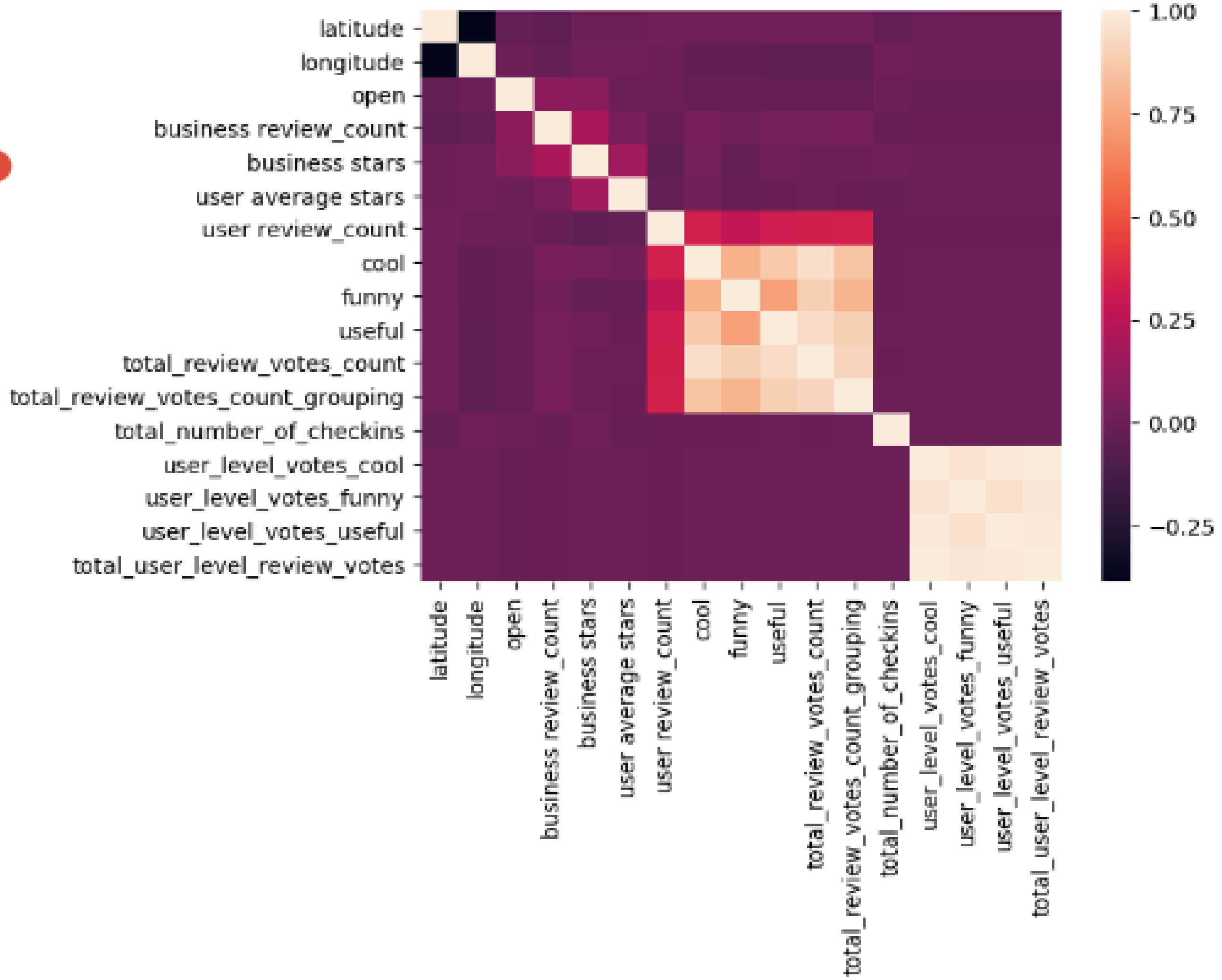
Most Frequently-used words in reviews



#3 Exploratory Analysis

Any relationship in the data?

