Natural Language Processing CS 585 Final Presentation

Title:Text Mining and Sentiment Analysis on Yelp Dataset Challenge

ARNAB MUKHOPADHYAY A20353463

GEORGE MATHEW A20352131

SMRITI RAJ A20364719

Abstract

- What is Yelp and what we are trying to achieve?
  - Yelp is a user service online which provides information and latest insights on local businesses
  - Yelp allows registered users located across the world to rate and review business
  - Yelp generates revenue by selling ads and sponsored listings to small businesses.
  - We are trying to solve certain aspects of the yelp data set challenge which contains
  - 4.1M reviews and 947K tips by 1M users for 144K businesses
  - 1.1M business attributes, e.g., hours, parking availability, ambience.
  - Aggregated check-ins over time for each of the 125K businesses
  - We build a classifier to classify the relevant and non-relevant sentences from the yelp review text for the bad reviews and using the relevant sentences we analysed the reason behind the bad reviews for each restaurant.
  - We also performed some experimental analysis on the key factors of the main reasons and plotted some graph for comparative analysis of the key factors.

Introduction

- For our study, we considered only restaurant data.
- We have considered out only those businesses that are categorized as food or restaurants.
- We have filtered out restaurants that have ratings below 3 stars.
- We tried to understand why the ratings are so low. Example: “The service is awful”, “The tables were not clean”
- This feedback helps business owners to improve upon certain areas
- This feedback also enables business owners to get an edge over their competition
- We are left with 48485 restaurants

