Natural Language Processing CS 585 Final Report

Text Mining and Sentiment Analysis on Yelp Dataset Challenge

AUTHOR	**ARNAB MUKHOPADHYAY, GEORGE MATHEW, SMRITI RAJ**
CWID	**A20353463, A20352131, A20364719**
DATE	**04/28/2017**

Table of Contents

1. 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.

2. 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
- Number of Reviews 528134
- Number of Restaurants with bad reviews 42009

3. 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
- restaurant chains have had a declined market share as Yelp penetration increased. This suggests that online consumer reviews have a more powerful effect than more traditional forms of reputation.

- **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.
- They use a combination of three feature generation methods and four machine learning models to find the best prediction. This approach creates a bag of words from the top frequent words in all raw text reviews, or top frequent words/adjectives from results of Part-of-Speech analysis

- **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.
- Our project tries to analyse restaurant reviews and understand the reasons behind a lower review. Though our review analysis was inspired from this paper, our objectives are completely different

- **Link 4:**https://pdfs.semanticscholar.org/8b2b/ada22181916196116f1711d456ea212f2b3b.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. In our project, we are taking the original ratings to filter out the data which has ratings above 3. Then we worked on those reviews with lower rating to classify the sentences and to analyse the reason behind the bad ratings.

4. APPROACH

- **Method:**

- We load the json data and create a CSV of the data relevant to us
- The code extracts only those values we need ['business_id','type','stars','text']
- We then filter those reviews with 1 and 2 stars
- We use a tokenizing function that removes hyperlinks and stop words
- Then we manually labelled the training dataset as relevant and non-relevant.
- We used SVM and Logistic Regression as our first classifier to experiment which one is effective to classify the review text as relevant and non-relevant.
- Finally, we got better accuracy for SVM and used the same to classify all the review text.
- Then we have used Logistic regression as the second classifier for POS tagging of the relevant sentences.
- Then relevant sentences are grouped according to business ID and then most frequently used nouns are found out. We assumed that these nouns are addressed in the review.
- In a sentence, we tried to find out the adjectives following that noun(subject). We assumed that these adjectives give us a better insight about the noun and the reason for a negative review.
- We have used opinion lexicon from NLTK corpus to get a list of negative words and then this list is used to find the negative adjectives associated with each subject.

- **Library:**

- Scikitlearn
- Numpy
- pandas
- itertools
- NLTK
- Genism
- Matplotlib
- Collections

- **Methods used to generate summary output:**

- **Sample Output for one restaurant:**

```
Using the classifier to find the most popular keywords in the entire yelp negative
reviews
Picking 5 random restaurant and analyzing 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"]
}
```