- Useful, funny and cool votes for the reviews are closely related.
- The count of the user reviews is related to the total number of review votes



#4 Modeling & Testing

Step 1: Transform all the review texts into 31,193 sets of keywords using count vectorizer and tfidf vectorizer

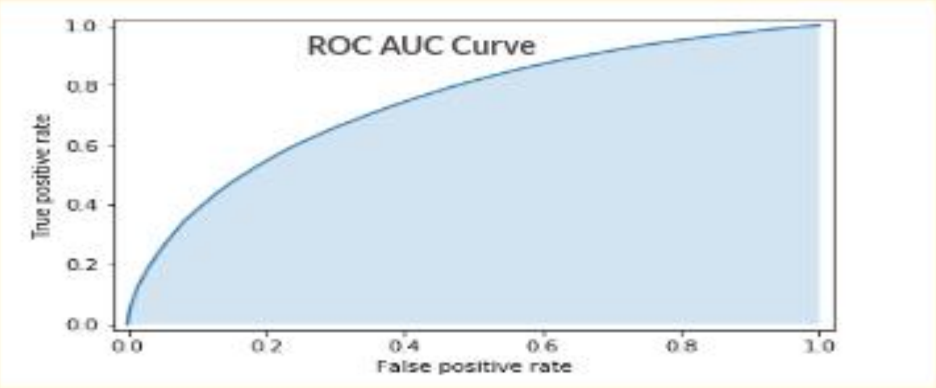
```
## Use the tfidf transformer to convert the review text to the vectors:
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf_transformer = TfidfVectorizer(ngram_range =(1,3), stop_words='english', lowercase=True, min_df=50)
X_text_train_tfidf = tfidf_transformer.fit_transform(X_text_train)
X_text_train_tfidf.shape #(200471, 31193) # 200471 samples with 31193 features
```

```
(200471, 31193)
```

Step 2: Classify reviews into categories using Naive Bayes, Random Forest and Logistic Regression

Step 3: Identify the right metric and evaluate the success of the model using cross validation