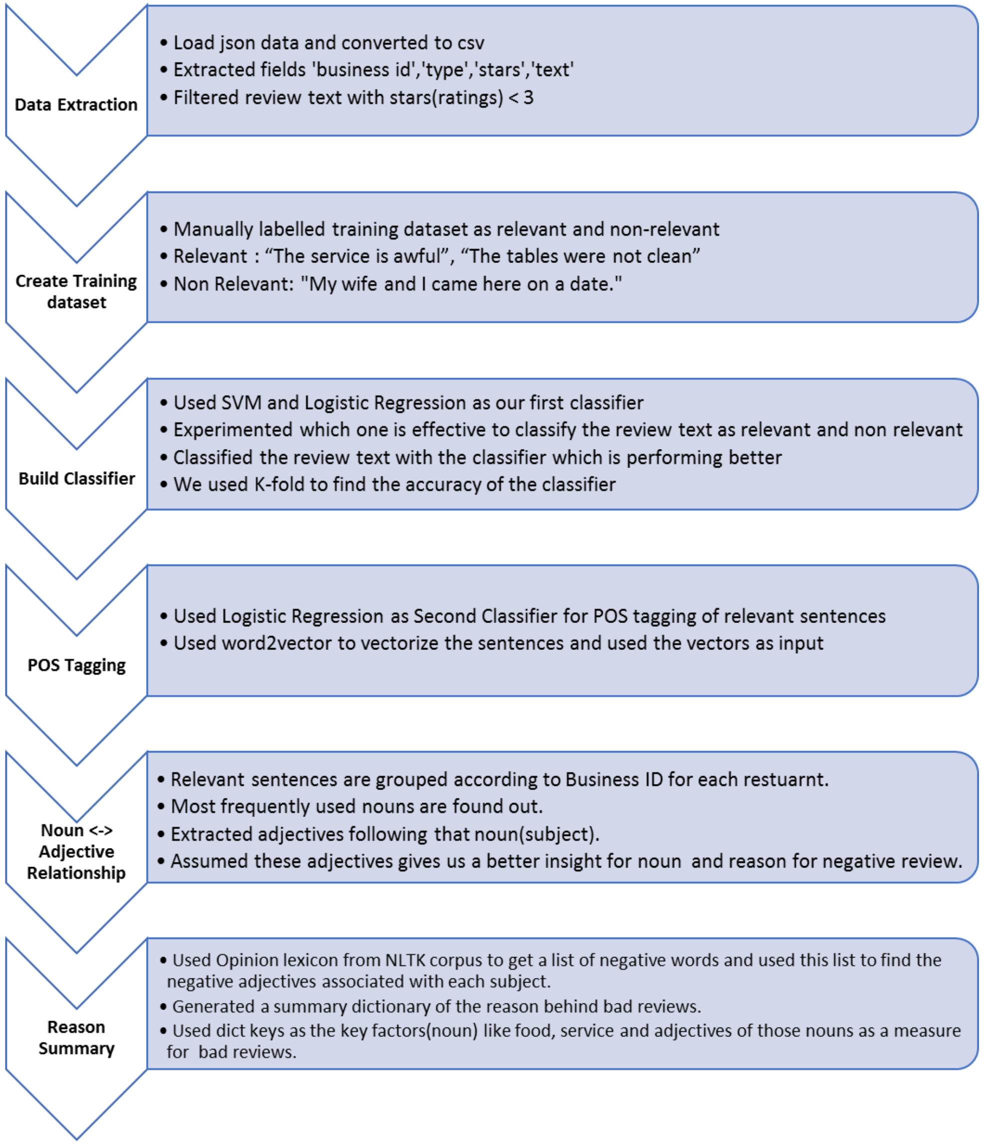
Background/related work

- **Link 1**: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1928601
  - In this paper three main findings are presented about the impact of consumer reviews on the restaurant industry:
  - An increase by one star in Yelp rating leads to a 5-9 percent increase in revenue, this effect noticeable in independent restaurants not a chain
- **Link 2:** https://arxiv.org/abs/1401.0864
  - There are an overwhelming number of reviews for businesses especially restaurants, it is very hard for users to go through all reviews and find the related information
  - This rating is mostly subjective and biased toward users' personality. In this paper, a business rating is predicted based on user-generated reviews texts.
- **Link 3:** http://dl.acm.org/citation.cfm?id=2631784
  - This paper analyses restaurant reviews to extract food words and create food recommendation. We used this paper to start analysing our reviews and to understand the general nature of restaurant reviews.

- **Link4:**https://pdfs.semanticscholar.org/8b2b/ada2218191619611 6f1711d456ea212f2b3b.pdf
    - This paper discusses a food recommendation model based on the yelp data. K-nearest algorithm is used to build the model to recommend food to users. This helped us in understanding yelp data. We decided not to use the k-nearest algorithm and devised our own methods to build a model that understands the reasons behind lower reviews

- **Link 5: http://varianceexplained.org/r/yelp-sentiment/**
