Yelp Data Analysis
&
Brunch Review

Jiaming Zhou, Hongyi Jin, Jingyu Ji

Outline

- Data Cleaning
- Data Visualization
- Future Work

Data Cleaning

- Tokenize
- Lemmatization vs Stemming
- Negation
- Other Works
- Stop Words

Tokenize

Split sentences into word and punctuation (tokens) by regular expression.

Tricky problem:

(1) aren't:

are n't / aren ' t / are not

(2) co-worker:

co worker / co-worker

Lemmatization

Use a dictionary to return the original word

was/were	be
Ran	run
Timing	time
	Timing

Accurate but time-consuming

Detect the part-of-speech tag

Stemming

Cut off the ends of the word directly

Trouble	troubl
Troubling	troubl
Troubled	troubl

Simple fast but unreadable

Summarize into one token

Negation

How to treat the words after a negation?

It isn’t a good idea

It isn’t good_NEG idea_NEG

Different kind of punctuation to denote the end of a negation:

comma or period?

Other works

Transfer to lower case

Delete foreign language

Combine the special punctuations back together

Stop word

Most common words in a language

Sort the phrases by frequency ⟶ Hand-filtered

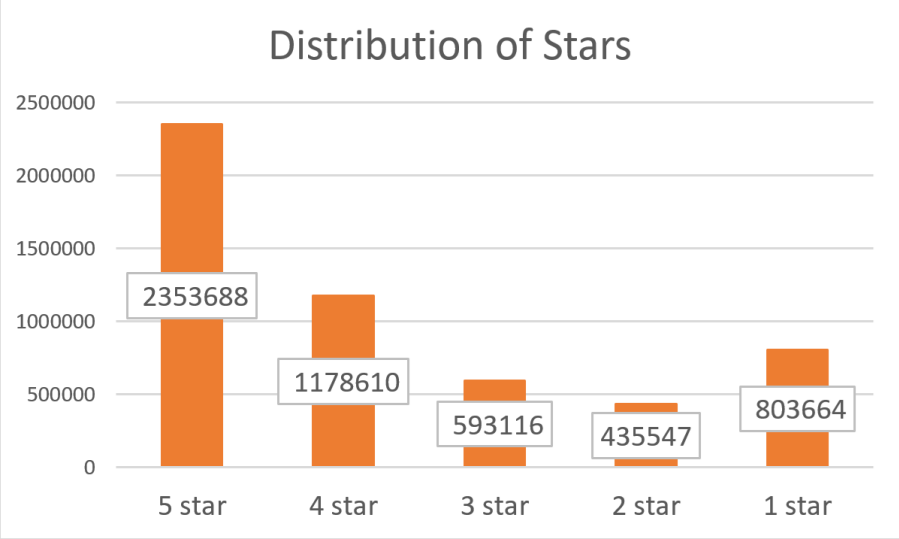
Combine nltk stopwords and the frequency in our data.

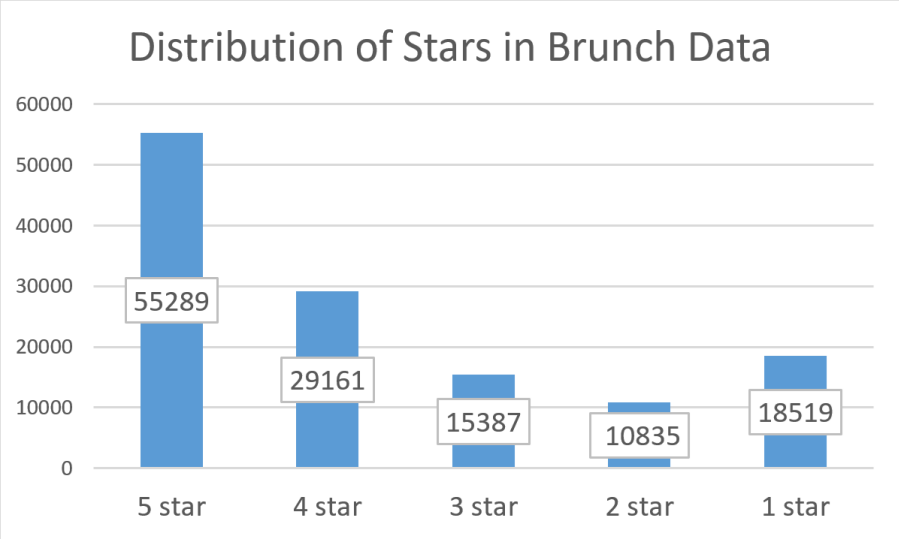
'yourself', 'there', 'about', 'with', 'an', 'be', 'some', 'for', 'do', 'such', 'of', 'am', 'or', 'as', 'from', 'the', 'below', 'are', 'your', 'through', 'me', 'this', 'above', 'both', 'up', 'to', 'had', 'at', 'any', 'same', 'and', 'been', 'have', 'in', 'will', 'on', 'does', 'yourselves', 'then', 'that', 'so', 'did', 'you', 'has', 'just', 'where', 'too', 'myself', 'which', 'i', 'few', 'whom', 'being', 'if', 'my', 'a', 'by', 'doing', 'it', 'how', 'further', 'was', 'here', ...

Data Visualization

- Whole Dataset
- Brunch Dataset

Star Distribution





Almost have the same distribution

Punctuations

.	,	'	!	-	(	"	…	$	)
35510545	19678129	11096689	4641524	3154390	2040996	1605218	1219153	1079932	1032210
19.84%	11.00%	6.20%	2.59%	1.76%	1.14%	0.90%	0.68%	0.60%	0.58%
876660	496862	271619	112957	80048	40169	40169	31051	27045	25889
19.69%	11.16%	6.10%	2.54%	1.80%	0.90%	0.90%	0.70%	0.61%	0.58%

- 24224 punctuations in the whole dataset
  First ten punctuations achieves 45.29% occurrence.

