
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.

The spam message is in HTML form, which is hard to read. And the ham message doesn't seems like HTML form. So HTML may be a factor of considering if an email is spam or not.

Create your bar chart with the following cell:

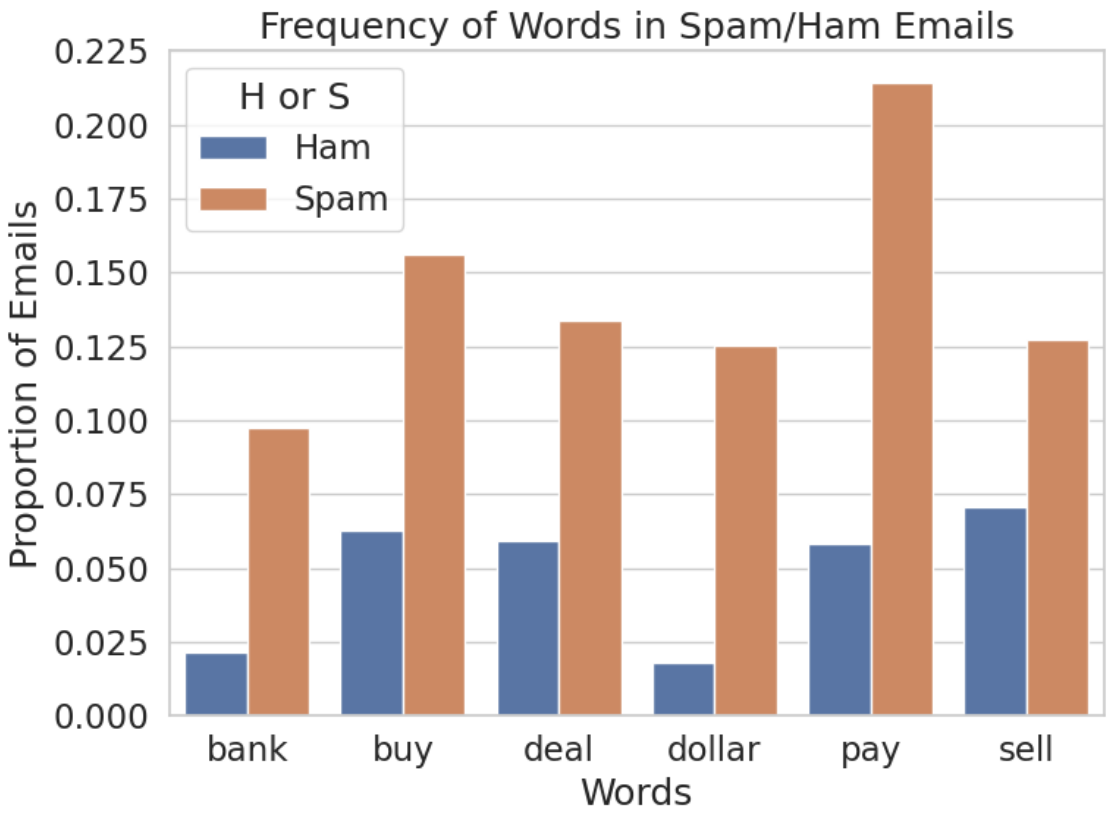
```
In [12]: train = train.reset_index(drop=True) # We must do this in order to preserve the ordering of em
         plt.figure(figsize=(8,6))

         words = ['buy', 'bank', 'dollar', 'deal', 'sell', 'pay']
         train_classfier = words_in_texts(words, train['email'])
         word_classfier_df = pd.DataFrame(train_classfier, columns = words)
         word_classfier_df['H or S'] = train['spam']
         word_classfier_df = word_classfier_df.replace({'H or S':{0 : 'Ham',1 : 'Spam'}})


         (sns.barplot(x = "variable", y = "value", hue = "H or S",
                      data = (word_classfier_df.melt('H or S').groupby(['H or S', 'variable']).mean().re

         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

Comment on the results from 6a and 6b. For **each** of FP, FN, accuracy, and recall, briefly explain why we see the result that we do.

FP is we predict positive but it was wrong. In our case, all our predictions are 0, so it is 0. FN is we predict negative but it was wrong. In our case, all our predictions are 0. In order to find where we predict wrong, we just sum our real data where is one, so we can check how many it is.
Accuracy means what persentage of our predict which is correct. So we check real data and see how many 0 we got, or just check what persentage that 0 weighted. Recall = TP / (TP + FN). In our case, our TP is 0, because we never predict 1. Therefore our recall is 0.

0.3 Question 6e

Are there more false positives or false negatives when using the logistic regression classifier from Question 5? Take a look at your result in 6d!

There are more false negatives than false positives. We got FN=1699 and FP 122

0.4 Question 6f

Our logistic regression classifier got 75.76% prediction accuracy (number of correct predictions / total). How does this compare with predicting 0 for every email?

Our zero predictor with accuracy around 74.47%. It means not big difference on accuracy between thoese two. Our logistic regression classifier is only little better than zero predictor.

0.5 Question 6g

Given the word features we gave you above, name one reason this classifier is performing poorly. **Hint:** Think about how prevalent these words are in the email set.

One reason our features are performing poorly is that maybe the features we get is not appear in most email. So if that happened, classifier can't tell which email is spam based on thoese features.

0.6 Question 6h

Which of these two classifiers would you prefer for a spam filter and why? Describe your reasoning and relate it to at least one of the evaluation metrics you have computed so far.

I would prefer our original logistic regression classifer, because it has higher recall than the zero predictor. Our logistic regression classifer performs poorly, but we can improve it by adding more powerful features.

Zero predictor predict all the thing 0, so it will have a bad evaluation under situation that most emails are spam. It cannot classify if the email is spam or not.