sungbin.andy.kang@berkeley.edu Â¶

InÂ [7]:

```
output:

url: http://boingboing.net/#85534171
 date: not supplied

 arts and letters daily, a wonderful and dense blog, has folded up its tent due
 to the bankruptcy of its parent company. a&l daily will be auctioned off by the
 receivers. link[1] discuss[2] (_thanks, misha!_)

 [1] http://www.aldaily.com/
 [2] http://www.quicktopic.com/boing/h/zlfterjnd6jf



<html>
 <head>
 </head>
 <body>
 <font size=3d"4"><b> a man endowed with a 7-8" hammer is simply<br>
  better equipped than a man with a 5-6"hammer. <br>
 <br>would you rather have<br>more than enough to get the job done or fall =
 short. it's totally up<br>to you. our methods are guaranteed to increase y=
 our size by 1-3"<br> <a href=3d"http://209.163.187.47/cgi-bin/index.php?10=
 004">come in here and see how</a>
 </body>
 </html>
```

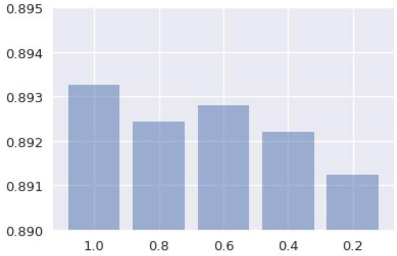
The spam email has a html formatting.

1. Sensitivity: 0, Specificity: 1
2. 0.744
3. It is barely better than a classifier that predicts 0 (ham) for every email.
4. Sensitivity: 0.077, Specificity: 0.989, False Negatives
5. This classifier is saying ham for most emails.

1. I followed general guidelines from the question itself, and looked through the emails to find better words for the bag-of-word method.
2. Using my intuition for words that immediately suggest spam to me worked pretty well. Using length of the email didn't work.
3. Some words didn't seem to increase the accuracy of the classifer at all.

InÂ [259]:

```
output:
```

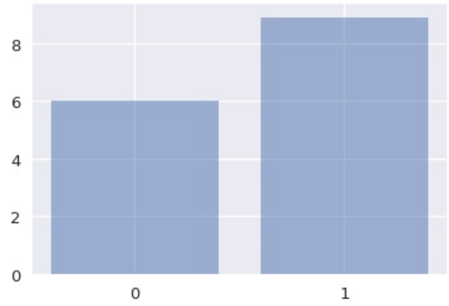


Out[259]:

```
'\nThis graph shows the change in accuracy depending on the regularization value, or rather the inverse of regularization value.\nThe graph shows two modes, with the higher mode around 1.0.\n'
```

InÂ [268]:

```
output:
```

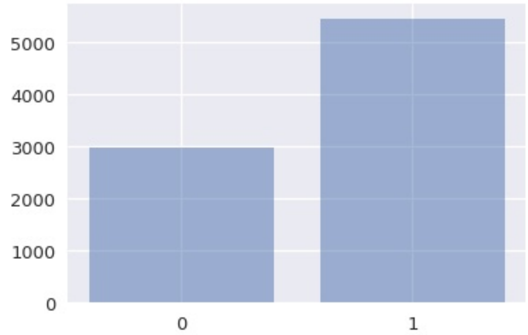


Out[268]:

```
'\nThis graph shows that spam emails had an average of more than 8 "!" or "?" per email whereas ham emails had an average of around 6 "!" or "?" per email.\n'
```

InÂ [271]:

```
output:
```



Out[271]:

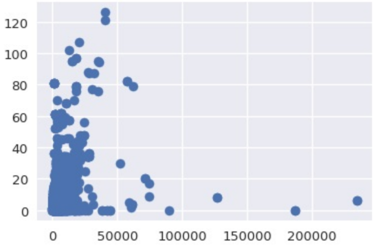
```
'\nThis bar plot shows that spam emails on average were more than 5000 characters long, whereas ham emails were about 3000\ncharacters long.\n'
```

InÂ [279]:

```
output:
```

Out[279]:

```
'\nThis scatter plot shows a very steep increase in the number of exclamations and question marks as the spam emails increased in length.\nThis correlation is also found in ham emails, but to a mu
```

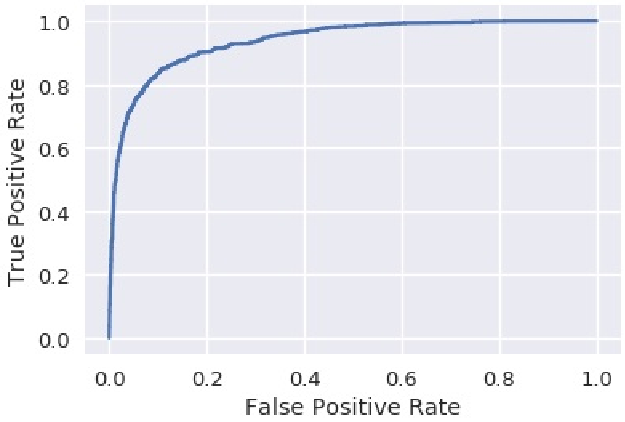


InÂ [287]:

```
output:
```

Out[287]:

```
Text(0,0.5,'True Positive Rate')
```


True Positive Rate
False Positive Rate