Sentiment Analysis

Qiao Hu

Source Data

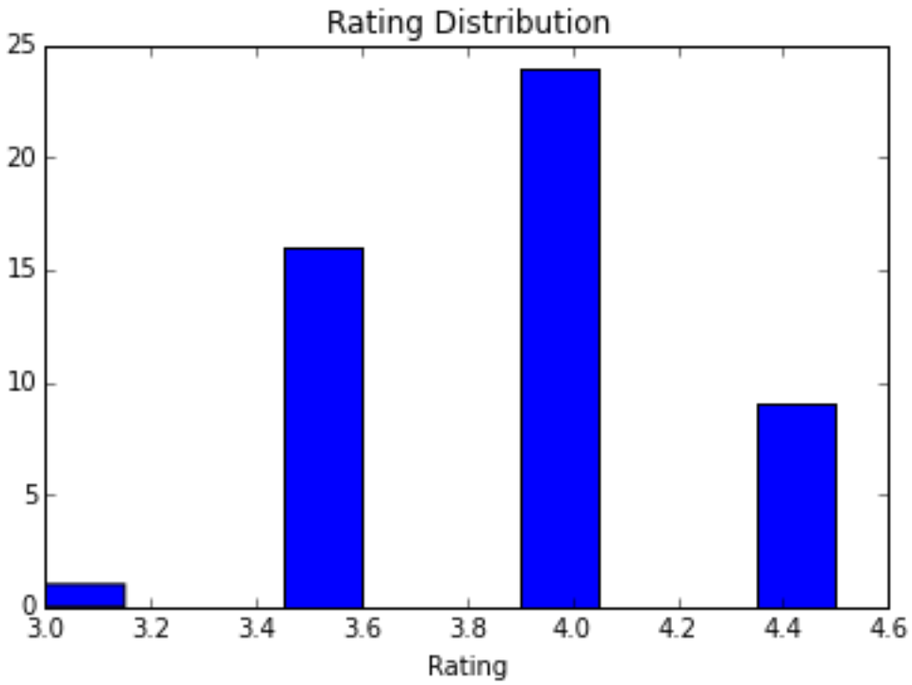
- JSON from Yelp API
  - **50 restaurants**

https://www.yelp.com/developers/documentation/v2/overview

- HTML from Restaurant URLs
  - **21675 reviews**

http://www.yelp.com/biz/gochi-japanese-fusion-tapas-cupertino

Distribution of restaurant ratings



correlation
(0.35434870439527888, 0.011577708119893972)

