0.1 Question 1

Discuss one attribute or characteristic you notice that is different between the two emails that may allow you to uniquely identify a spam email.

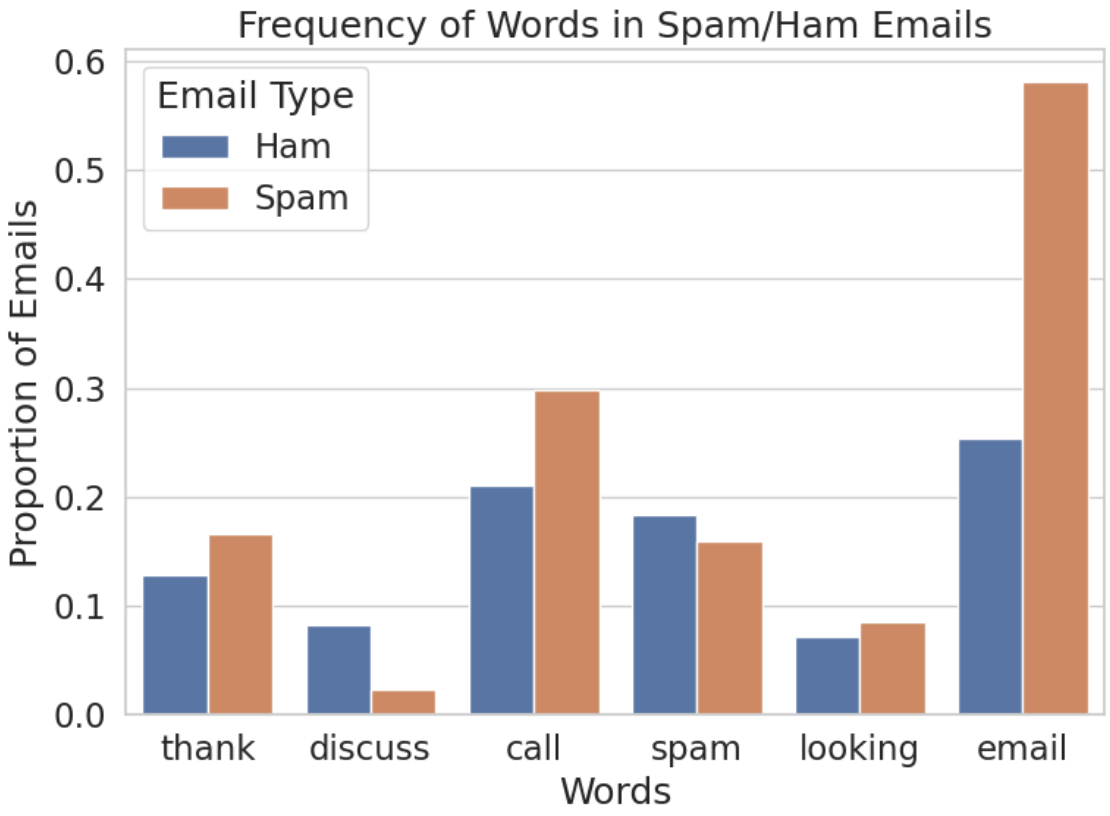
For the spam email, there are a lot of HTML tags such as $< html >$ or $< br >$ in between the texts, unlike the ham email, which doesn't have any.

Create your bar chart in the following cell:

```
In [40]: train = train.reset_index(drop=True) # We must do this in order to preserve the ordering of em
         plt.figure(figsize=(8,6))
         words = ['thank', 'discuss', 'call', 'spam', 'looking', 'email']
         ham_texts = train[train['spam'] == 0]['email'].str.lower()
         spam_texts = train[train['spam'] == 1]['email'].str.lower()
         ham_indicator = words_in_texts(words, ham_texts)
         spam_indicator = words_in_texts(words, spam_texts)
         ham_proportions = np.mean(ham_indicator, axis=0)
         spam_proportions = np.mean(spam_indicator, axis=0)

         df_proportions = pd.DataFrame({'Words': words,'Ham': ham_proportions,'Spam': spam_proportions}
         df_melt = pd.melt(df, id_vars='Words', var_name='Email Type', value_name='Proportion')

         sns.barplot(x='Words', y='Proportion', hue='Email Type', data=df_melt)
         plt.title('Frequency of Words in Spam/Ham Emails')
         plt.ylabel('Proportion of Emails')
         plt.tight_layout()
         plt.show()
```



0.2 Question 6c

Explain your results in `q6a` and `q6b`. How did you know what to assign to `zero_predictor_fp`, `zero_predictor_fn`, `zero_predictor_acc`, and `zero_predictor_recall`?

I was able to assign zero_predictor_fp, zero_predictor_fn, knowing that the zero predictor always predicts 0 and that there will be no cases of True Positives or False Positives. With that, we know that the False Negatives would just be all the cases of spam, since none of the spam emails are classified as spam. Therefore, I can assign the accuracy by calculating (TP)/(TP + FP), which is 0 since TP is 0 and the recall by calculating TP/(TP + FN), which also 0 due to the same reason.

0.3 Question 6f

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

The accuracy is significantly higher (0.67) as the zero predictor had an accuracy of 0 due to not being able to predict any emails as spam and only as ham emails.

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.

It might be performing poorly due to the choice of words that are used to classify the emails. The words that were chosen, which is ['drug', 'bank', 'prescription', 'memo', 'private'], are not really a determining factor of a spam or a ham email. Moreover, these are words that aren't very prevalent in emails. Perhaps choosing more crucial identifier words could have improved the performance of the model.

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.

The logistic regression classifier my_model would rather be used as a classifier since it at least has some precision and recall compared to the zero predictor classifier. While the model doesn't seem to have the best performance, its ability to still have some True Positives makes it seem better than a zero predictor classifier that can't possibly give any True Positives.