Notebook

April 19, 2020

0.0.1 Question 1c

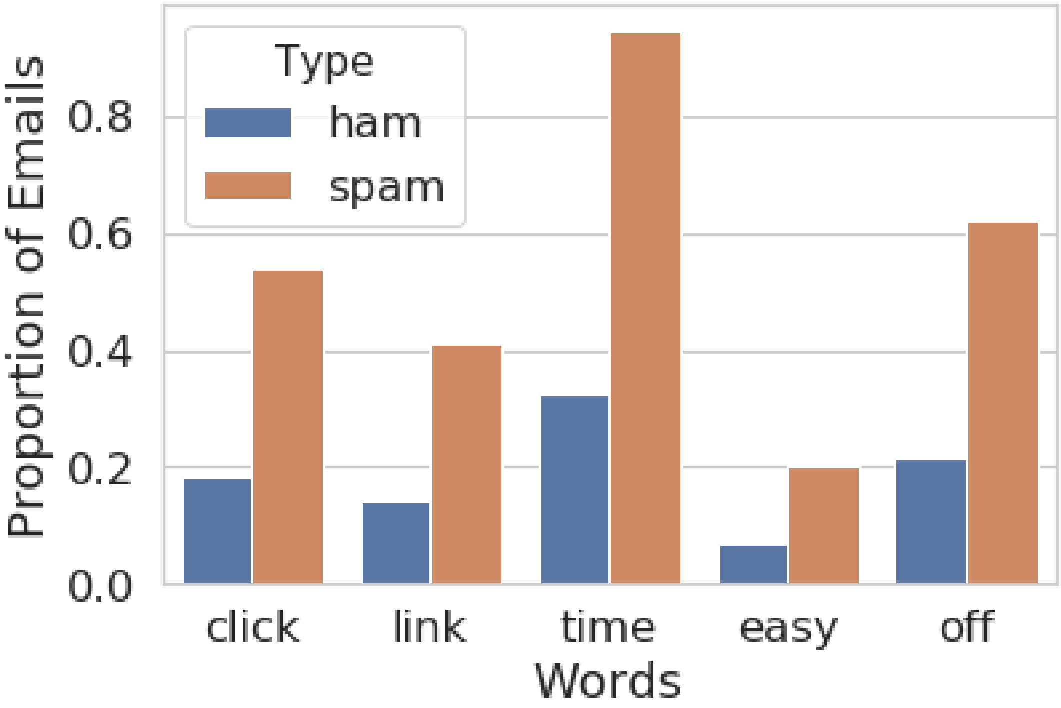
Discuss one thing you notice that is different between the two emails that might relate to the identification of spam.

Spam may be more inclined to use html content type.

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`.

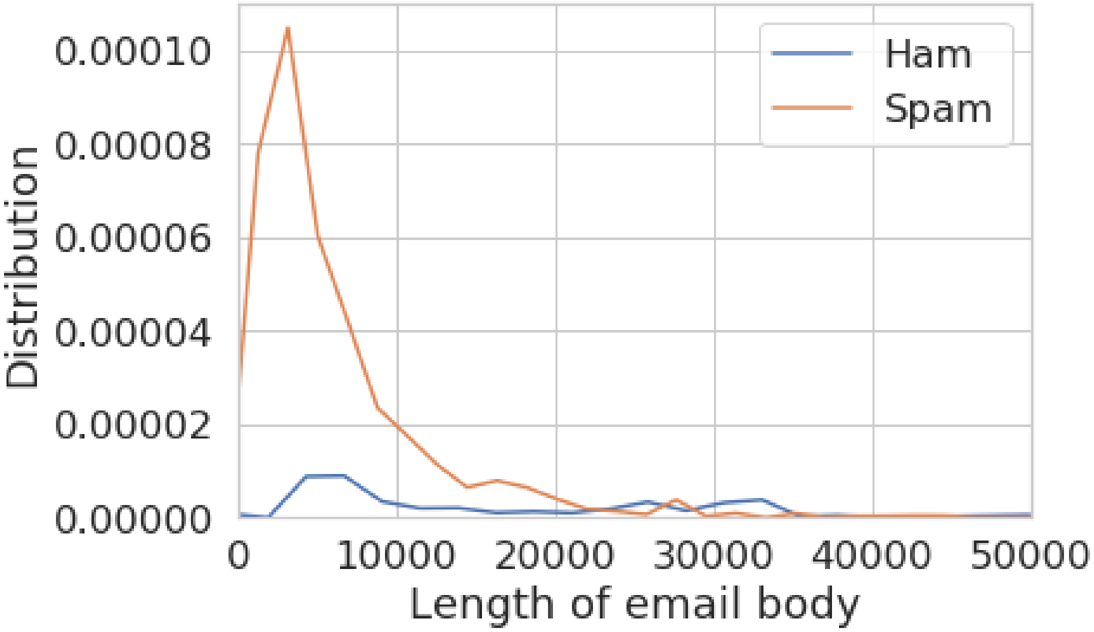
```
In [12]: train=train.reset_index(drop=True) # We must do this in order to preserve the ordering of emai

         some_words = ['click', 'link', 'time', 'easy', 'off']
         ham_count = np.sum(words_in_texts(some_words, train[train['spam']==0]['email']),axis=0)
         ham_proportion = ham_count/train[train['spam']==0].shape[0]
         spam_count = np.sum(words_in_texts(some_words, train[train['spam']==1]['email']),axis=0)
         spam_proportion = ham_count/train[train['spam']==1].shape[0]
         plot_data = pd.DataFrame([ham_proportion,spam_proportion], columns=some_words, index = ['ham',
         plot_data = plot_data.stack()
         plot_data = plot_data.reset_index()
         plot_data.columns = ['Type','Words','Proportion of Emails']
         #print(plot_data)
         sns.barplot(data=plot_data, hue = 'Type', x = 'Words', y = 'Proportion of Emails');
```



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 [13]: train[train['spam']==1]['email'].str.len()
         sns.distplot(train[train['spam']==0]['email'].str.len(), hist=False, label='Ham')
         sns.distplot(train[train['spam']==1]['email'].str.len(), hist=False, label='Spam')
         plt.xlim(0, 50000)
         plt.xlabel("Length of email body")
         plt.ylabel("Distribution");
```


Distribution
Length of email body
Ham
Spam