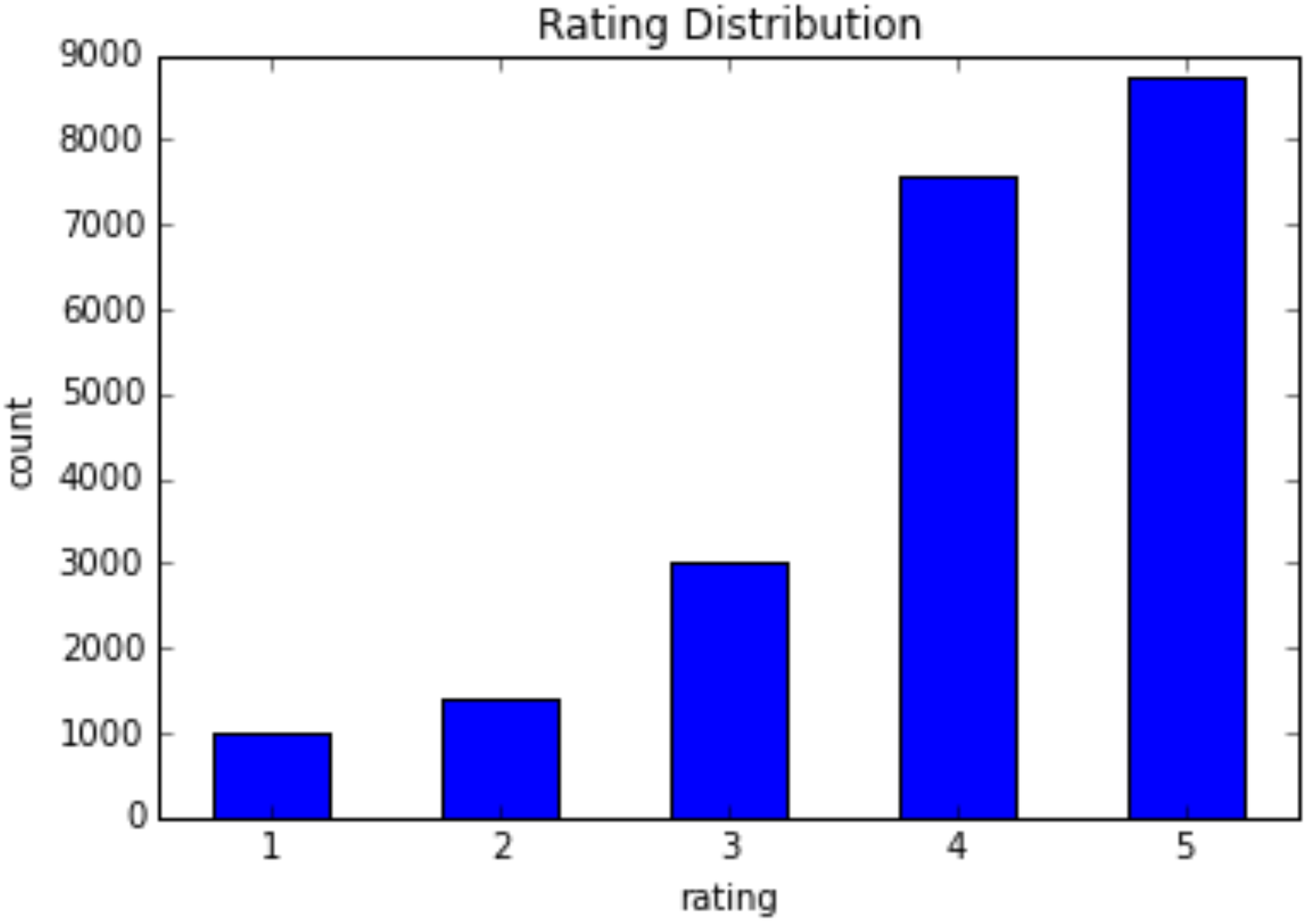
Data

Always one of my favorite places to eat. Too bad the only way to get a table here is to reserve a week beforehand or arrive right when the restaurant opens for the dinner.

The food there is just ok, but the service is awful! One waiter was even very rude to the customers! Don't go there if you want to be treated nicely!

The good: Free tea Service Fresh food The bad: Bland tuna Cramped tables Unclean bathroom The ugly: Bethesda at night