- **Approach:**

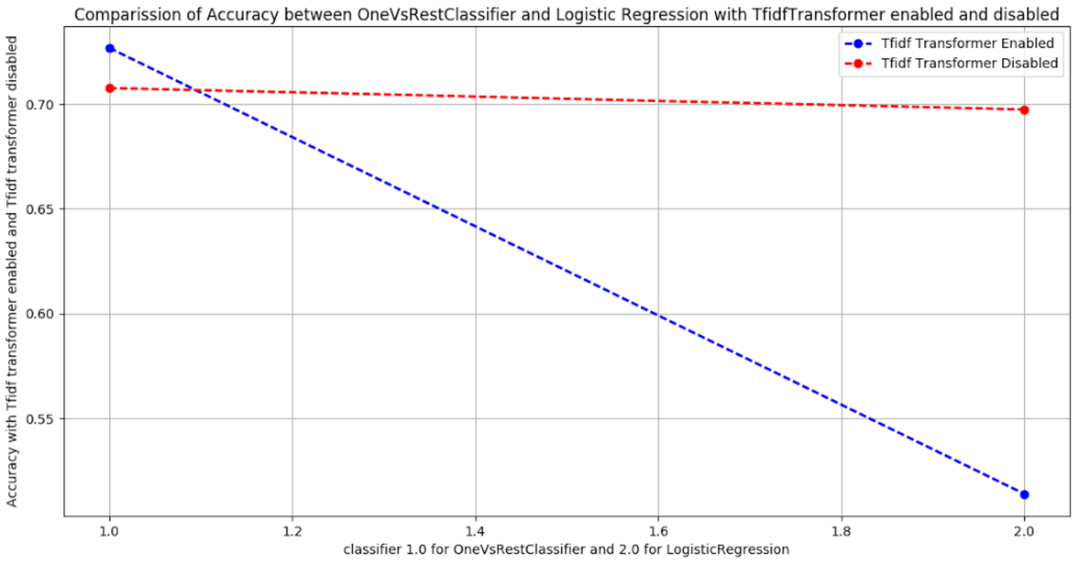
- Using the classifier to find the most popular keywords in the entire yelp negative reviews for a restaurant.
- Collecting Significant Negative Reviews for the Restaurant and analysing the negative reviews.
- Extracting comments(adjectives) about the most frequent subjects(nouns).
- Generate a summary dictionary of the reason behind bad reviews. Used keys as the key factors(noun) like food, service and adjectives of those nouns as a measure for bad reviews.

5. EXPERIMENT

I. SVM VS LOGISTIC REGRESSION CLASSIFIER:

- We built one Logistic Regression and One v/s All classifier and to classify sentences as relevant or non-relevant
- We tried various combinations and chose the classifier with the highest accuracy
- We sifted through several reviews and picked the most common ones or the ones that have most common attributes

- We experimented with TFID enabled and disabled and found the when it is disabled leads to higher accuracy
- We adapted different combinations of min_df, max_df, binary in our classifiers the highest accuracy is 0.746794871795 using a One v/s Rest Classifier with min_df, max_df, binary as 1, 2, True respectively



	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

II. POS CLASSIFIER AND WORD TO VECTOR

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

- Confusion Matrix:

```
          "   $  ''    (    )     ,     .    :   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
RB        0   0   0    0    0     0     0    0    0     3  ...    7     8
RBR       0   0   0    0    0     0     0    0    0     0  ...    0     0
RBS       0   0   0    0    0     0     0    0    0     0  ...    0     0
RP        0   0   0    0    0     0     0    0    0     0  ...    0     0
SYM       0   0   0    0    0     0     0    0    0     0  ...    0     1
TO        0   0   0    0    0     0     0    0    0     0  ...    0     0
UH        0   0   0    0    0     0     0    0    0     0  ...    0     0
VB        0   0   0    0    0     0     0    0    0     3  ...  517     8
VBD       0   0   0    0    0     0     0    0    0     3  ...    7  1018
VBG       0   0   0    0    0     0     0    0    0     2  ...    7    16
VBN       0   0   0    0    0     0     0    0    0     1  ...    2    36
VBP       0   0   0    0    0     0     0    0    0     2  ...   19     5
VBZ       0   0   0    0    0     0     0    0    0     1  ...    0    36
WDT       0   0   0    0    0     0     0    0    0     0  ...    0     0
WP        0   0   0    0    0     0     0    0    0     0  ...    0     0
WP$       0   0   0    0    0     0     0    0    0     0  ...    0     0

WRB       0   0   0    0    0     0     0    0    0     0  ...    0     0



        VBG  VBN  VBP  VBZ  WDT  WP  WP$  WRB
"         0    0    0    0    0   0    0    0
$         0    0    0    0    0   0    0    0
''        1    0    0    0    0   0    0    0
(         0    0    0    0    0   0    0    0
)         0    0    0    0    0   0    0    0
```

```
,          0    0    0    0    0   0    0    0
.          0    0    0    0    0   0    0    0
:          1    0    0    0    0   0    0    0
CC         0    0    0    0    0   0    0    0
CD         2    3    0    0    0   0    0    0
DT         0    0    0    0    0   0    0    0
EX         0    0    0    0    0   0    0    0
FW         0    0    0    0    0   0    0    0
IN         2    0    0    1    1   0    0    0
JJ         6   45    0    1    0   0    0    0
JJR        0    0    0    0    0   0    0    0
JJS        0    0    0    0    0   0    0    0
LS         0    0    0    0    0   0    0    0
MD         0    0    0    1    0   0    0    0
NN         6    9    0    0    0   0    0    0
NNP        4    4    0    0    0   0    0    0
NNPS       1    0    0    0    0   0    0    0
NNS        5    3    0    1    0   0    0    0
NN|SYM     0    0    0    0    0   0    0    0
PDT        0    0    0    0    0   0    0    0
POS        0    0    0    0    0   0    0    0
PRP        1    0    0    0    0   0    0    0
PRP$       0    0    0    0    0   0    0    0
RB         8   13    1    0    0   0    0    0
RBR        0    0    0    0    0   0    0    0
RBS        0    0    0    0    0   0    0    0
RP         0    0    0    0    0   0    0    0
SYM        0    0    0    0    0   0    0    0
TO         0    0    0    0    0   0    0    0
UH         0    0    0    0    0   0    0    0
VB         2    4   16    0    0   0    0    0
VBD        6   25    1    2    0   0    0    0
VBG      154   33    0    0    0   0    0    0
VBN        8  379    0    0    0   0    0    0
VBP        0    0  148    0    0   0    0    0
VBZ        1    0    0  229    0   0    0    0
WDT        0    0    0    0   78   0    0    0
WP         2    0    0    0    1  58    0    0
WP$        0    0    0    0    0   0    0    0
WRB        0    0    1    0    0   1    0   36

[45 rows x 45 columns]
```

