

Yelp Review Data Prediction Part I

Jiyun Chen

Shiwei Cao

Jing Guo



Content

- Overview
- Text
 - Language
 - Cleaning Procedure
 - Length of Characters
 - Special Punctuation
- Categories
- Future work plan

Data Overview

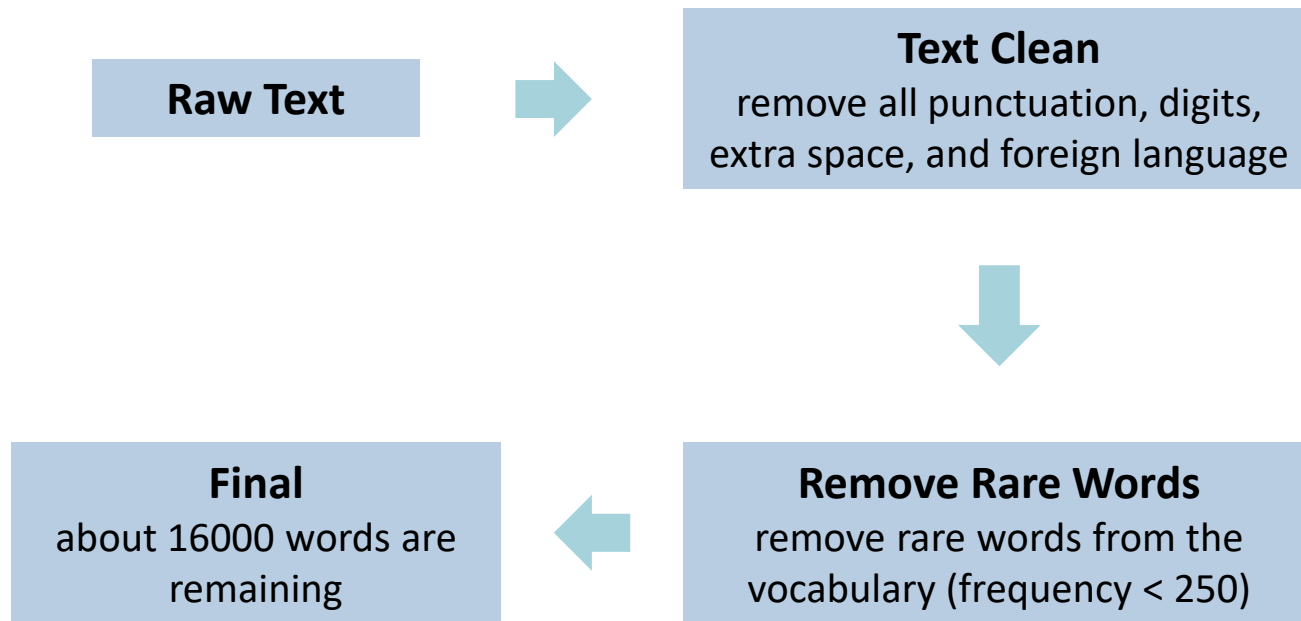
- **Overall:** 1546379 obs. of 8 variables
 - stars; name; text; date; city; longitude; latitude; categories
- **Names & location & date:** not considered
- **Categories:** 606 in train data and 611 in total
- **Text:** 464938 words in total after first step data cleaning
- **Focus on text and categories**

Text - Language

Language	English	German	French	Spanish	Italian	Japanese
Counts	1524126	13547	7190	553	187	143