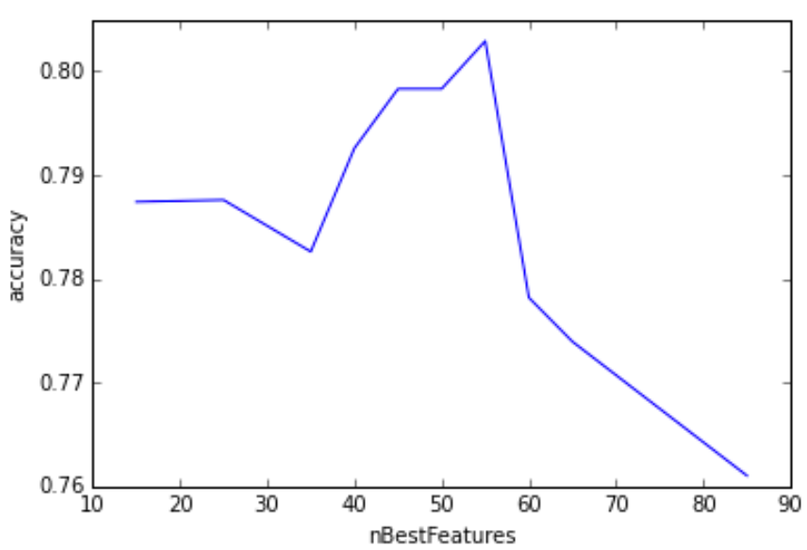
Distribution of review ratings

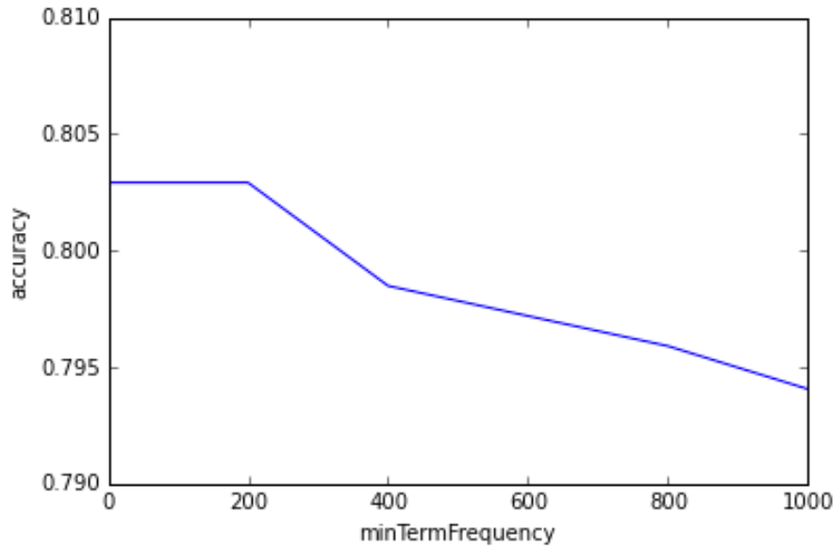


Naive Bayes

```
negsample = 5372
possample = 16303
training size:  16256, testing size: 5419
accuracy: 0.547887064034
pos precision: 0.962457337884
pos recall: 0.415112855741
neg precision: 0.348811800055
neg recall: 0.950856291884
```

Thresholds





nBestFeatures: the top n most significant features identified by chi.square

minTermFrequency: minimum frequency of feature

Naive Bayes - on Best Features

Most Influential Terms

word	label	probability	word	label	probability
worst	neg	18.5	bland	neg	4.6
rude	neg	10.6	average	neg	4.3
horrible	neg	10.1	ok	neg	3.9
terrible	neg	9.6	perfect	pos	3.9
mediocre	neg	9.5	okay	neg	3.8
disappointing	neg	8.1	asked	neg	3.3
meh	neg	6.9	amazing	pos	3
overpriced	neg	6.3	nothing	neg	3
understand	neg	5	seattle	pos	2.9
manager	neg	4.7	delicious	pos	2.9

```
accuracy: 0.802915667097
pos precision: 0.891666666667
pos recall: 0.840039254171
neg precision: 0.587080430652
neg recall: 0.690245718541
```

Naive Bayes - on Best N-Grams

Most Influential Terms

	word	labe	probability		word	labe	probability		word	labe	probabil
0	will definitely be	pos	34.8	**10**	terrible	neg	9.6	**20**	had to ask	neg	6.3
1	worst	neg	18.5	**11**	mediocre	neg	9.5	**21**	melts in your	pos	6.3
2	can't wait	pos	15.5	**12**	be sure to	pos	8.8	**22**	overpriced	neg	6.3
3	3 stars	neg	15.2	**13**	disappointing	neg	8.1	**23**	of my favorite	pos	6.2
4	definitely be back	pos	13.5	**14**	can't go wrong	pos	7.4	**24**	used to be	neg	6.2
5	hands down	pos	12	**15**	i really enjoyed	pos	7.1	**25**	love this place	pos	6.2
6	my favorites	pos	11.3	**16**	meh	neg	6.9	**26**	to die for	pos	6.2
7	rude	neg	10.6	**17**	restaurant in seattle	pos	6.9	**27**	must try	pos	5.8
8	melts in	pos	10.4	**18**	a must try	pos	6.9	**28**	best sushi i've	pos	5.6
9	horrible	neg	10.1	**19**	highly recommend	pos	6.8	**29**	i was expecting	neg	5.5

```
accuracy: 0.816940394907
pos precision: 0.887048192771
pos recall: 0.867026496565
neg precision: 0.622299651568
neg recall: 0.664929262844
```