- Evaluation Matrix:

```
           "         $         ''  (  )  ,  .         :        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]
```

- Average F1s: 0.692201

➢ **Analysis:**

- The word to vector model is used to vectorise all the terms for use of computation and for better accuracy.
- POS tagging is used to determine the reason for bad reviews. From the relevant sentences an inference was drawn between the verbs and the nouns. Such as:Bad food,horrible services.

III. 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 frequesnt 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", "ala
rming", "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"]}

Restaurant Name :Cracker Barrel Old Country Store
Collecting Significant Negative Reviews for the Restaurant
Analyzing the negative reviews
Most Frequent Subjects in negative Reviews
[('food', 297), ('time', 103), ('barrel', 96), ('service', 87), ('cracker', 78)]
Extracting comments about the most frequesnt subjects
food
-----
['sick', 'cold', 'worse', 'rude', 'cold', 'unfortunately', 'bad', 'boring', 'bad',
'bad', 'cold', 'poor', 'bland', 'poor', 'bland', 'cold', 'mediocre', 'mediocre', 'c
old', 'unhappy', 'cold', 'disappointing', 'bad', 'weird', 'cold', 'mediocre', 'blan
d', 'disappointment', 'worse', 'awful', 'bland', 'frozen', 'cold', 'cold', 'hard',
'cold', 'hard', 'inconsistent', 'worse']
time
-----
```

```
['creepy', 'cold', 'wrong', 'cold', 'downhill', 'ridiculous', 'disgusted', 'horribl
e', 'bad']
barrel
-----
['bust', 'bad', 'bad', 'worse', 'horrible', 'worst', 'disgusting', 'bad', 'gross',
'hard']
service
-----
['sick', 'slow', 'bad', 'horrible', 'horrible', 'slow', 'boring', 'poor', 'bad', 's
low', 'bad', 'cold', 'bad', 'bad', 'ridiculously', 'cold', 'poor']
cracker
-----
['horrible']
```

The Summary

