Notebook

April 15, 2020

0.0.1 Question 1c

Discuss one thing you notice that is different between the two emails that might relate to the identification of spam.

The second email that is denoted as spam has alot of greater than and less than signs as well as a link that starts with numbers.

0.0.2 Question 3a

Create a bar chart like the one above comparing the proportion of spam and ham emails containing certain words. Choose a set of words that are different from the ones above, but also have different proportions for the two classes. Make sure to only consider emails from `train`.

```
In [164]: train
```

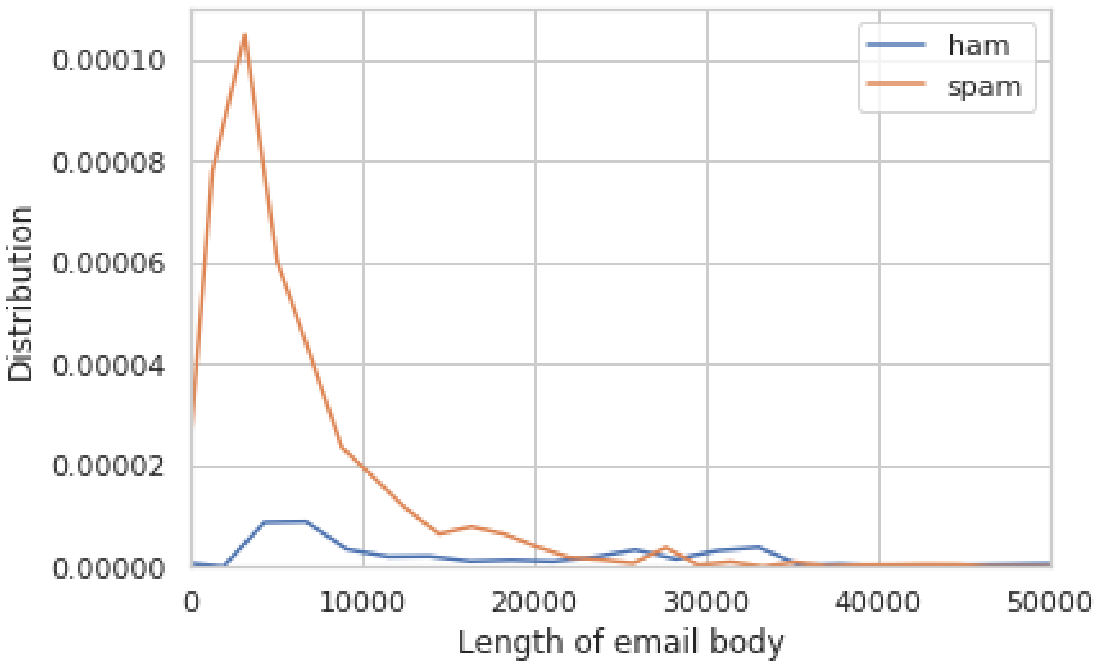
```
Out[164]:         id                                            subject  \
          7657  7657             Subject: Patch to enable/disable log\n
          6911  6911        Subject: When an engineer flaps his wings\n
          … Omitting 20 lines …
          7270  chris haun wrote:\n > \n > we would need someo…     0

          [7513 rows x 4 columns]
```

0.0.3 Question 3b

Create a *class conditional density plot* like the one above (using `sns.distplot`), comparing the distribution of the length of spam emails to the distribution of the length of ham emails in the training set. Set the x-axis limit from 0 to 50000.

```
In [178]: new = train[['email', 'spam']]
          new['email'] = new['email'].apply(len)
          new_ham = new.loc[new['spam'] ==0]
          new_spam = new.loc[new['spam'] ==1]
          sns.set(style="whitegrid")
          ax = sns.distplot(new_ham[['email']], hist = False, label = "ham")
          ax = sns.distplot(new_spam[['email']], hist = False, label = "spam")
          plt.xlim(0,50000)
          ax= ax.set(xlabel="Length of email body", ylabel="Distribution")
```

```
/srv/conda/envs/data100/lib/python3.7/site-packages/ipykernel_launcher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
… Omitting 0 lines …
See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.h
```


ham
spam
Distribution
Length of email body