Random Forest

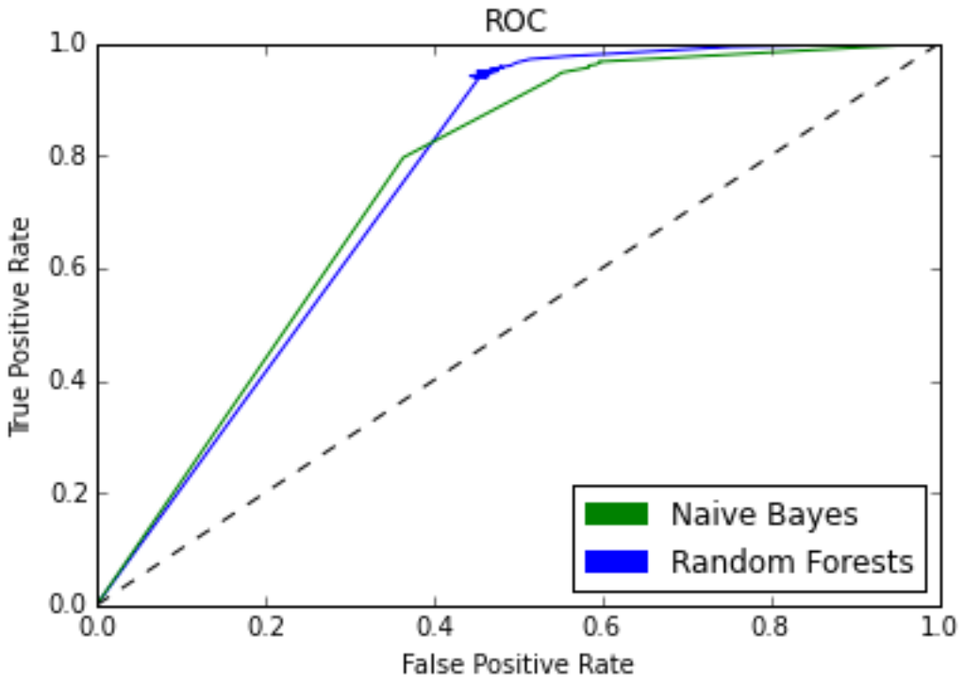
```
Estimators:   10
nBestWords:   100
Accuracy:   0.844225741965
tpr:   0.957674131909
fpr:   0.493276283619
```

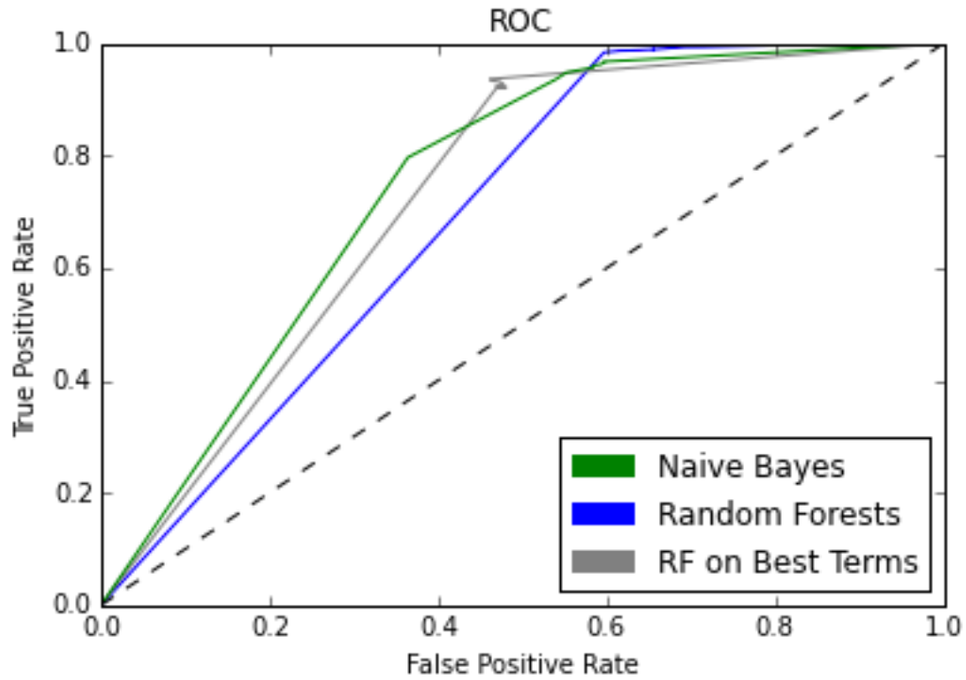
```
Feature ranking:
('1. feature 53 (0.045691) ', u'not')
('2. feature 4 (0.042109) ', u'and')
('3. feature 85 (0.036812) ', u'was')
('4. feature 39 (0.033020) ', u'is')
('5. feature 76 (0.032030) ', u'to')
('6. feature 16 (0.027622) ', u'but')
('7. feature 71 (0.024626) ', u'sushi')
('8. feature 92 (0.024447) ', u'you')
('9. feature 34 (0.023072) ', u'great')
('10. feature 23 (0.021592) ', u'delicious')
('11. feature 55 (0.020578) ', u'ok')
('12. feature 74 (0.020307) ', u'that')
('13. feature 47 (0.019772) ', u'mediocre')
('14. feature 87 (0.019534) ', u'we')
('15. feature 6 (0.017825) ', u'are')
('16. feature 88 (0.017330) ', u'were')
('17. feature 12 (0.016305) ', u'best')
('18. feature 25 (0.014696) ', u'didn')
('19. feature 40 (0.013747) ', u'just')
('20. feature 13 (0.013616) ', u'better')
('21. feature 91 (0.013428) ', u'worst')
('22. feature 32 (0.012949) ', u'fresh')
('23. feature 54 (0.012866) ', u'nothing')
('24. feature 35 (0.012791) ', u'happy')
('25. feature 46 (0.012770) ', u'me')
('26. feature 22 (0.012663) ', u'definitely')
('27. feature 3 (0.012497) ', u'amazing')
('28. feature 42 (0.012402) ', u'love')
('29. feature 56 (0.012239) ', u'okay')
('30. feature 11 (0.012145) ', u'bad')
('31. feature 9 (0.011993) ', u'average')
```

Tf-idf Random Forest

```
Estimators:   10
nBestWords:   100
Accuracy:   0.84530216823
tpr:   0.953201970443
fpr:   0.47700797057
```

Model comparison





Conclusion

Most Influential Terms

	word	label	probability		word	label	probability		word	label	probabilit
0	will definitely be	pos	34.8	**10**	terrible	neg	9.6	**20**	had to ask	neg	6.3
1	worst	neg	18.5	**11**	mediocre	neg	9.5	**21**	melts in your	pos	6.3
2	can't wait	pos	15.5	**12**	be sure to	pos	8.8	**22**	overpriced	neg	6.3
3	3 stars	neg	15.2	**13**	disappointing	neg	8.1	**23**	of my favorite	pos	6.2
4	definitely be back	pos	13.5	**14**	can't go wrong	pos	7.4	**24**	used to be	neg	6.2
5	hands down	pos	12	**15**	i really enjoyed	pos	7.1	**25**	love this place	pos	6.2
6	my favorites	pos	11.3	**16**	meh	neg	6.9	**26**	to die for	pos	6.2
7	rude	neg	10.6	**17**	restaurant in seattle	pos	6.9	**27**	must try	pos	5.8
8	melts in	pos	10.4	**18**	a must try	pos	6.9	**28**	best sushi i've	pos	5.6
9	horrible	neg	10.1	**19**	highly recommend	pos	6.8	**29**	i was expecting	neg	5.5