```
-----------
{"food": ["sick", "cold", "worse", "rude", "cold", "unfortunately", "bad", "boring"
, "bad", "bad", "cold", "poor", "bland", "poor", "bland", "cold", "mediocre", "medi
ocre", "cold", "unhappy", "cold", "disappointing", "bad", "weird", "cold", "mediocr
e", "bland", "disappointment", "worse", "awful", "bland", "frozen", "cold", "cold",
"hard", "cold", "hard", "inconsistent", "worse"], "service": ["sick", "slow", "bad"
, "horrible", "horrible", "slow", "boring", "poor", "bad", "slow", "bad", "cold", "
bad", "bad", "ridiculously", "cold", "poor"], "time": ["creepy", "cold", "wrong", "
cold", "downhill", "ridiculous", "disgusted", "horrible", "bad"], "cracker": ["horr
ible"], "barrel": ["bust", "bad", "bad", "worse", "horrible", "worst", "disgusting"
, "bad", "gross", "hard"]}
```

Restaurant Name: Rhythm Kitchen

```
Collecting Significant Negative Reviews for the Restaurant
Analyzing the negative reviews
Most Frequent Subjects in negative Reviews
[('food', 143), ('service', 102), ('place', 94), ('restaurant', 63), ('table', 61)]
Extracting comments about the most frequesnt subjects
food
-----
['dirty', 'slow', 'bad', 'disappointing', 'bad', 'fried', 'ridiculous', 'worst', 'w
rong', 'disappointing', 'annoyed', 'horrible', 'cold', 'expensive', 'disappointed']
service
-----
['bad', 'slow', 'overrated', 'slow', 'horrible', 'slow', 'worst', 'wrong', 'lacking
', 'cold', 'bad', 'bad', 'cold', 'poor', 'joke', 'poor', 'reprehensible', 'worst']
place
-----
['falls', 'dead', 'bad', 'frozen', 'terrible', 'terrible', 'foul', 'rollercoaster',
'fail', 'joke']
restaurant
-----
['horrible', 'disappointed']
table
-----
['freaking', 'wrong']
```

The Summary

```
-----------
{"food": ["dirty", "slow", "bad", "disappointing", "bad", "fried", "ridiculous", "w
orst", "wrong", "disappointing", "annoyed", "horrible", "cold", "expensive", "disap
pointed"], "service": ["bad", "slow", "overrated", "slow", "horrible", "slow", "wor
st", "wrong", "lacking", "cold", "bad", "bad", "cold", "poor", "joke", "poor", "rep
rehensible", "worst"], "place": ["falls", "dead", "bad", "frozen", "terrible", "ter
rible", "foul", "rollercoaster", "fail", "joke"], "restaurant": ["horrible", "disap
pointed"], "table": ["freaking", "wrong"]}
```

Restaurant Name: StripSteak

```
Collecting Significant Negative Reviews for the Restaurant
Analyzing the negative reviews
Most Frequent Subjects in negative Reviews
[('steak', 182), ('food', 107), ('service', 89), ('fries', 86), ('restaurant', 81)]
Extracting comments about the most frequesnt subjects
steak
```

```
-----
['bad', 'bloody', 'fatty', 'puny', 'difficult', 'bad', 'awful', 'poor', 'greasy', '
disappointing', 'poor']
food
-----
['terrible', 'weak', 'horrible', 'disappointingly', 'lacked', 'disappointing', 'awf
ul', 'hard']
service
-----
['slow', 'slow', 'slow', 'terrible', 'rude', 'horrible', 'bad', 'poor', 'poor', 'po
or', 'slow', 'slow']
fries
-----
['greasy', 'cold', 'cold', 'cold', 'cold']
restaurant
-----
['loud', 'crap', 'pricey', 'misunderstood', 'raving']
```

The Summary