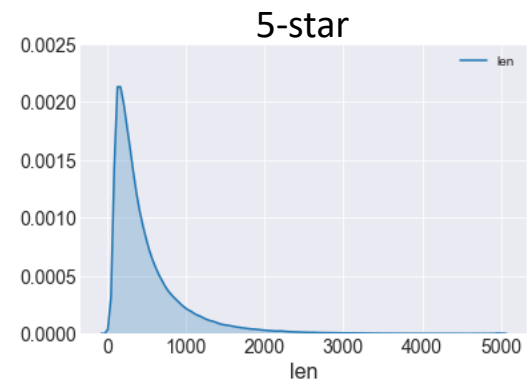
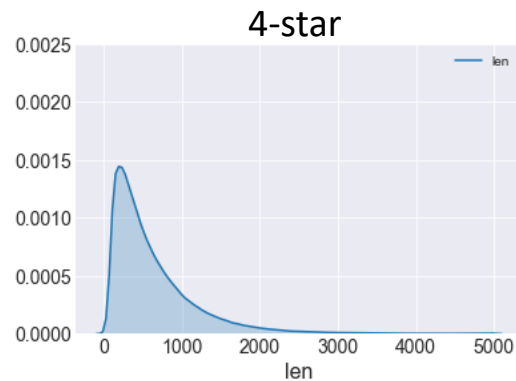
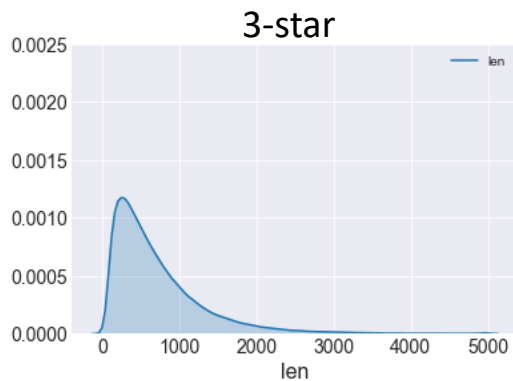
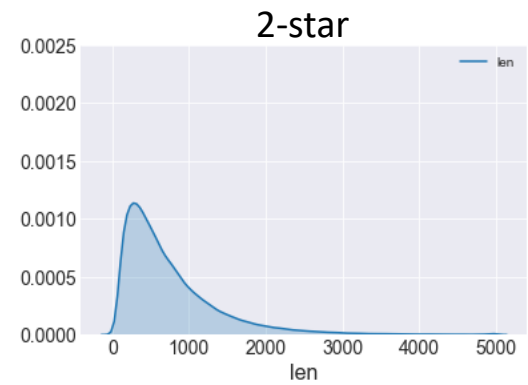
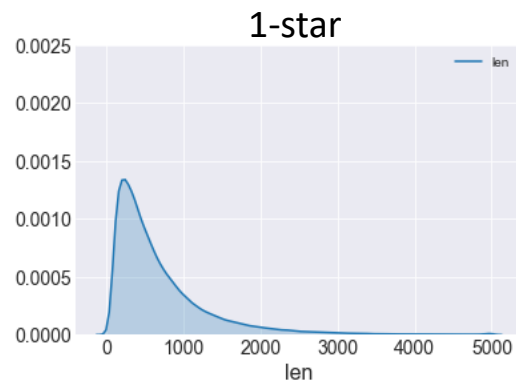
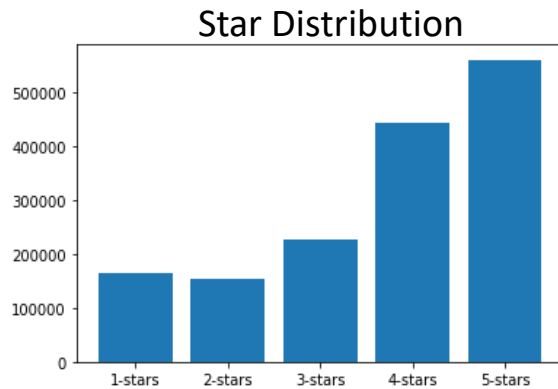
- Foreign Language: 32
- English: 98.56%