    - In this project, they used sentiment analysis methods to classify words as “positive” or “negative”, then to average the values of each word to categorize the entire document. Then they predicted a customer’s rating based on their written opinion. Finally, they compared the predicted ratings with the original ratings.

Approach



Data Extraction

- Load json data and converted to csv
- Extracted fields 'business id','type','stars','text'
- Filtered review text with stars(ratings) < 3

Create Training dataset

- Manually labelled training dataset as relevant and non-relevant
- Relevant : “The service is awful”, “The tables were not clean”
- Non Relevant: "My wife and I came here on a date."

Build Classifier

- Used SVM and Logistic Regression as our first classifier
- Experimented which one is effective to classify the review text as relevant and non relevant
- Classified the review text with the classifier which is performing better
- We used K-fold to find the accuracy of the classifier

POS Tagging

- Used Logistic Regression as Second Classifier for POS tagging of relevant sentences
- Used word2vector to vectorize the sentences and used the vectors as input

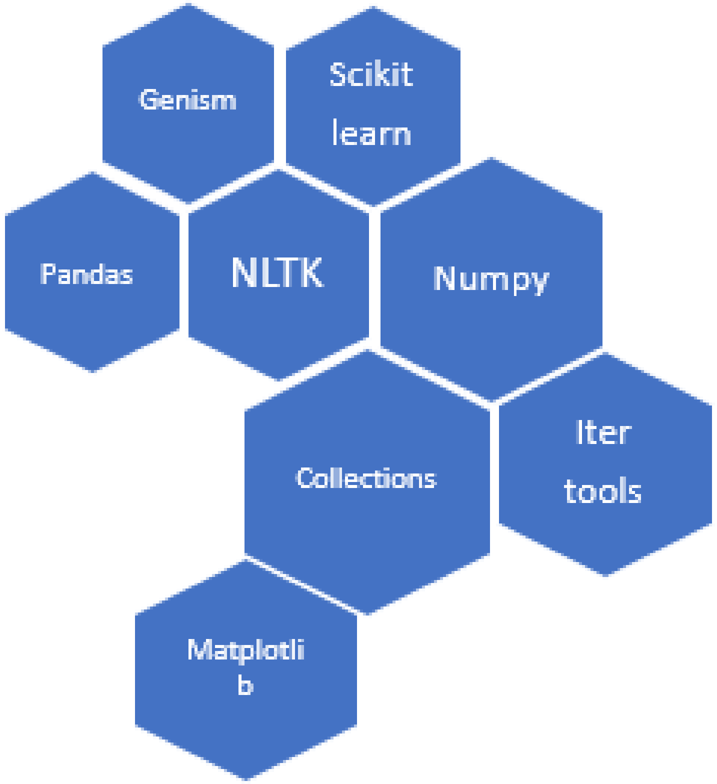
Noun <-> Adjective Relationship

- Relevant sentences are grouped according to Business ID for each restuarnt.
- Most frequently used nouns are found out.
- Extracted adjectives following that noun(subject).
- Assumed these adjectives gives us a better insight for noun and reason for negative review.

