Groupon Reviews Analysis

Students: Haitao Liu , Tianyu Liu, Jhaohan Chen, Lingyao Kong
Instructor: Rong Liu

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Web Mining
Spring, 2019

Introduction

- Find the advantages and drawbacks for each wine bar in New York City according to their customers' rating and reviews.
- Help the merchants have a better understanding of their wine bar and the aspects that need to be improved.
- Find the relation between the rating score , rating count and reviews

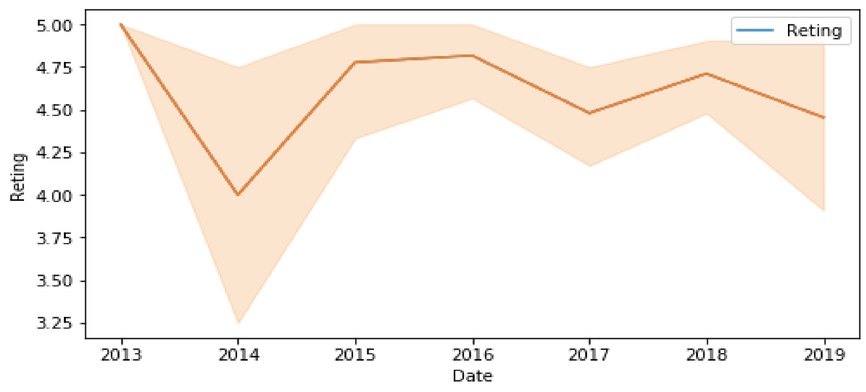
Data Understanding

- The data we used is base on 30 wine bars in New York City on the Groupon website including review, user, rating, date, rating count and review count
- We manually labeled 1000 reviews(label P/N according to emotion of reviews and add four aspects of descriptions based on each sentence.

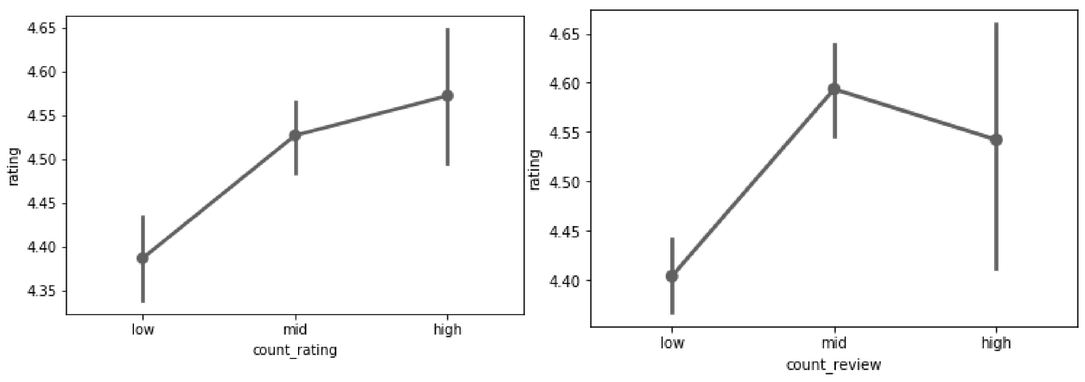
label(p/n)	Price	Service	Food	Atmosphere	Location
1	0	1	1	1	0
1	1	0	1	1	0
1	0	0	1	1	0
1	0	1	0	0	0
1	0	0	1	0	0
1	0	1	0	1	0

Exploratory Data Analysis

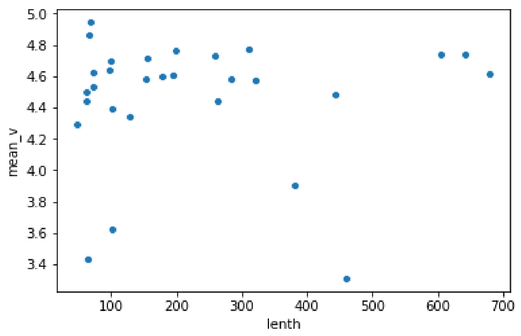
The tendency for the average rating in one store by year



The relation between the number of reviews or rating that each people has with their rating value:



The relation between the number of rating and the rating score



The word cloud shows that what kind of words appear most in each bars' reviews

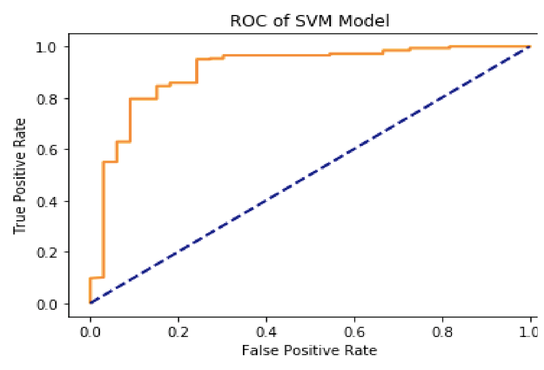
Sentiment Analysis

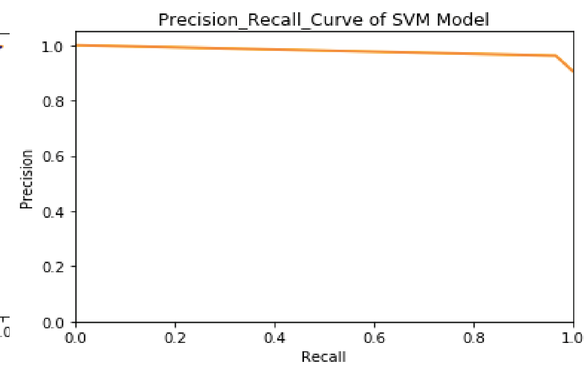
Naïve Sentiment Analysis

- We used the two dictionaries that contain positive and negative words to filter out the other words.
- If more positive words than negative, then this review is positive.
  If more positive words than negative, then this review is negative.

Supervised Sentiment Analysis

- We chose Support Vector Machine as classifier and Tf-idf matrix as a feature.
- ROC Curve and Precision-Recall Curve:





- To measure accurately, we can also created a feature according to the rating score.(rating: >4 Positive,<4 Negative)
- We can calculate the confusion matrix and accuracy according to the label

Method	SVM	Naïve Sentiment Analysis	Rating
AUC	0.84	0.86	0.92
Confusion matrix	col_0 0 1 row_0 0 14 18 1 36 361	col_0 0 1 row_0 0 60 57 1 137 1176	col_0 0 1 row_0 0 112 109 1 5 1204

Prepare for the Reviews

Data Scraping and Cleaning:

- We scraped all the data using python selenium webdriver, stored each wine bar in CSV file.
- 30 wine bars, 9000 reviews in total
- Tokenization, involves splitting sentences and words from the body of the text. Remove Punctuations, Stopword, Emoji

Part of Speech Tagging:

We can find the specific aspects that get the positive or negative reviews.

- Example:

```
[('great', 'host'), ('great', 'date'), ('cheese', 'chocolate')] [('robert', 'great'), ('host', 'great')]
```

Feature Extraction

- Performed Multi-label classification on the aspects of the reviews
- Accuracy: 0.65
- The accuracy in each aspect: