Sentiment Analysis of Yelp Reviews

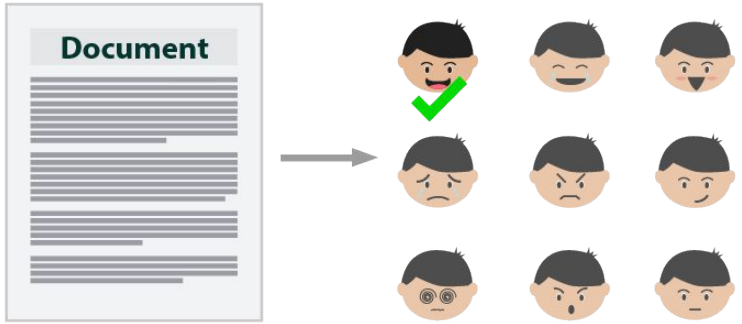
Derek Hu
- Maile Naito -
Bryan Ton

Overview

- Background
- Naive Bayes (Main Algorithm)
- Logistic Regression (Comparison Algorithm)
- Visualizations
- Demo

Background

- Why sentiment analysis?
  - Understand consumer attitudes using feedback data
  - With large consumer feedback, it helps to quantify data
  - Especially helpful for market research
- Data Set used:
  - Yelp reviews (UCI dataset)
  - 1000 reviews
  - Associated with rating: 0 (negative) /1 (positive)



Naive Bayes Classifier

- Why Naive Bayes?
  - Simple
  - Doesn't need a huge set of data
  - Easy to train and update
  - Performs fairly accurately

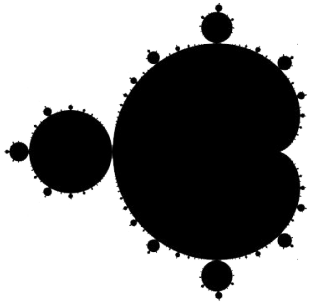
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- What are some pitfalls?
  - Data scarcity
    - Likelihood of word
  - Assumes independence of features
    - I like pasta and burgers, but I don't like pasta burgers

Implementation

- Used textblob and nltk libraries
- Removed stopwords from nltk and punctuation
  - Added words like “vegas”, “food”, “restaurant” because they have little effect on sentiment
- Stemmed words with Porter Stemmer
- List of reviews was split for training and testing
  - 70% used for training and 30% used for testing
- Achieved 82% accuracy for test data

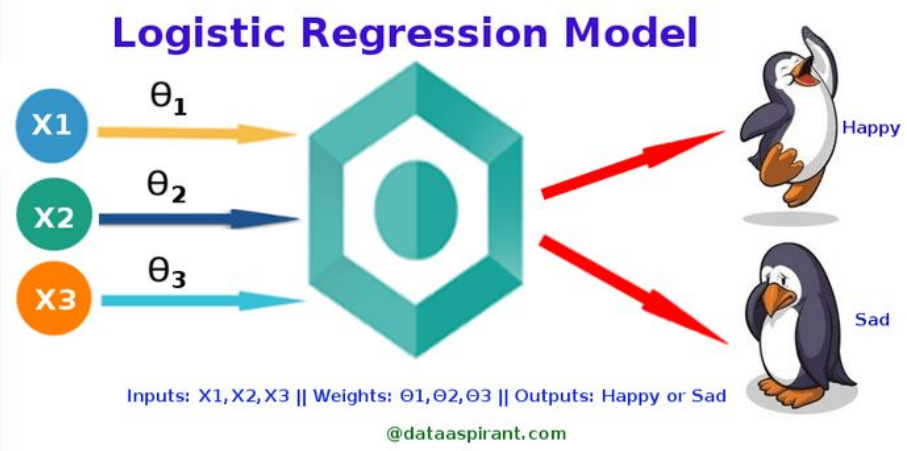
Natural Language Analysis
with Python NLTK



TextBlob

Logistic Regression

- Why Logistic Regression?
  - Useful for classifying binary dependent variables (yes vs. no, positive vs. negative)
  - Easy to implement
  - Efficient to train
  - Robust to noise/avoid overfitting



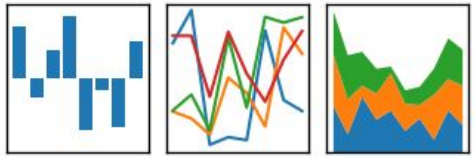
- What are some pitfalls?
  - Assumes independence of features
  - Feature selection impacts accuracy
  - Doesn't handle large number of categorical features/variables well
  - Relies on transformations for non-linear features