Text - Cleaning Procedure



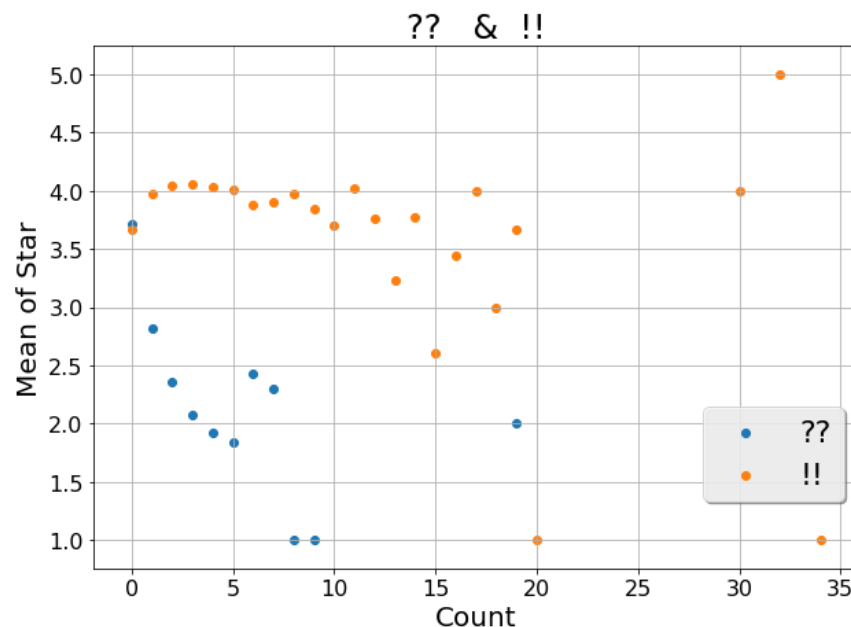
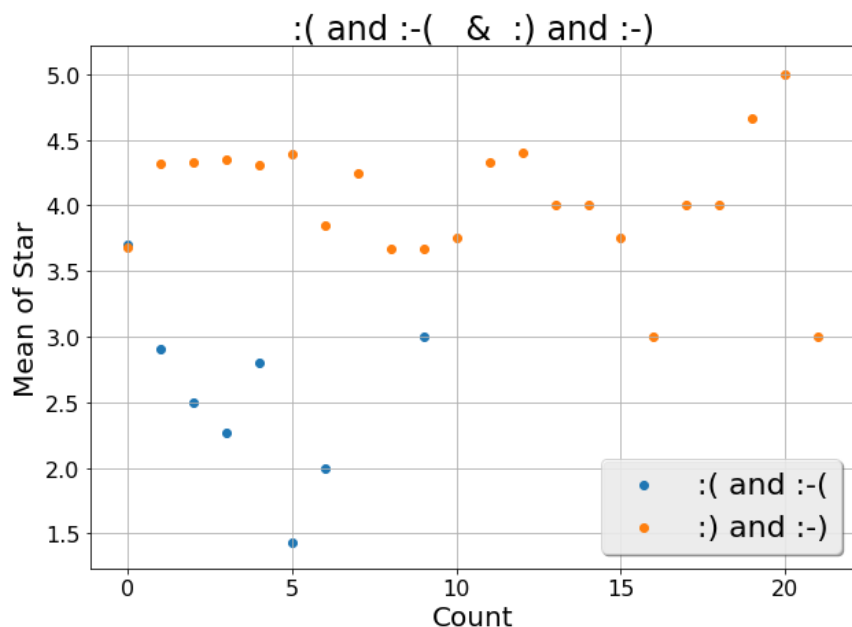
Text – Length of Characters

