
0.1 Question 1

Discuss one attribute or characteristic you notice that is different between the two emails that might relate to the identification of a spam email.

One email is in html code and has a suspicious URL. Whereas the ham email is more personalized and is not in HTML code.

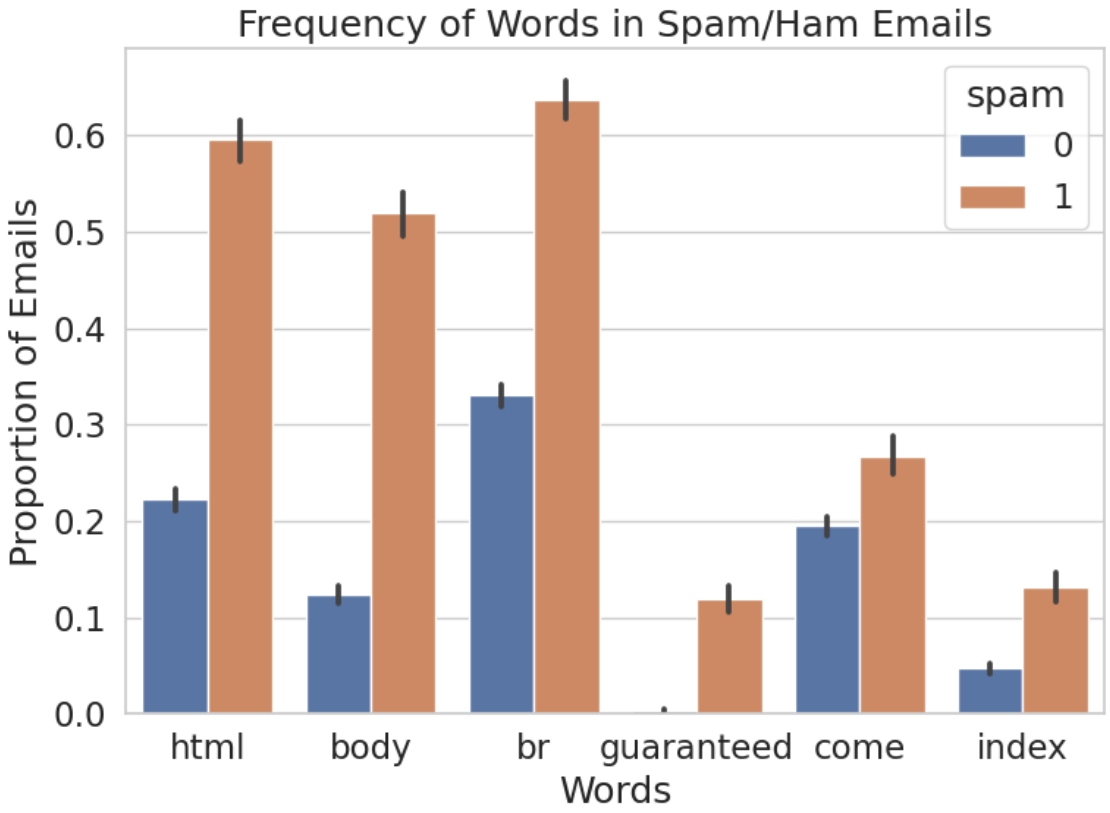
Create your bar chart with the following cell:

```
In [51]: train = train.reset_index(drop=True) # We must do this in order to preserve the ordering of em
         plt.figure(figsize=(8,6))

         array = ['html', 'body', 'br', 'guaranteed', 'come', 'index']
         df = pd.DataFrame(words_in_texts(array, train['email']))
         df['spam'] = train['spam']
         df = df.melt('spam')

         sns.barplot(x = df['variable'], y = df['value'], hue = df['spam']).set_xticklabels(array)
         plt.title('Frequency of Words in Spam/Ham Emails')
         plt.xlabel('Words')
         plt.ylabel('Proportion of Emails')

         plt.tight_layout()
         plt.show()
```



0.2 Question 6c

Explain your results in Question 6a and Question 6b. How did you know what to assign to `zero_predictor_fp`, `zero_predictor_fn`, `zero_predictor_acc`, and `zero_predictor_recall`?

zero_predictor_fp and zero_predictor_recall are both 0 since the amount of true positives = 0. Since TP = 0, zero_predictor_fn and zero_predictor_acc are just variations of sum(Y_train) which gives us the number of false negatives. So, zero_predictor_fn = sum(Y_train) and zero_predictor_acc is the amount of len-sum/len for Y_train.

0.3 Question 6f

How does the accuracy of the logistic regression classifier `my_model` compare to the accuracy of the zero predictor?

Less accurate than the zero_predictor.

0.4 Question 6g

Given the word features provided in Question 4, discuss why the logistic regression classifier `my_model` may be performing poorly.

Hint: Think about how prevalent these words are in the email set.

The words that were used in question 4 might not have been frequent enough in the Ham email set or in the Spam email set. In other words, they are not good at differentiating.

0.5 Question 6h

Would you prefer to use the logistic regression classifier `my_model` or the zero predictor classifier for a spam filter? Why? Describe your reasoning and relate it to at least one of the evaluation metrics you have computed so far.

my_model. The reason is that the zero_predictor is not really predicting anything it is just assuming that all emails are Ham, if there is an increase of Spam emails, this zero_predictor becomes poor. On the other hand, the my_model predictor would continue to work at a similar level to it is now and would be more reliable in the long run with more emails of each category.