Implementation

- Used pandas and sklearn library
- Vectorized reviews
- Used tf-idf term weighting for training and test set
- K-fold cross validation for testing
- Achieved 83% accuracy

pandas

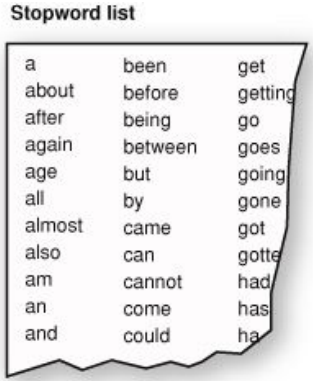
$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



scikit learn

How could we improve accuracy?

- Didn't use POS tagging, which may improve accuracy
- Implement spell correction for reviews with typos
- Try different stemmers (i.e. Snowball, Lancaster)
- Improve stopword list (Add irrelevant words, remove relevant words)



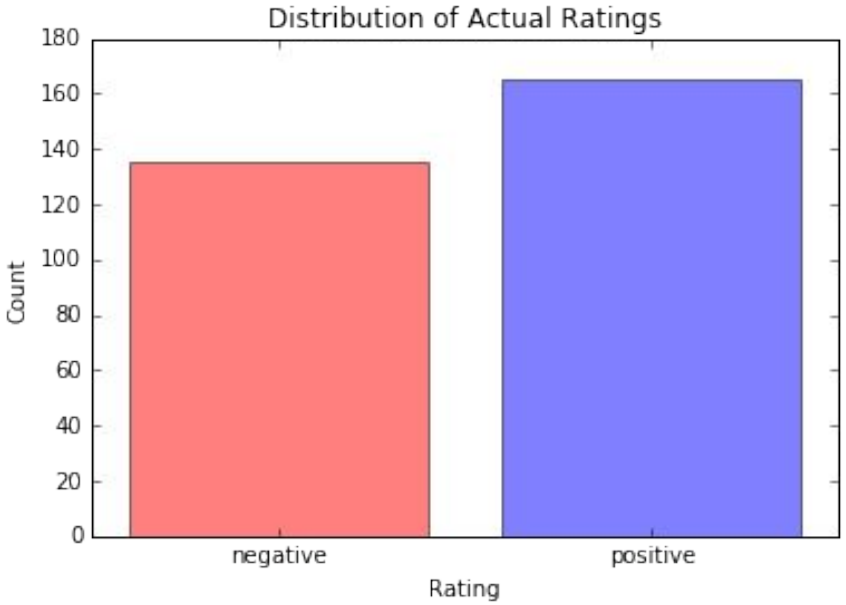
Probability Distributions Naive Bayes

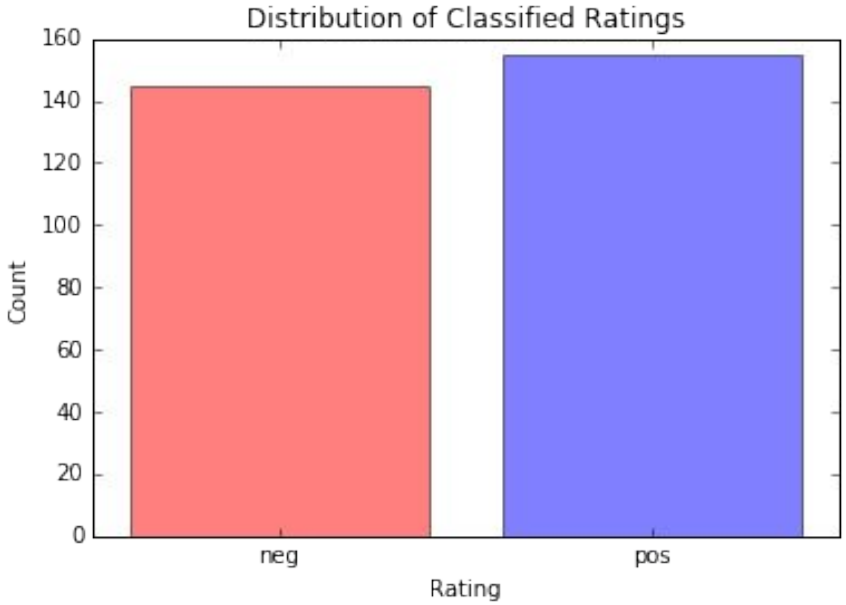
```
Wow... Loved this place.
        Positive: 96.74%; Negative: 3.26%

Crust is not good.
        Positive: 18.07%; Negative: 81.93%

Not tasty and the texture was just nasty.
        Positive: 8.45%; Negative: 91.55%

Stopped by during the late May bank holiday off Rick Steve recommendation and loved it.
        Positive: 97.28%; Negative: 2.72%

The selection on the menu was great and so were the prices.
        Positive: 99.95%; Negative: 0.05%

Now I am getting angry and I want my damn pho.
        Positive: 19.16%; Negative: 80.84%

Honeslty it didn't taste THAT fresh.)
        Positive: 67.19%; Negative: 32.81%

The potatoes were like rubber and you could tell they had been made up ahead of time being kept under a warmer.
        Positive: 47.77%; Negative: 52.23%

The fries were great too.
        Positive: 98.47%; Negative: 1.53%

A great touch.
        Positive: 97.61%; Negative: 2.39%
```

Probability Distributions Logistic Regression

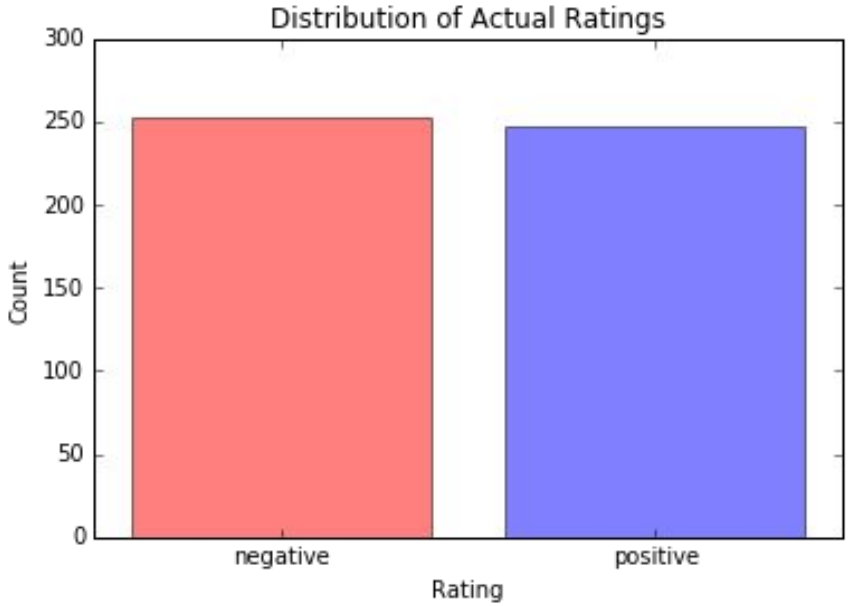
```
Crust is not good.
        Positive: 15.19%; Negative: 84.81%

Stopped by during the late May bank holiday off Rick Steve recommendation and l
oved it.
        Positive: 86.75%; Negative: 13.25%

Now I am getting angry and I want my damn pho.
        Positive: 14.84%; Negative: 85.16%

The potatoes were like rubber and you could tell they had been made up ahead of
time being kept under a warmer.
        Positive: 32.23%; Negative: 67.77%

A great touch.
        Positive: 98.27%; Negative: 1.73%

Would not go back.
        Positive: 3.15%; Negative: 96.85%

I tried the Cape Cod ravoli, chicken,with cranberry...mmmm!
        Positive: 83.02%; Negative: 16.98%

I was shocked because no signs indicate cash only.
        Positive: 7.71%; Negative: 92.29%

Waitress was a little slow in service.
        Positive: 16.08%; Negative: 83.92%

did not like at all.
        Positive: 1.85%; Negative: 98.15%
```

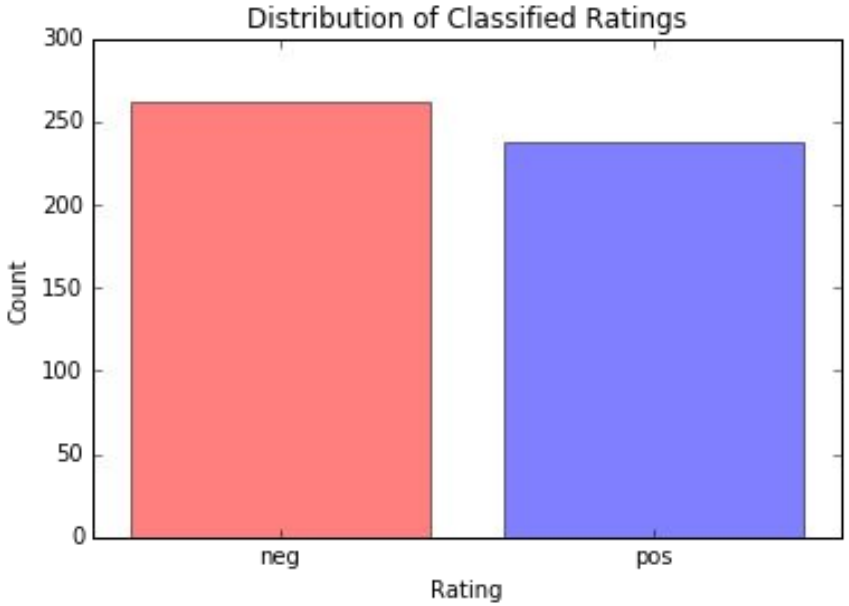
Rating Distributions Naive Bayes



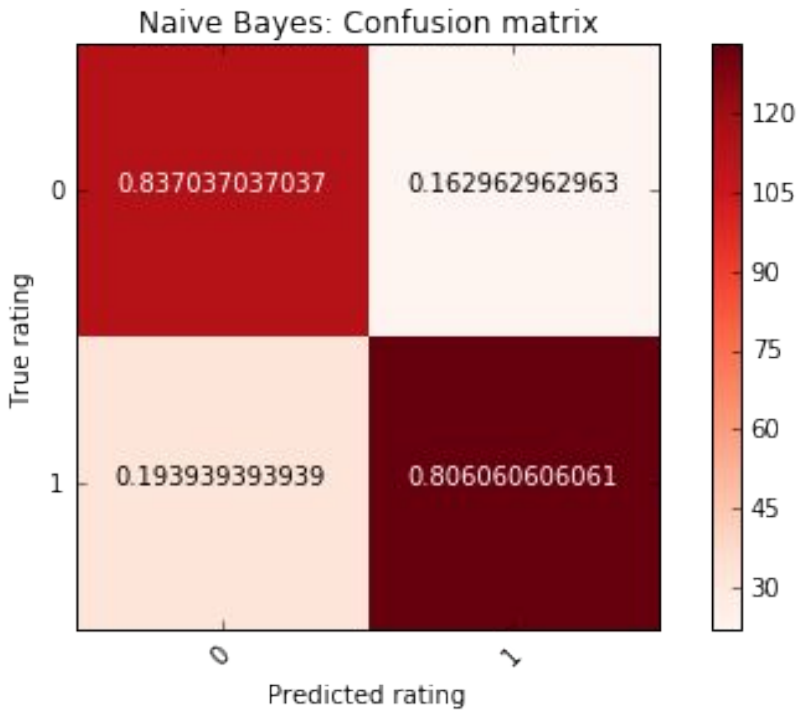


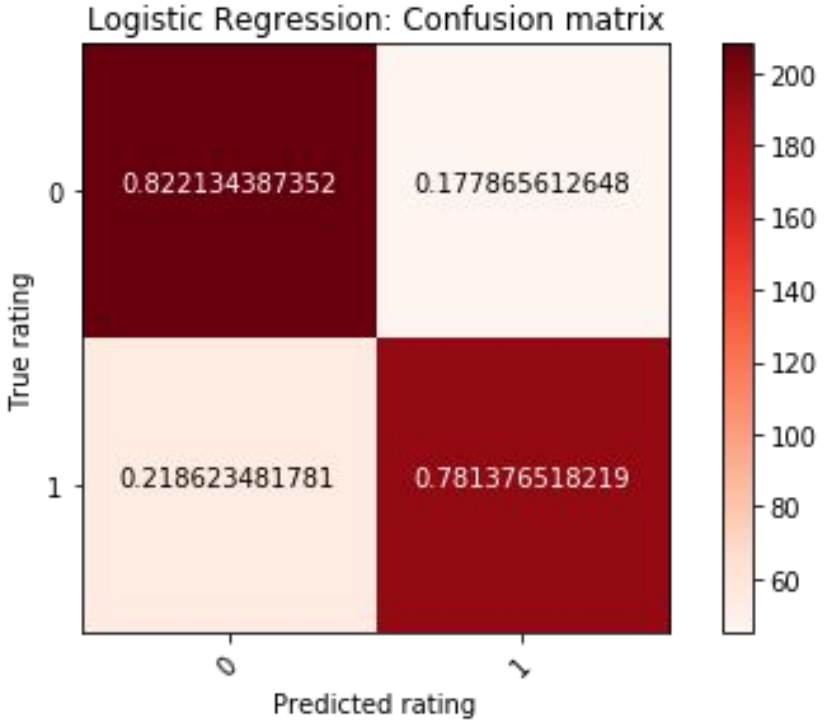
Rating Distributions Logistic Regression





