Improving Yelp predictions using sentiment analysis

General Assembly Data Science
Daniel Chung
October 6th, 2014

Agenda

- Introduction
- Analytical data set
- Python libraries used
- KNN results
- Logistic regression results
- Next steps

yelp

Find tacos, cheap dinner, Max's | Near Washington, District of Columbia

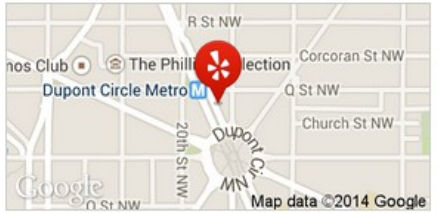
Home | About Me | Write a Review | Find Friends | Messages | Talk | Events

Willie T's Lobster Shack

42 reviews | Details

Write a Review | Add Photo | Share

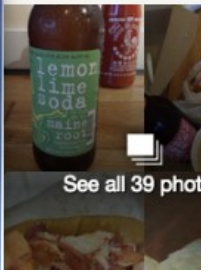
$$ · Seafood | Edit



1511 Connecticut Ave NW
Washington, DC 20009
at N Q St in Dupont Circle
Edit
Get Directions
Dupont Circle Metro More info
(202) 332-3690
Message the business
willietslobstershack.com

"Maine or Connecticut style, gr
Luke's!" in 10 reviews

"I had the **clam chowder**, and
from Boston a year ago." in 7 re



Mokes A.
Washington, DC
4 friends
29 reviews

Share review
Compliment
Send message
Follow Mokes A.

10/5/2014

Yum yum yum. Loved this place when I went a few weeks ago. I'm looking forward to going again.

They have a daily special which at about $22 gets you a lobster roll, pickle, drink, chips, and a moon pie. It was superb!

The people were super nice, the music was jammin. Customers were happy. Staff was happy.

Happy all around. Don't miss it.

Monday Special - Lobster Roll, Maine soda, Cape Cod chils, pickle, and whoopie pie - $22 for bliss!

Was this review ...?

Hours

Mon	11:00 am - 9:00 pm
Tue	11:00 am - 9:00 pm
Wed	11:00 am - 9:00 pm
Thu	11:00 am - 9:00 pm
Fri	11:00 am - 10:00 pm
Sat	11:00 am - 10:00 pm
Sun	11:00 am - 9:00 pm Open now

Edit business info

More business info

Takes Reservations **No**
Delivery **No**
Take-out **Yes**
Accepts Credit Cards **Yes**
Good For **Lunch**
Parking **Street**
Bike Parking **Yes**
Good for Kids **Yes**
Good for Groups **Yes**
Attire **Casual**
Ambience **Casual**
Noise Level **Very Loud**
Alcohol **No**
Outdoor Seating **No**
Wi-Fi **No**
Has TV **No**
Waiter Service **No**
Caters **No**

