
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 first email has a sign-off from a familiar name, while the spam email has no sign-off.

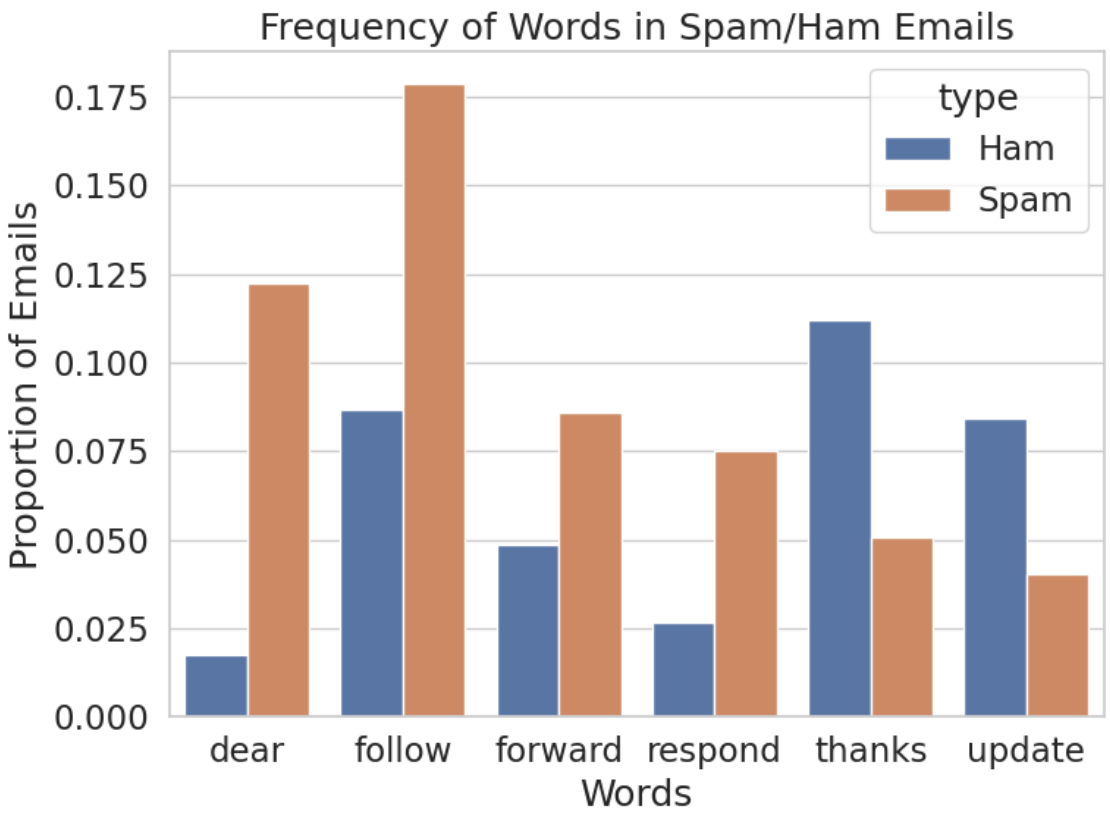
Create your bar chart with the following cell:

```
In [72]: train = train.reset_index(drop=True) # We must do this in order to preserve the ordering of em
         plt.figure(figsize=(8,6))

         words = ['follow','respond','update','forward','dear','thanks']
         d2_array = words_in_texts(words, train['email'])
         df = pd.DataFrame(d2_array,columns=words)
         df['type'] = train['spam'].map({1:'Spam',0:'Ham'})
         melted = df.melt('type')

         props = melted.groupby(['type','variable']).agg('mean').reset_index()
         sns.barplot(data=props, x='variable', y='value', hue='type')
         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 measures how many ham emails are marked as spam. Since the zero predictor always predicts ham, none will get marked as spam. FN measures how many spam emails are marked as ham. Since all the spam emails will be predicted as ham, FN is the number of spam emails in the data set. Accuracy measures the proportion of data points that are predicted correctly, which would be the proportion of ham emails to all emails since we always predict ham. Recall measures the proportion of spam emails that were correctly marked as spam, which is none since the zero predictor only predicts ham.

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 using the logistic regression classifier.

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?

The logistic regression classifier is only 1% more accurate than the 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.

The word features that I gave above don't appear often in the training set, which is why it's hard to classify just based on those.

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 the logistic regression as a spam filter because it has a higher predictor and would stil flag some spam whereas the zero-predictor would never mark spam.