- 2109 punctuations in the brunch dataset
  First ten punctuations are the same; achieve 44.99% occurrence.

Special Punctuations

Multiple Punctuation Combination

!!	!!!	:)	!!!!	?!
491043	323907	120686	90888	39355
0.27%	0.18%	0.07%	0.05%	0.02%
12009	8144	2967	2143	1018
0.27%	0.18%	0.07%	0.05%	0.02%

Why special? (stronger emotion tendency)

"This drink is terrible!" *vs* "This drink is terrible!!!"
"I love the atmosphere." *vs* "I love the atmosphere:)"

Word Cloud

Brunch Dataset

One-Star Brunch Dataset

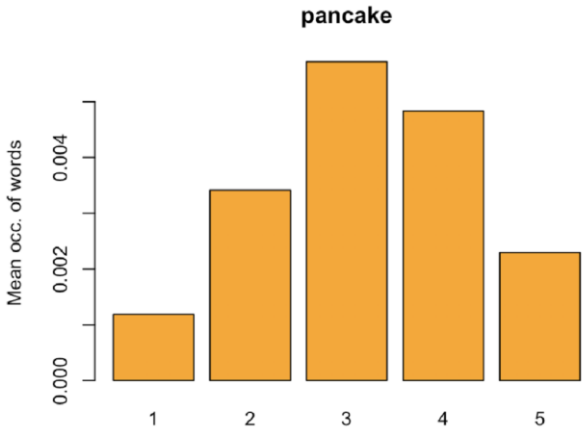
Circle the words with large difference in frequency from overall brunch review.

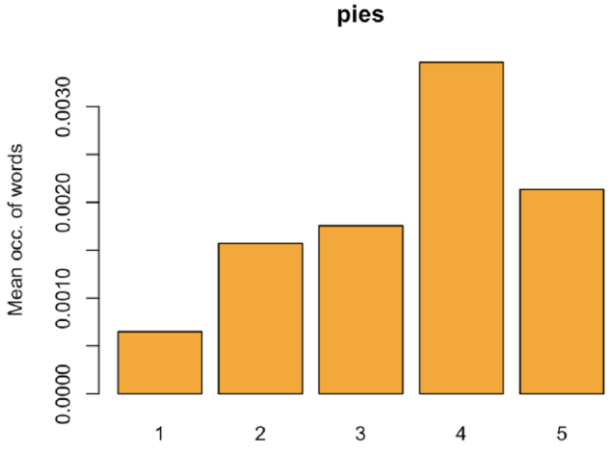
Word Cloud

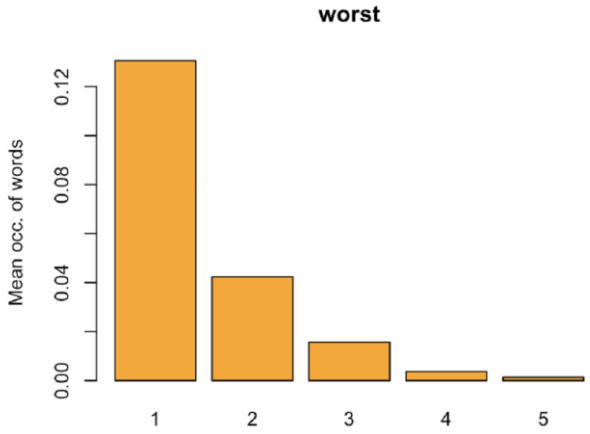
Three-Star Brunch Dataset

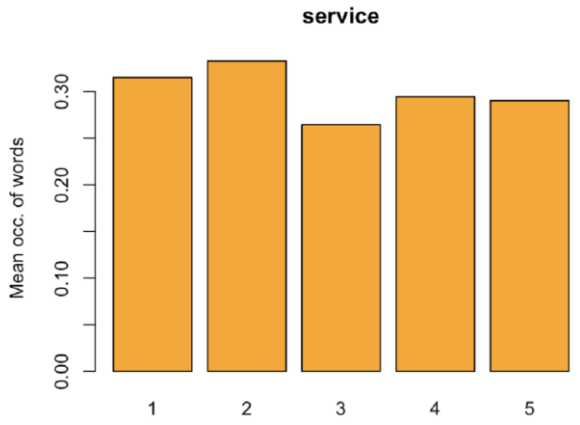
Five-Star Brunch Dataset

Mean Occurrence of Words









Future Work

- Suggestion
- Prediction

Suggestion

47506 restaurants in the dataset

4322 restaurants operate brunch business

Attributes Occurrence Frequency

High Rate Restaurant *vs* Brunch
High Rate Brunch *vs* Low Rate Brunch
High Rate Area *vs* Low Rate Area

For Low Rate Brunch Restaurant:

Whether Attributes match the Review?

The high professional frequency words' differences between High Rate Brunch Restaurant

Prediction

Overall prediction

Predict the overall sentiment of one review

Feature construction:

Frequency

tf-idf (*term frequency-inverse document frequency)*

Model

Multinomial Logistics Regression

Naive Bayes

LSTM

Specific attributes estimation:

Estimate the sentiment toward specific attributes of one restaurant, such as parking, service, atmosphere

Feature construction:

- A word list to describe each attribute
- Sentences related to each attribute
- The overall sentiment of corresponding reviews

Model
The trained model form overall prediction