Analytical data set

- Yelp academic data set
- Limited to Las Vegas restaurants

- 2,583 businesses

Python libraries

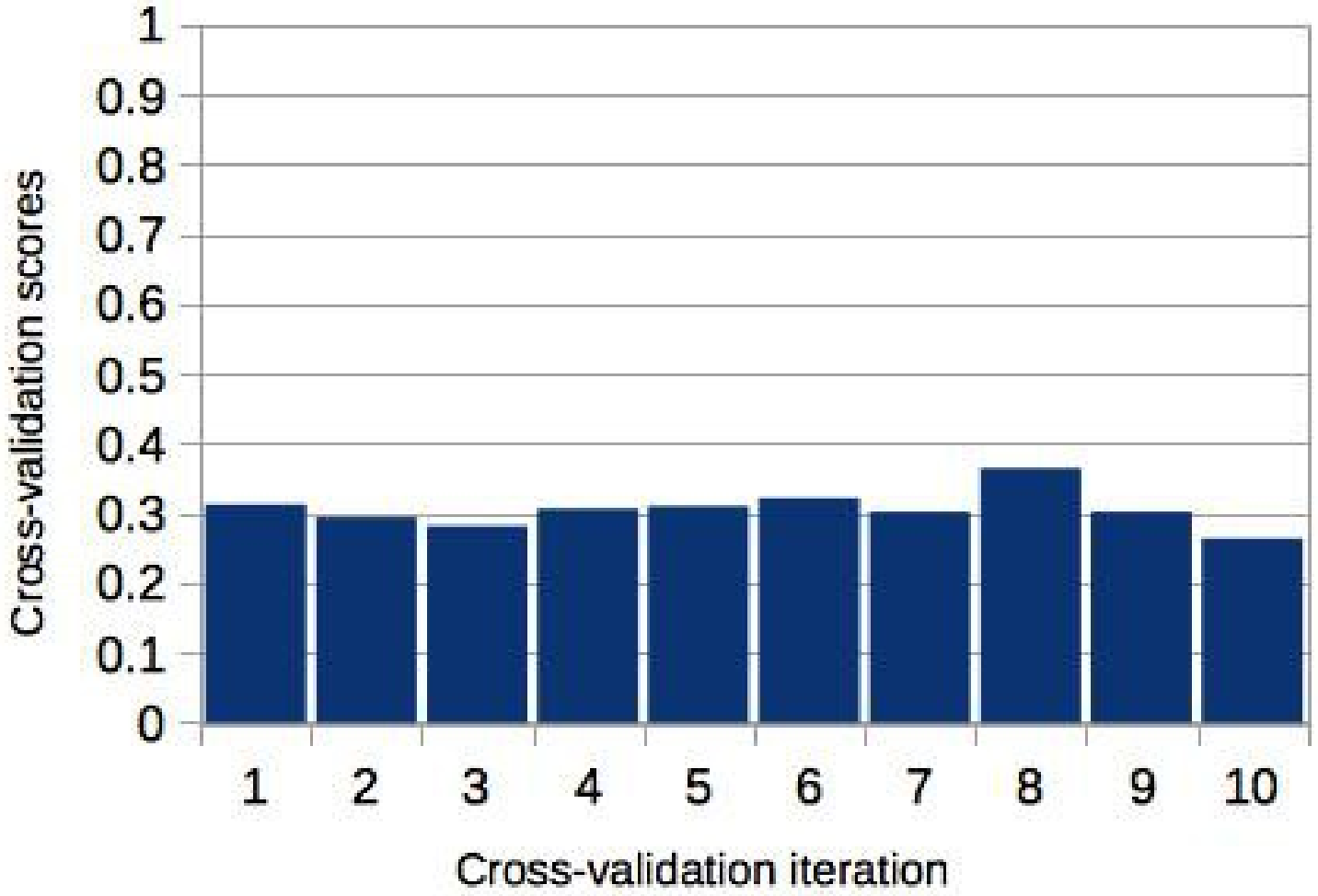
```
import json
import pickle
import pandas as pd
import statsmodel.formula.api as smf
from sklearn import neighbors
from sklearn.cross_validation import cross_val_score
```

AFINN

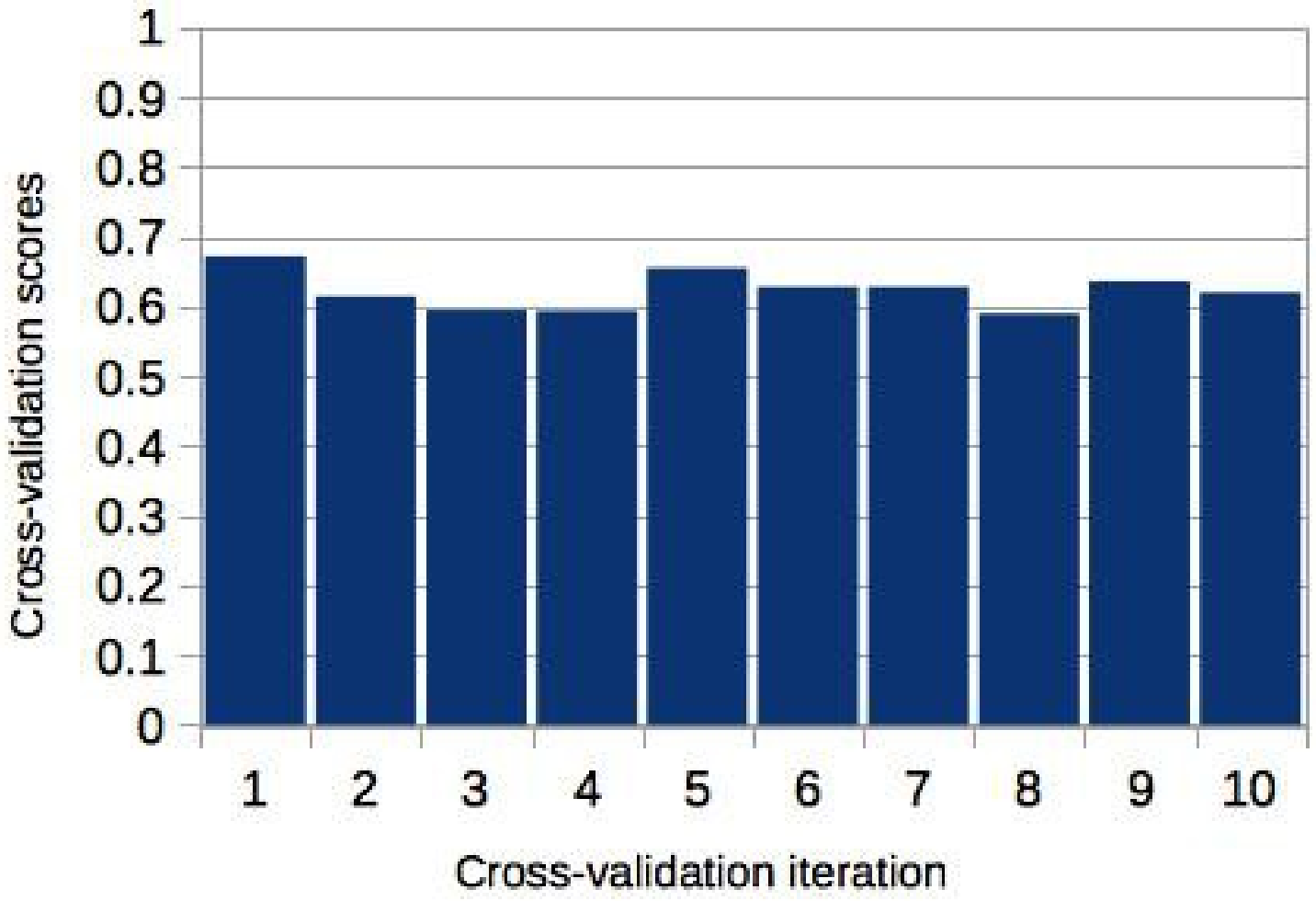
- 2,477 words with scores
- Scores range from -5 to +5

```
fire       -2
fired      -2
firing     -2
fit         1
fitness     1
flagship            2
flees      -1
flop       -2
flops      -2
flu        -2
flustered           -2
focused     2
fond        2
fondness            2
fool       -2
foolish    -2
fools      -2
forced     -1
foreclosure         -2
foreclosures        -2
forget     -1
forgetful           -2
forgive     1
forgiving           1
forgotten           -1
fortunate           2
frantic    -1
fraud      -4
frauds     -4
```