Reason Summary

- Used Opinion lexicon from NLTK corpus to get a list of negative words and used this list to find the negative adjectives associated with each subject.
- Generated a summary dictionary of the reason behind bad reviews.
- Used dict keys as the key factors(noun) like food, service and adjectives of those nouns as a measure for bad reviews.

Python Libraries



Genism

Scikit learn

Pandas

NLTK

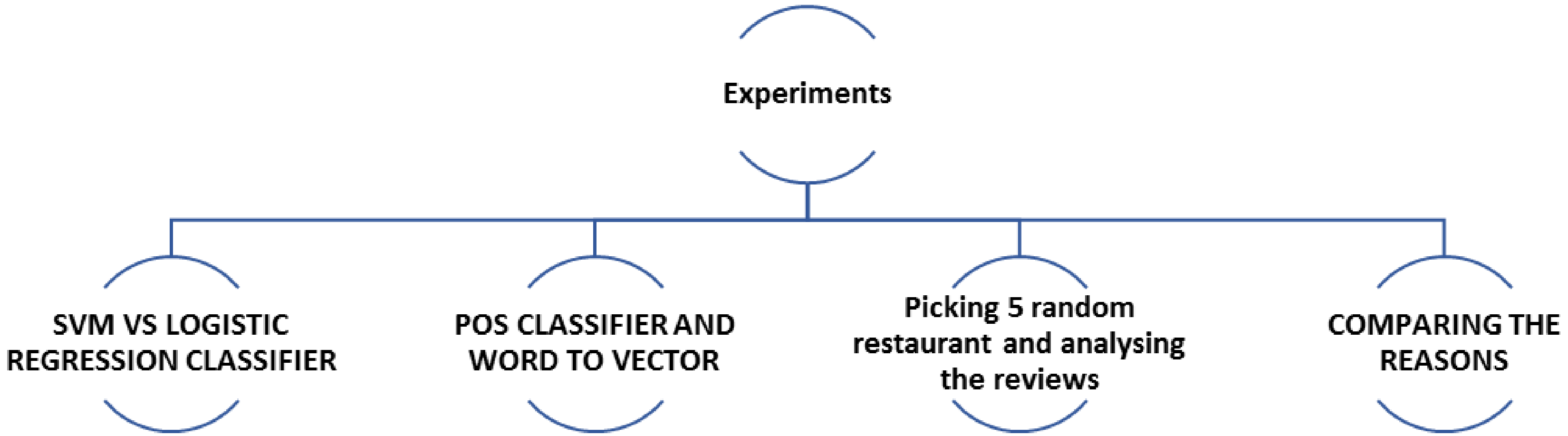
Numpy

Collections

Iter tools

Matplotli b

EXPERIMENT



Experiments

SVM VS LOGISTIC REGRESSION CLASSIFIER

POS CLASSIFIER AND WORD TO VECTOR

Picking 5 random restaurant and analysing the reviews

COMPARING THE REASONS

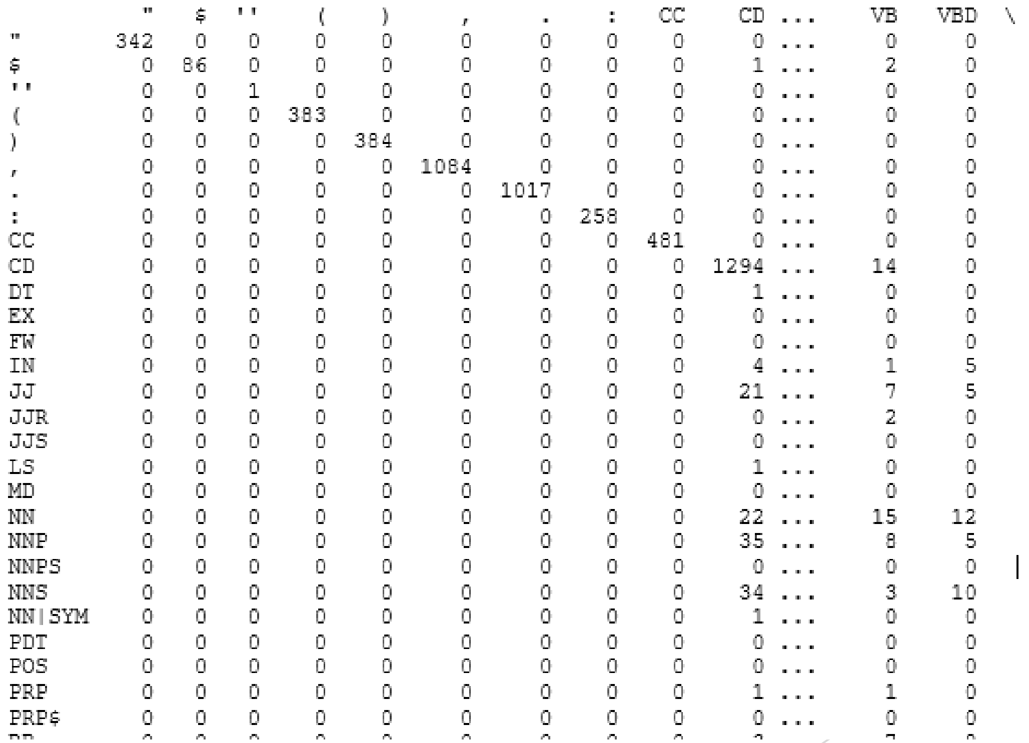
SVM VS LOGISTIC REGRESSION CLASSIFIER:

	Logistic Regression Classifier	One v/s Rest Classifier
TfidfTransformer-enabled	0.513974358974	0.726794871795
TfidfTransformer-disabled	0.697435897436	0.707692307692
Highest Accuracy (One v/s Rest Classifier) min_df-1 max_df-2 binary -true		0.746794871795

POS CLASSIFIER AND WORD TO VECTOR

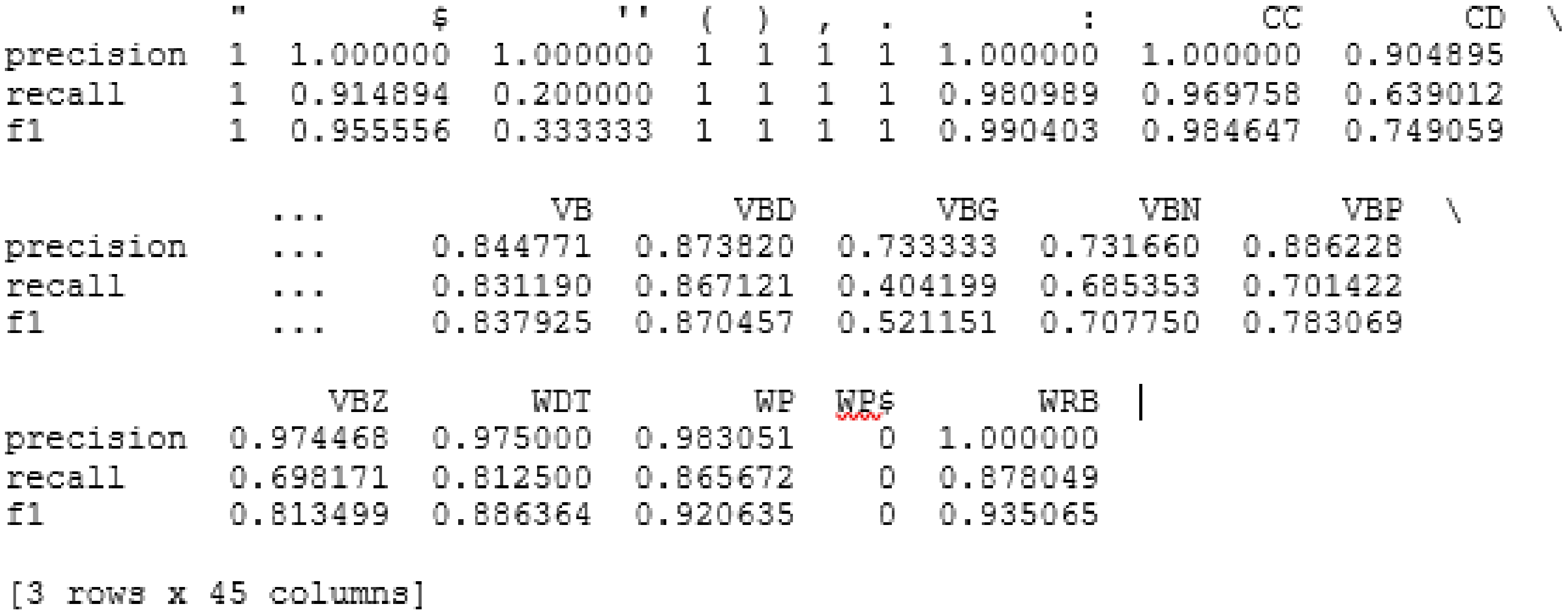
- training data shape: (27867, 18260)
- testing data shape: (28033, 18260)
- Confusion Matrix:

  (45 rows x 45 columns)



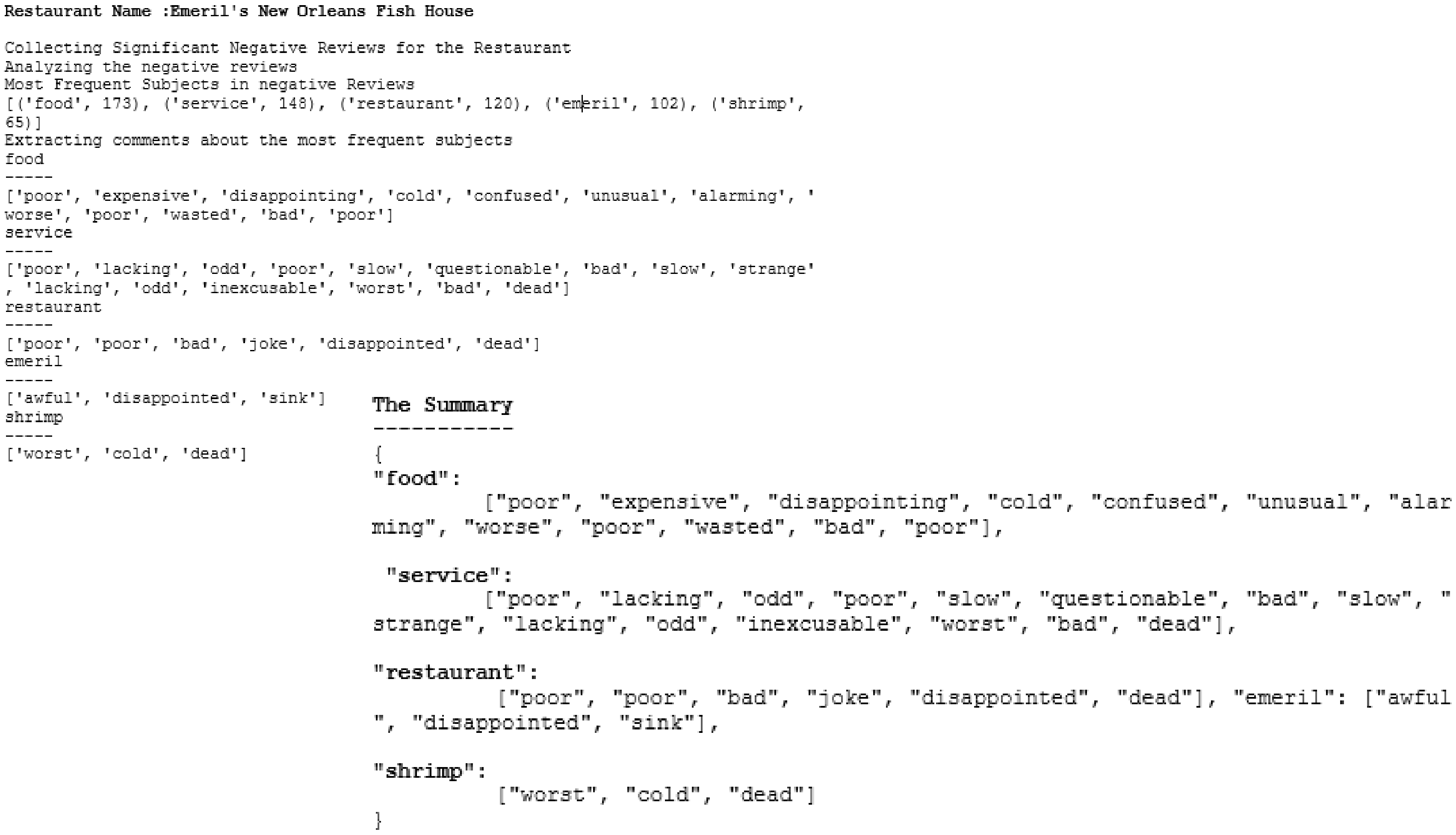
```
             "    $   ''     (     )      ,      .     :    CC     CD ...    VB   VBD  \
"          342    0    0     0     0      0      0     0     0      0 ...     0     0
$            0   86    0     0     0      0      0     0     0      1 ...     2     0
''           0    0    1     0     0      0      0     0     0      0 ...     0     0
(            0    0    0   383     0      0      0     0     0      0 ...     0     0
)            0    0    0     0   384      0      0     0     0      0 ...     0     0
,            0    0    0     0     0   1084      0     0     0      0 ...     0     0
.            0    0    0     0     0      0   1017     0     0      0 ...     0     0
:            0    0    0     0     0      0      0   258     0      0 ...     0     0
CC           0    0    0     0     0      0      0     0   481      0 ...     0     0
CD           0    0    0     0     0      0      0     0     0   1294 ...    14     0
DT           0    0    0     0     0      0      0     0     0      1 ...     0     0
EX           0    0    0     0     0      0      0     0     0      0 ...     0     0
FW           0    0    0     0     0      0      0     0     0      0 ...     0     0
IN           0    0    0     0     0      0      0     0     0      4 ...     1     5
JJ           0    0    0     0     0      0      0     0     0     21 ...     7     5
JJR          0    0    0     0     0      0      0     0     0      0 ...     2     0
JJS          0    0    0     0     0      0      0     0     0      0 ...     0     0
LS           0    0    0     0     0      0      0     0     0      1 ...     0     0
MD           0    0    0     0     0      0      0     0     0      0 ...     0     0
NN           0    0    0     0     0      0      0     0     0     22 ...    15    12
NNP          0    0    0     0     0      0      0     0     0     35 ...     8     5
NNPS         0    0    0     0     0      0      0     0     0      0 ...     0     0
NNS          0    0    0     0     0      0      0     0     0     34 ...     3    10
NN|SYM       0    0    0     0     0      0      0     0     0      1 ...     0     0
PDT          0    0    0     0     0      0      0     0     0      0 ...     0     0
POS          0    0    0     0     0      0      0     0     0      0 ...     0     0
PRP          0    0    0     0     0      0      0     0     0      1 ...     1     0
PRP$         0    0    0     0     0      0      0     0     0      0 ...     0     0
```

- Evaluation Matrix:
- Average F1s: 0.692201



```
           "         $        ''  (  )  ,  .         :        CC        CD  \
precision  1  1.000000  1.000000  1  1  1  1  1.000000  1.000000  0.904895
recall     1  0.914894  0.200000  1  1  1  1  0.980989  0.969758  0.639012
f1         1  0.955556  0.333333  1  1  1  1  0.990403  0.984647  0.749059

             ...        VB       VBD       VBG       VBN       VBP  \
precision    ...  0.844771  0.873820  0.733333  0.731660  0.886228
recall       ...  0.831190  0.867121  0.404199  0.685353  0.701422
f1           ...  0.837925  0.870457  0.521151  0.707750  0.783069

                VBZ       WDT        WP  WP$       WRB
precision  0.974468  0.975000  0.983051    0  1.000000
recall     0.698171  0.812500  0.865672    0  0.878049
f1         0.813499  0.886364  0.920635    0  0.935065

[3 rows x 45 columns]
```

Picking 5 random restaurant and analysing the reviews

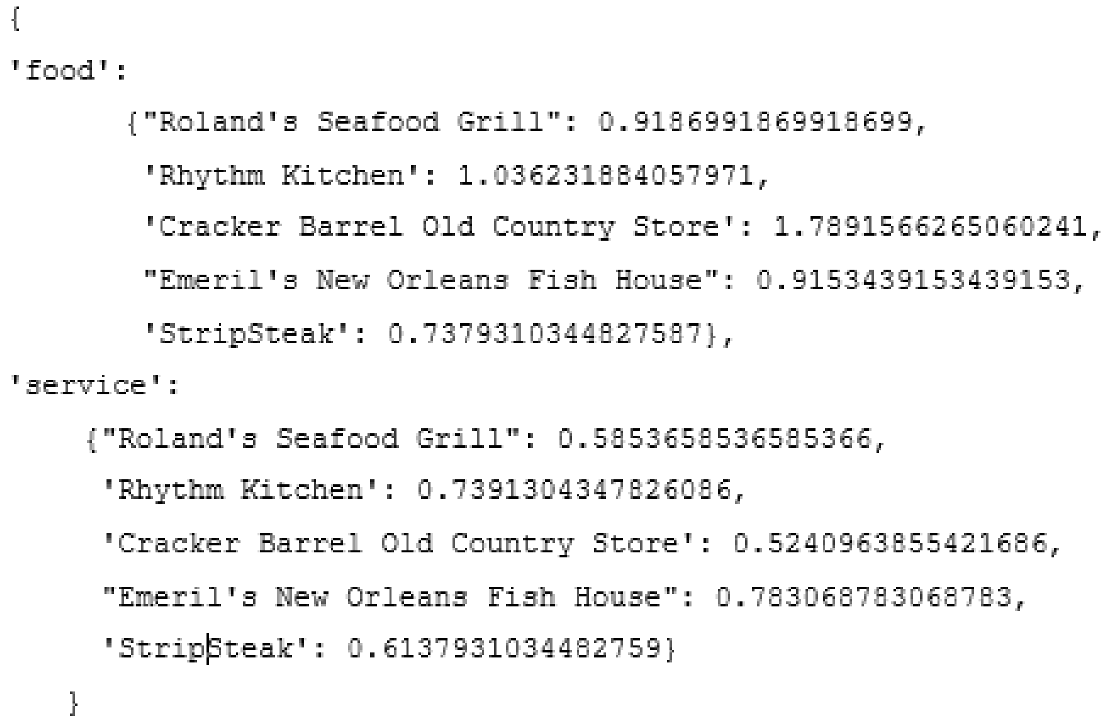


```
Restaurant Name :Emeril's New Orleans Fish House

Collecting Significant Negative Reviews for the Restaurant
Analyzing the negative reviews
Most Frequent Subjects in negative Reviews
[('food', 173), ('service', 148), ('restaurant', 120), ('emeril', 102), ('shrimp',
65)]
Extracting comments about the most frequent subjects
food
-----
['poor', 'expensive', 'disappointing', 'cold', 'confused', 'unusual', 'alarming', '
worse', 'poor', 'wasted', 'bad', 'poor']
service
-----
['poor', 'lacking', 'odd', 'poor', 'slow', 'questionable', 'bad', 'slow', 'strange'
, 'lacking', 'odd', 'inexcusable', 'worst', 'bad', 'dead']
restaurant
-----
['poor', 'poor', 'bad', 'joke', 'disappointed', 'dead']
emeril
-----
['awful', 'disappointed', 'sink']
shrimp
-----
['worst', 'cold', 'dead']
```

```
The Summary
-----------
{
"food":
         ["poor", "expensive", "disappointing", "cold", "confused", "unusual", "alar
ming", "worse", "poor", "wasted", "bad", "poor"],

 "service":
         ["poor", "lacking", "odd", "poor", "slow", "questionable", "bad", "slow", "
strange", "lacking", "odd", "inexcusable", "worst", "bad", "dead"],

"restaurant":
          ["poor", "poor", "bad", "joke", "disappointed", "dead"], "emeril": ["awful
", "disappointed", "sink"],

"shrimp":
          ["worst", "cold", "dead"]
}
```

COMPARING THE REASONS

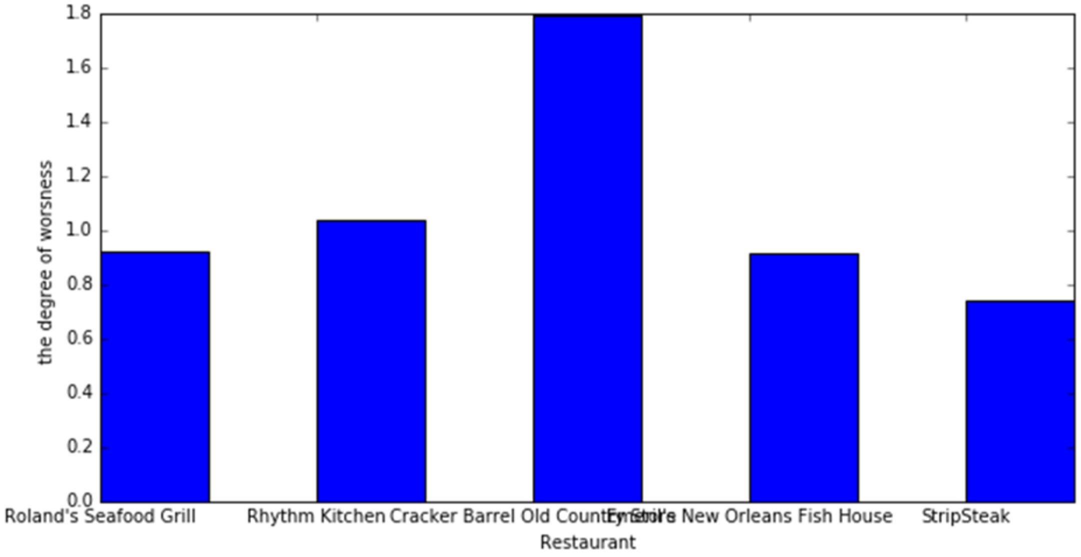
- Assigned degree of worseness for each subject depending on how many times the subject is mentioned with respect to the number of reviews
- Degree of worseness = (Number of times the subject is mentioned) / (Number of negative reviews)



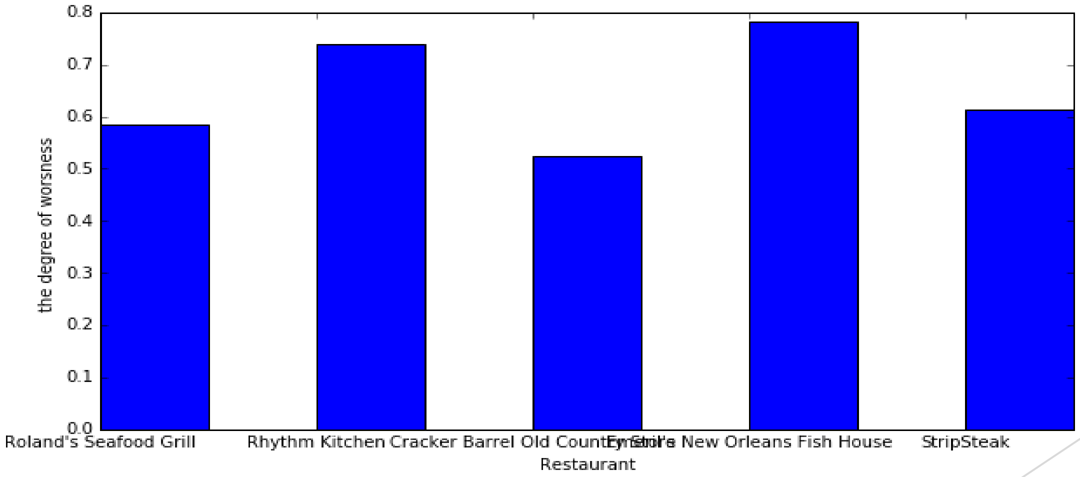
```
{
'food':
       {"Roland's Seafood Grill": 0.9186991869918699,
        'Rhythm Kitchen': 1.036231884057971,
        'Cracker Barrel Old Country Store': 1.7891566265060241,
        "Emeril's New Orleans Fish House": 0.9153439153439153,
        'StripSteak': 0.7379310344827587},
'service':
     {"Roland's Seafood Grill": 0.5853658536585366,
      'Rhythm Kitchen': 0.7391304347826086,
      'Cracker Barrel Old Country Store': 0.5240963855421686,
      "Emeril's New Orleans Fish House": 0.783068783068783,
      'StripSteak': 0.6137931034482759}
    }
```

Plot

- Graph for food



- Graph for service



Conclusion

- We learned how to use POS tags to make sense out of a sentence.
- As far as natural language is concerned most of the times there could be a lot of junk data which may not be relevant to the result. (Example: “My wife and I went on a date here”). This sentence doesn’t help us to understand the reason for bad reviews.
- SVM classifier is performing better than Logistic regression classifier for this dataset.
- When a restaurant rating falls below a certain threshold the restaurant owner can use this application to find out the reason for the decline and neighbouring or competitors can use this information to pull ahead in the business.

References

- https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1928601
- https://arxiv.org/abs/1401.0864
- http://dl.acm.org/citation.cfm?id=2631784
- http://dl.acm.org/citation.cfm?id=2507163
- https://pdfs.semanticscholar.org/8b2b/ada22181916196116f1711d456ea212f2b3b.pdf
- https://pdfs.semanticscholar.org/9c85/836ffaa9dfb3523b793f0d41198d13621b6a.pdf

Thank you…..