- Star Distribution & Text Length Based on Rating



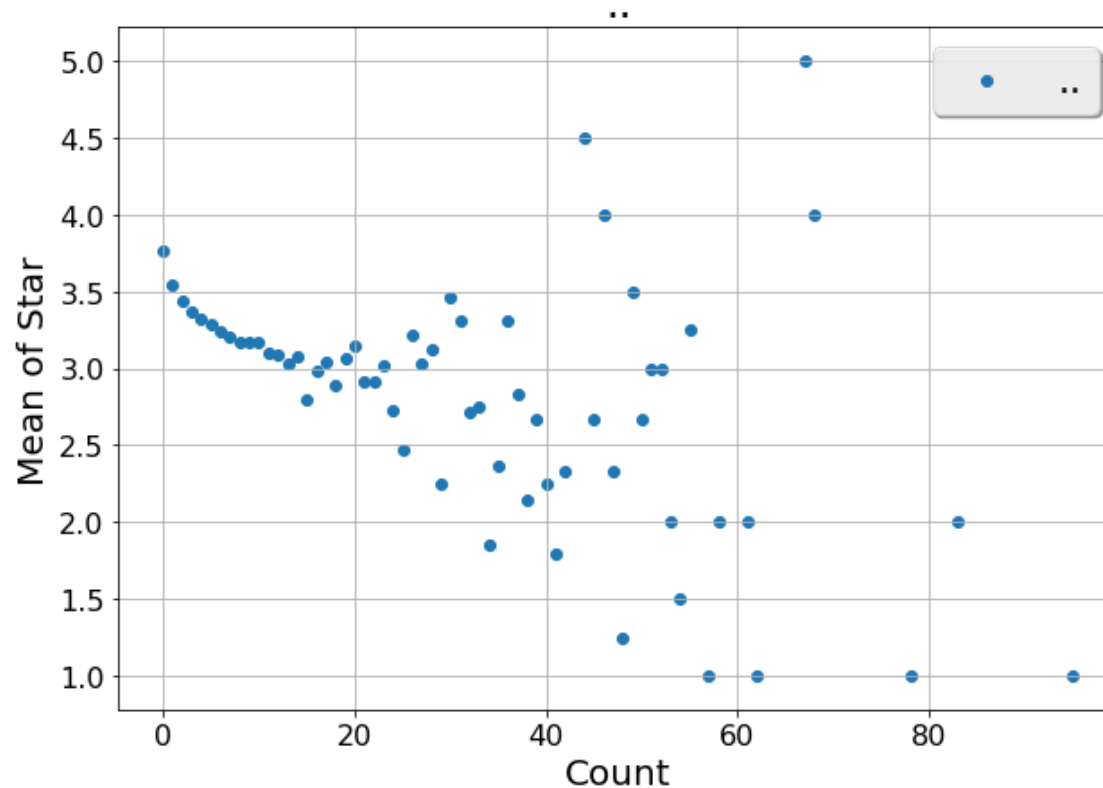
Text – Special Punctuation

Special Punctuation	..	!!	??	:) & :-)	:(& :-(
Counts	692215	253755	23956	47460	9332



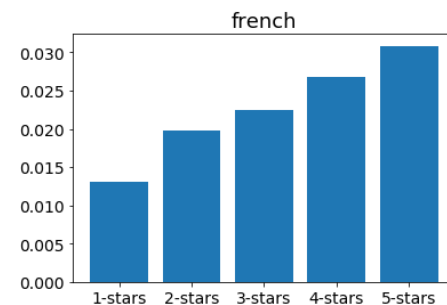
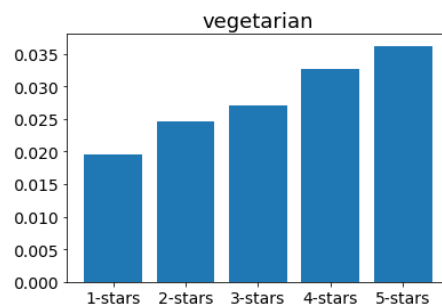
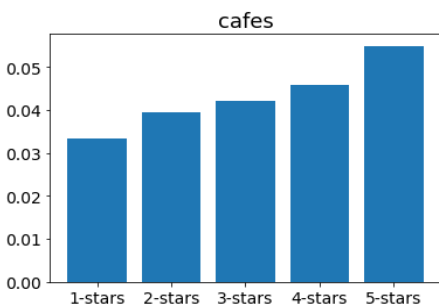
Text – Special Punctuation

Special Punctuation	..	!!	??	:) & :-)	:(& :-(
Counts	692215	253755	23956	47460	9332

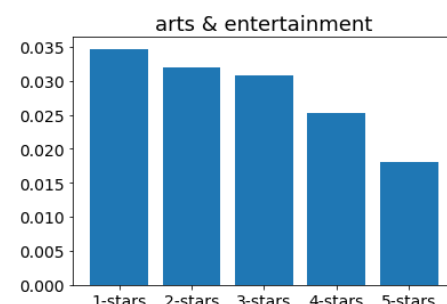
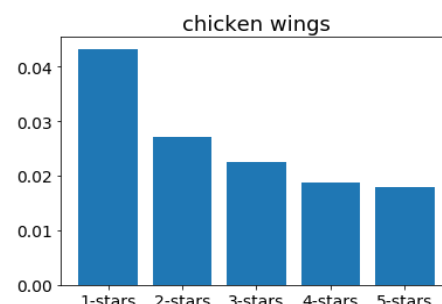
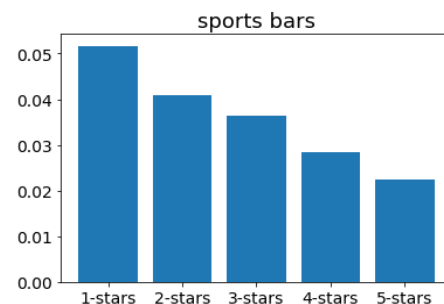
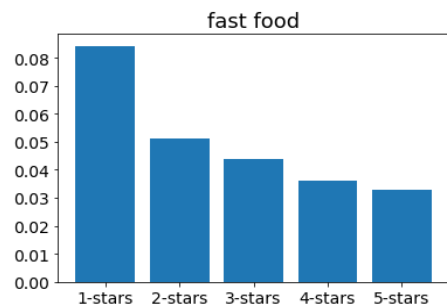


Categories

Positive Categories



Negative Categories



Categories – Word Frequency

Positive Categories Examples

Vegan & Vegetarian



French



Categories – Word Frequency

Negative Categories Examples



Future work

- **Linear model**

- LASSO: select important words
- Add interaction terms of adverbs(preposition) and adjectives:

e.g: not : good
 very : good
 however : fast

- **Neural network**

- A good method to do natural language processing

Thank You!