- Use ROC AUC score: Precision & Recall are both important in this case



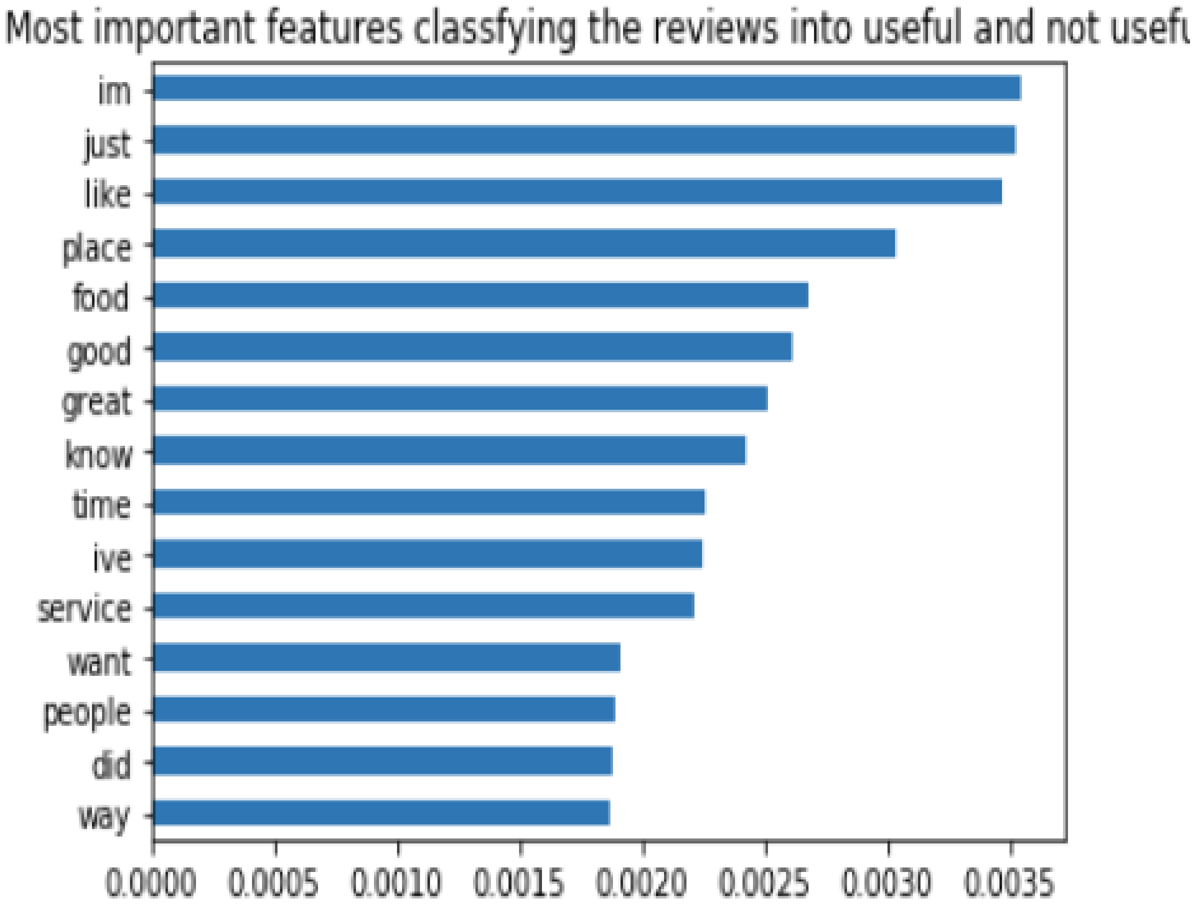
#5 Insights & Next Steps

Positive & Important keywords within the reviews for Classification

Positive/Negative keywords to group reviews into "Useful" vs "Not Useful"

(+) delicious, staff, new, come, friendly, salad, fresh, came, say, right, want, better, did, went, going, night, lunch, cheese

(-) food great ambiance, good seating, hour sushi, happy hour sushi, best indian food, wife chicken, great pho, good ribs, great bartender,toppings want, staff best, like spicy food, planning going



Yuka Abe 2/1/2018

#5 Insights & Next Steps

Positive & Important keywords within the reviews for Classification

Positive/Negative keywords to group reviews into "Cool" vs "Not Cool"

(+) eat, new, staff, come, salad, did, say, delicious, right, want, fresh, better, going, went, friendly, night, way, great, food, like, good, place

(-) rudest, half appetizers, restaurant closed, food old, inconveniencing, food needs, disappointing meal, disrespectful, server went, good seating, bad pizza, valley location, awful food

Positive/Negative keywords to group reviews into "Funny" vs "Not Funny"

(+) delicious, staff, new, come, friendly, salad, fresh, came, say, right, want, better, did, went, going, night, lunch, cheese, way, didnt, make, pizza

(-) said manager, did apologize, brought attention, recommend hotel, got orders, spoke manager, meals great, rudest, helpful service, ordered entrees, items order, manager didnt, lost business, waitress finally, little smaller, location north, lousy service

#5 Insights & Next Steps

Positive & Important keywords within the reviews for Classification

Top positive/Negative keywords to group reviews into "Above Average Votes" vs "Below Average Votes"

(+) try, dont, little, chicken, ive, staff, friendly, pizza, nice, best, time, really, just, love, like, service, place, food, good, great

(-) beautiful carin, carin, uye, bec, certainly dont, really dont think, server comes, rand, cane sugar, giggling, robyn, inevitable, wasi, forgiven, goats, translate, bottle champagne, todayi, bastards, ridden

Next Step

- Clean up the text using more advanced techniques like stemming the word
- Acquire more reivew data from other state for further analysis.

THANK YOU