```
-----------
{"food": ["terrible", "weak", "horrible", "disappointingly", "lacked", "disappointi
ng", "awful", "hard"], "service": ["slow", "slow", "slow", "terrible", "rude", "hor
rible", "bad", "poor", "poor", "poor", "slow", "slow"], "fries": ["greasy", "cold",
"cold", "cold", "cold"], "restaurant": ["loud", "crap", "pricey", "misunderstood",
"raving"], "steak": ["bad", "bloody", "fatty", "puny", "difficult", "bad", "awful",
"poor", "greasy", "disappointing", "poor"]}
```

Restaurant Name: Roland’s Seafood Grill

```
Collecting Significant Negative Reviews for the Restaurant
Analyzing the negative reviews
Most Frequent Subjects in negative Reviews
[('food', 113), ('lobster', 98), ('service', 72), ('place', 57), ('time', 54)]
Extracting comments about the most frequesnt subjects
food
-----
['slow', 'horrible', 'wrong', 'bad', 'bad', 'mediocre', 'disappointing', 'awful', '
cold', 'cold', 'poor', 'disgusting', 'cold', 'poorly']
lobster
-----
['odd', 'cold', 'dripping', 'doubt']
service
-----
['terrible', 'slow', 'worst', 'cold', 'slow', 'slow', 'terrible', 'hard', 'doubt',
'horrible', 'bad', 'horrible', 'poor', 'terrible']
place
-----
['terrible', 'terrible', 'smelly', 'overrated', 'racist']
time
-----
['cold']
```

The Summary