Confusion Matrices



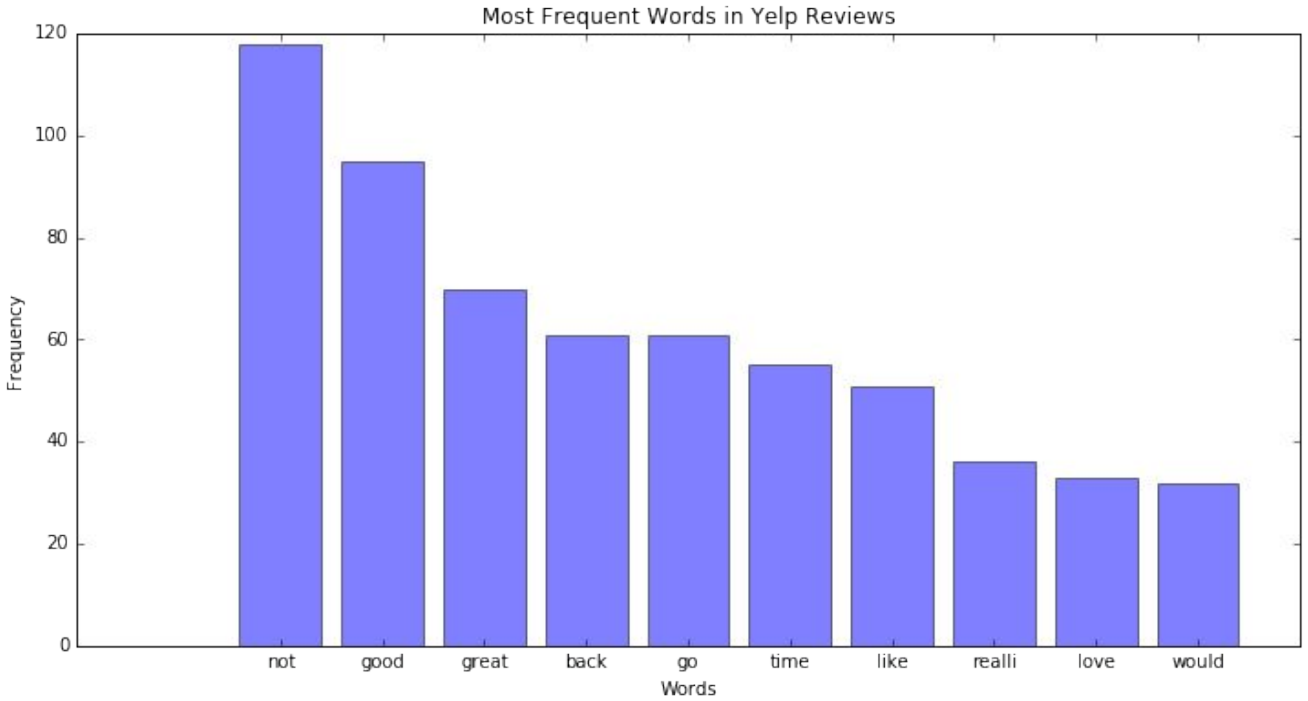


Sentiment Word Clouds

Positive Sentiment Word Cloud

Negative Sentiment Word Cloud

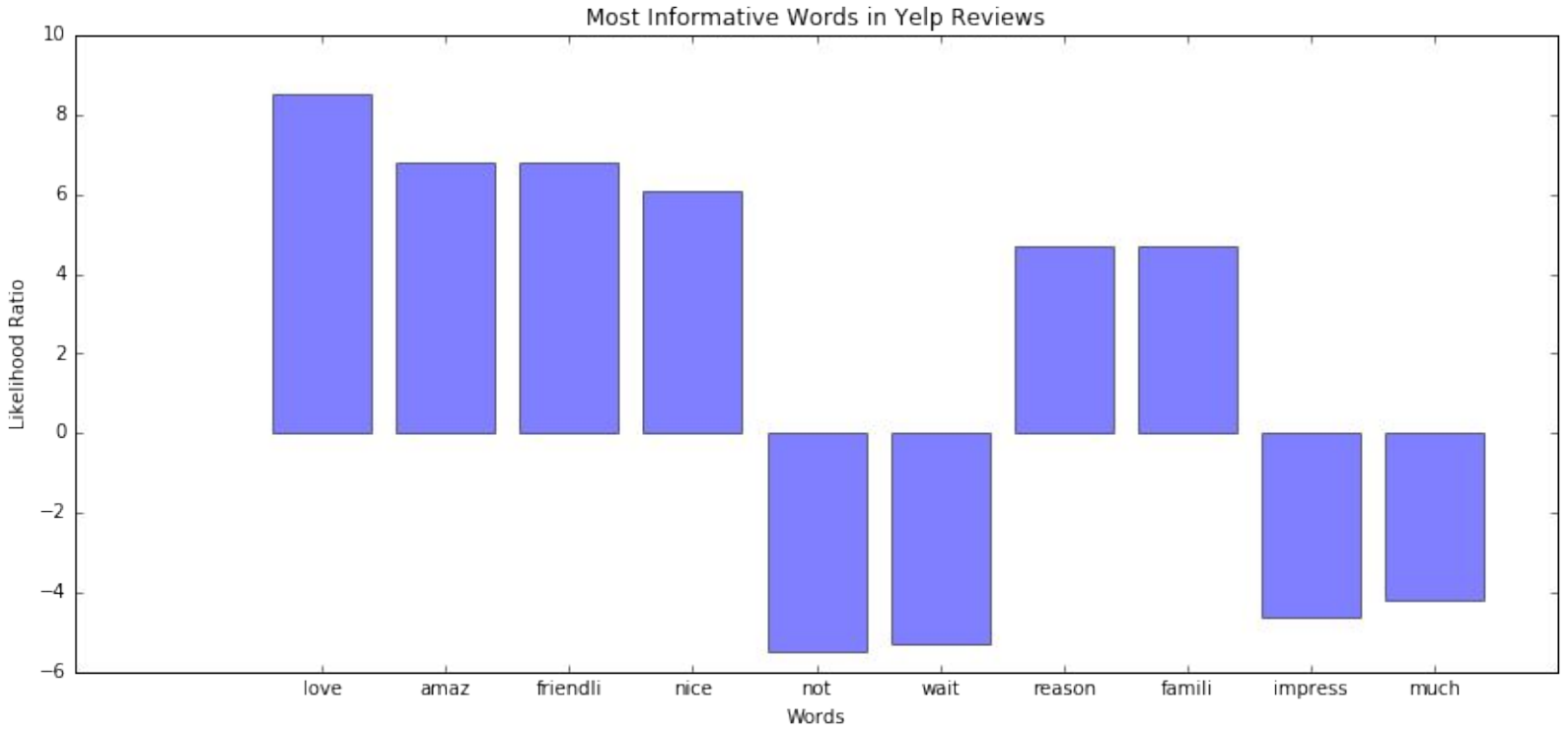
Most Frequent Words



Most Informative Words Data

contains(love) = True	1 : 0	=	8.5 : 1.0
contains(amaz) = True	1 : 0	=	6.8 : 1.0
contains(friendli) = True	1 : 0	=	6.8 : 1.0
contains(nice) = True	1 : 0	=	6.1 : 1.0
contains(not) = True	0 : 1	=	5.5 : 1.0
contains(wait) = True	0 : 1	=	5.3 : 1.0
contains(reason) = True	1 : 0	=	4.7 : 1.0
contains(famili) = True	0 : 1	=	4.7 : 1.0
contains(impress) = True	0 : 1	=	4.6 : 1.0
contains(much) = True	0 : 1	=	4.2 : 1.0

Most Informative Words Visualization



Demo