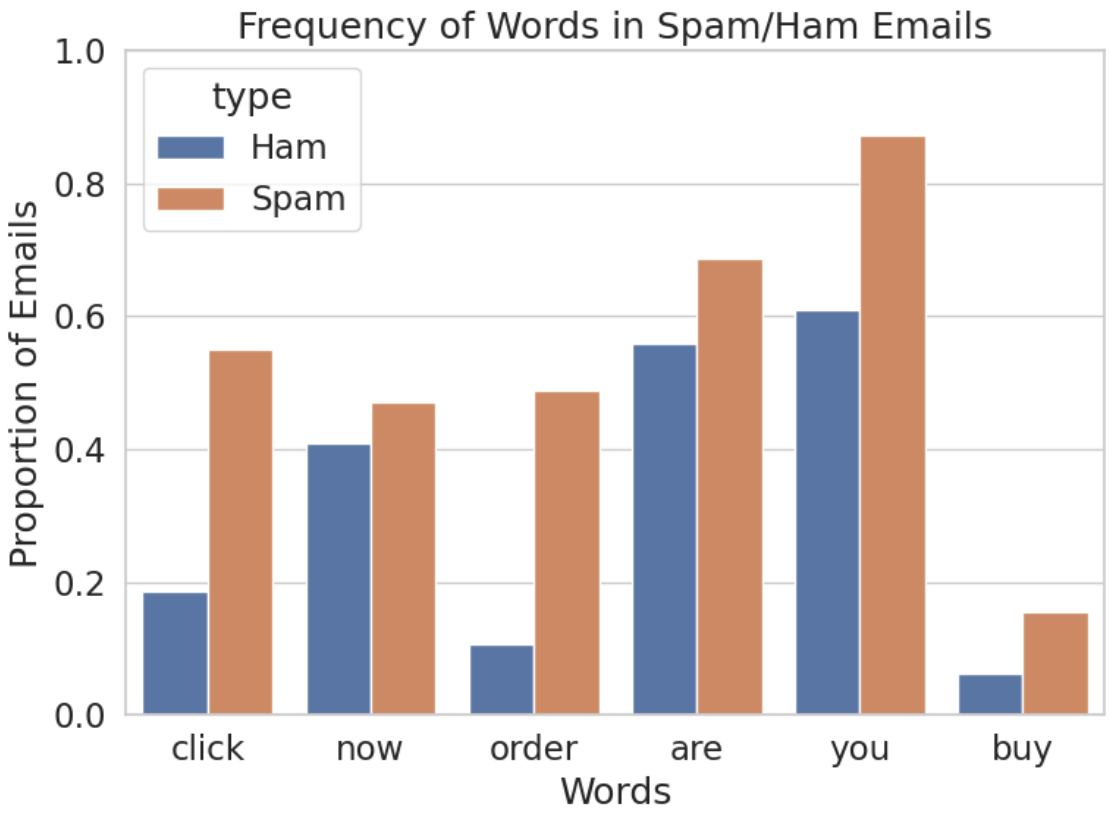

0.1 Question 1

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

I feel like the spam is in an html format and the email is not. It has the url, date, body, and the attatched links. The spam is in html formatting and it make it look very convoluted.

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 = ['click', 'now', 'order', 'are', 'you', 'buy']
         d = words_in_texts(words, train['email'])
         df = pd.DataFrame({words[0]: d[:, 0],
                            words[1]: d[:, 1],
                            words[2]: d[:, 2],
                            words[3]: d[:,3],
                            words[4]: d[:,4],
                            words[5]: d[:,5],
                            'type': train['spam']})
         df= df.replace({'type':{0: "Ham",
                                 1: "Spam"}})
         df = df.melt('type')
         sns.barplot(data= df, x='variable', y= 'value', ci= None, hue= 'type')
         plt.xlabel('Words')
         plt.ylabel('Proportion of Emails')
         plt.title('Frequency of Words in Spam/Ham Emails')
         plt.ylim(top= 1)
         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.

Since we are using a perdictor that always predicts zero, there won't be any false positives. There will be false negatives because all the spam will be classified as emails. Since there are no true positives or fals positives because we are predicting 0 for all, the accuracy is just true negative divided by the sum of thr true negatives and false negatives. Again since there won't be any true positives recall will be 0 since the equation is TP/(TP + FN).

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 positives when 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 does slightly better than the zero predictor. The zero predictor's accuracy is 74.47%. I would have thought that the zero predictor would have been way worse than the logistic regressor, but I guess this means that many of the emails are actually emails and very few are spam.

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.

I think it could be because many of the words we are using are found more in spam than emails. This means that we are training our model to find these words and calssify more emails as spam, which causes our precision to drop.

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 still choose the the logistic regressor, because both its accuracy and recall are higer than the zero predictor's. Since the zero predictor has a 0 for recall this means it lets all the spam pass through as emails, which is not what we want.