```
-----------
{"food": ["slow", "horrible", "wrong", "bad", "bad", "mediocre", "disappointing", "
awful", "cold", "cold", "poor", "disgusting", "cold", "poorly"], "service": ["terri
ble", "slow", "worst", "cold", "slow", "slow", "terrible", "hard", "doubt", "horrib
le", "bad", "horrible", "poor", "terrible"], "place": ["terrible", "terrible", "sme
lly", "overrated", "racist"], "time": ["cold"], "lobster": ["odd", "cold", "drippin
g", "doubt"]}
```

- **Analysis:**
- We pick 5 random restaurants and our algorithm has been applied.
- We were able to successfully identify the subjects addressed in the negative reviews and the comments made regarding the subjects.
- Using these data, we were able to generate a summary for each restaurant.

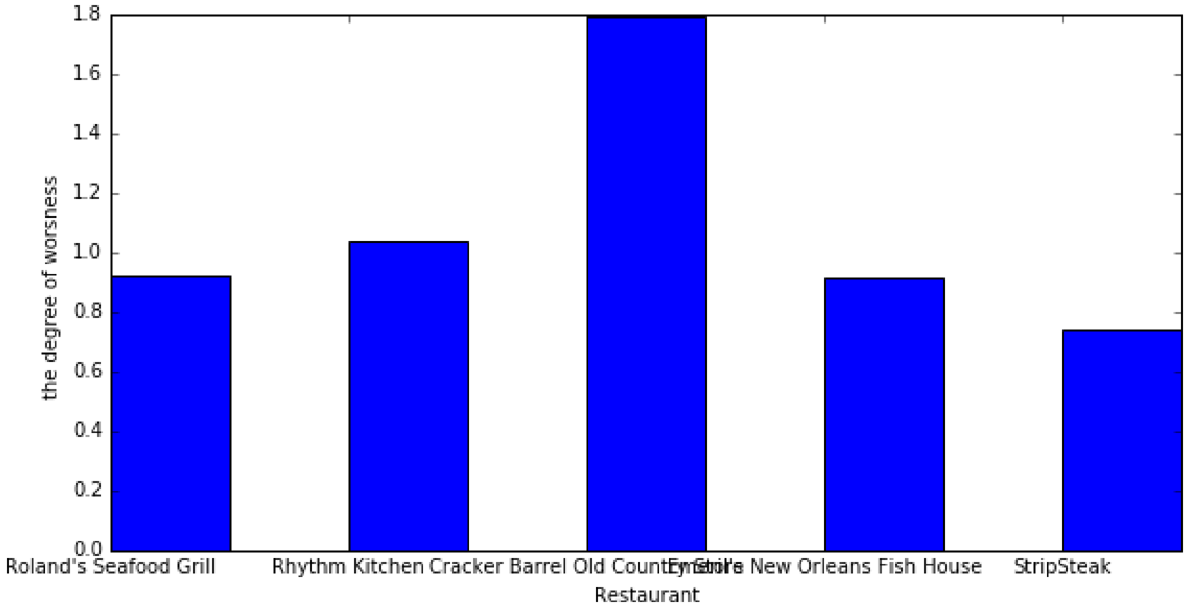
IV. COMPARING THE REASONS

- We will try to assign 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)

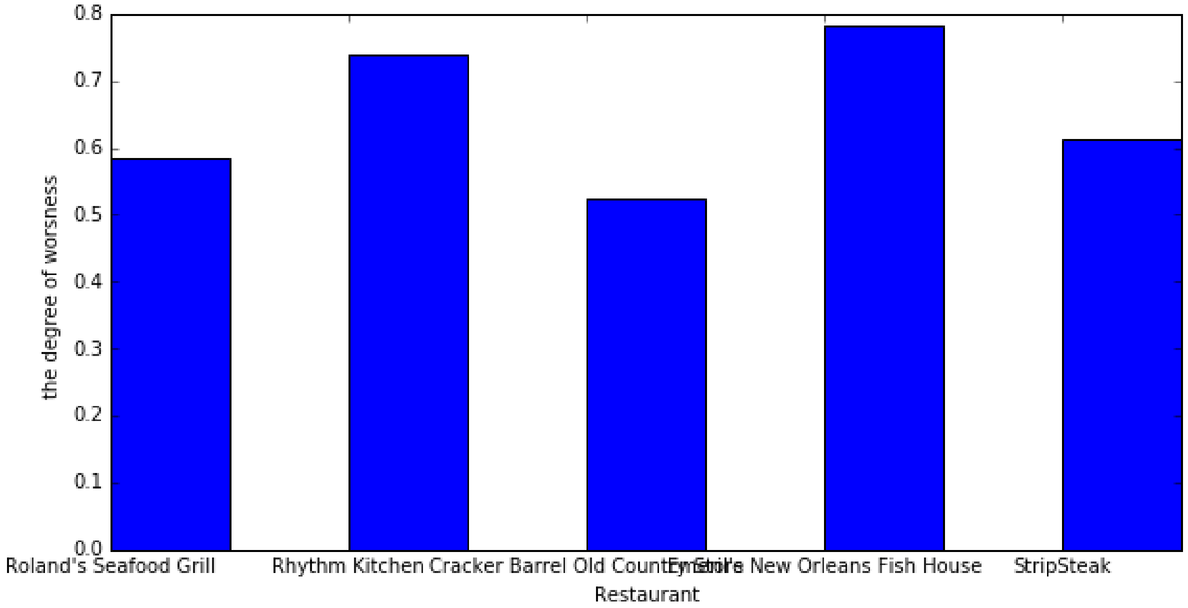
```
{
'food':
       {"Roland's Seafood Grill": 0.9186991869918699,
        'Rhythm Kitchen': 1.036231884057971,
        'Cracker Barrel Old Country Store': 1.7891566265060241,
        "Emeril's New Orleans Fish House": 0.9153439153439153,
        'StripSteak': 0.7379310344827587},
'service':
     {"Roland's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



- **Analysis:**

- We have identified common subjects among the five randomly chosen restaurants to compare the degree of worseness for the subjects.
- Here the X labels refer to the name of the restaurants and the Y labels refer to the degree of worseness.

- Cracker Barrel Old Country Store has the worst review for food among the five restaurants because it has the highest degree of worseness for food.
- Emeril's New Orleans Fish House has the worst review for service among the five restaurants because it has the highest degree of worseness for service.

6. 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.

7. 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