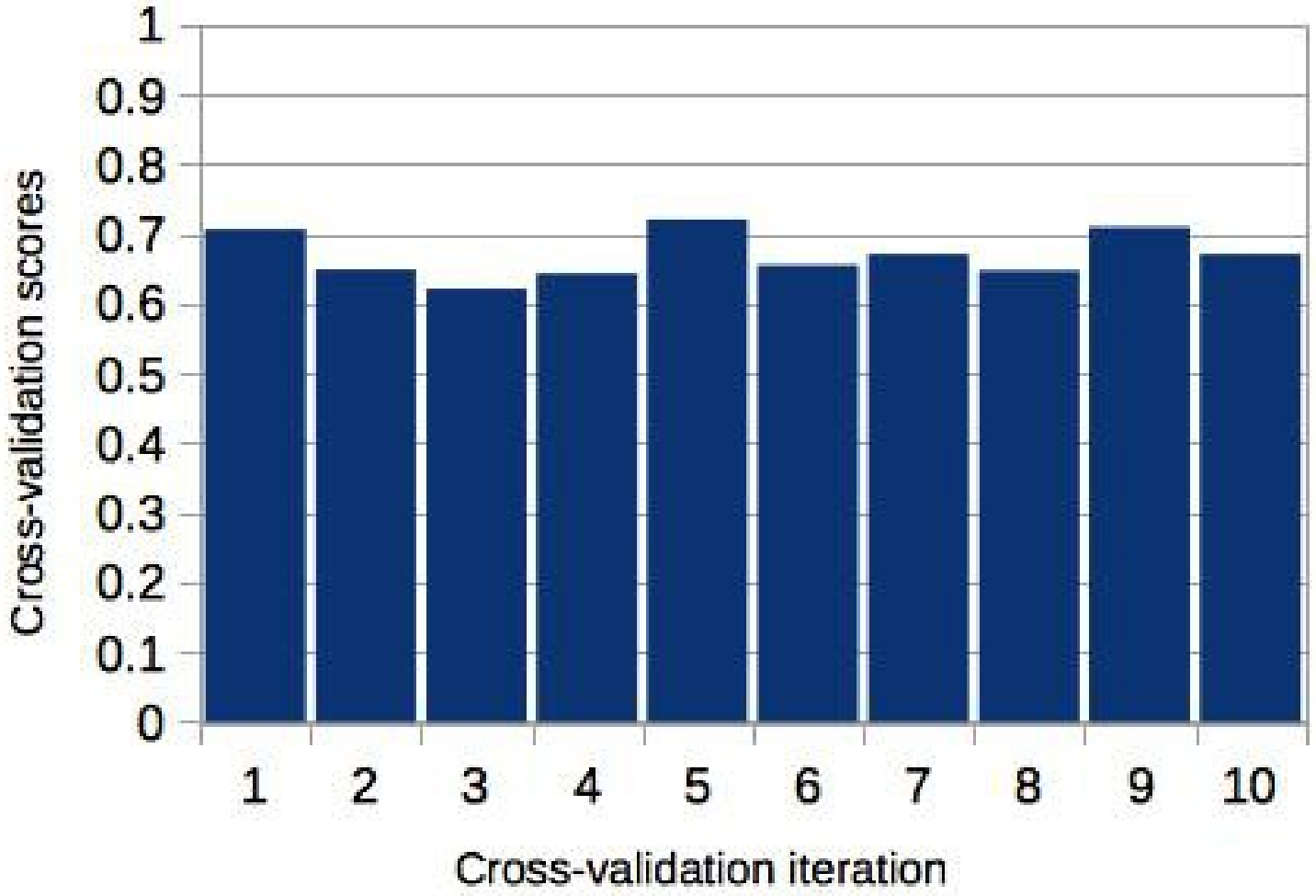
KNN with stars label



KNN with high rating label



KNN with high rating label and text score



Logistic regression

```
Optimization terminated successfully.
         Current function value: 0.528846
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:             high_rating   No. Observations:                 2583
Model:                          Logit   Df Residuals:                     2575
Method:                           MLE   Df Model:                            7
Date:                Fri, 03 Oct 2014   Pseudo R-squ.:                  0.2037
Time:                        22:22:47   Log-Likelihood:                -1366.0
converged:                       True   LL-Null:                       -1715.5
                                        LLR p-value:                1.139e-146
==============================================================================
                    coef    std err          z      P>|z|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     -2.7473      0.356     -7.715      0.000        -3.445    -2.049
romantic[T.True]    -0.0560      0.345     -0.162      0.871        -0.733     0.621
touristy[T.True]    -2.5305      0.797     -3.175      0.001        -4.092    -0.969
trendy[T.True]      -0.5872      0.259     -2.269      0.023        -1.094    -0.080
review_count         0.0013      0.000      5.231      0.000         0.001     0.002
credit_card          0.0050      0.333      0.015      0.988        -0.647     0.657
Alcohol             -0.4729      0.056     -8.493      0.000        -0.582    -0.364
text_scores          0.4447      0.023     19.585      0.000         0.400     0.489
==============================================================================
```

Next steps

- Improving sentiment analysis
- Refining sentiment analysis feature
- Incorporating different models
- Expanding analysis to more cities

Improving Yelp predictions using sentiment analysis

General Assembly Data Science
Daniel Chung
October 6th, 2014