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.

ATTRIBUTE: The presennce of an IP address in a website link. **REASONING:** This is usually a sign of a sketchy link, or at least a link to a non-professional webpage. I would be willing to bet that sketchy links are correlated with spam emails. For that reason, IP addresses in links mught be a good proxy for spam.

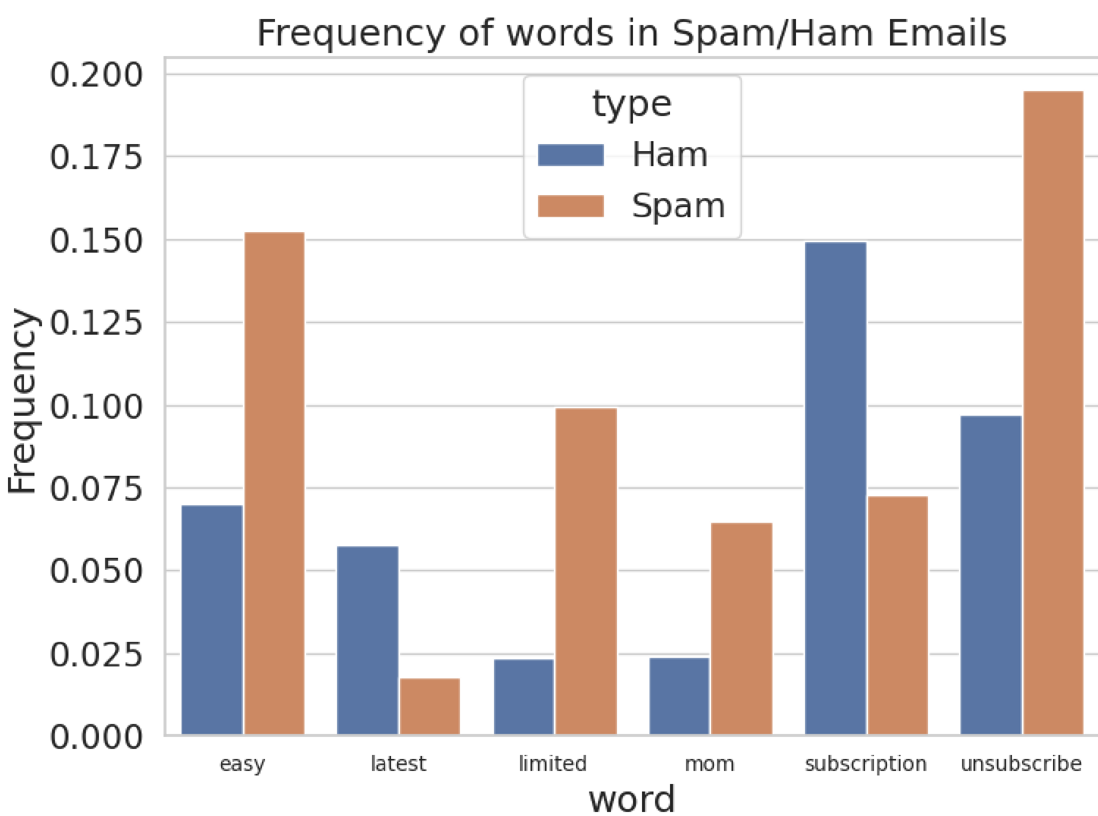
Create your bar chart in the following cell:

```
In [70]: train = train.reset_index(drop=True) # We must do this in order to preserve the ordering of em
         plt.figure(figsize=(8,6))

         words = ['easy', 'unsubscribe', 'subscription', 'limited', 'latest', 'mom']

         df = (pd.DataFrame(words_in_texts(words, train['email']), columns=words)
          .join(train.reset_index()[['spam']].replace({0:'Ham', 1:'Spam'}))
          .melt('spam')
          .groupby(['spam', 'variable'])
          .mean()
          .reset_index()
          .rename(columns={'spam':'type','variable':'word','value':'proportion'})
         )

         sns.barplot(df, x='word', y='proportion', hue='type')
         plt.title('Frequency of words in Spam/Ham Emails')
         plt.ylabel('Frequency')
         plt.xticks(fontsize=10)
         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`?

The predictor is better than randomly guesssing. It gets about a three quarters of its binary classifications right. (however this is just because only 25% of the emails are spam) The predictor never returns positive, so it will never return a positive value when the true value is negative. Thus, FP=0. The predictor always returns negative, so it will return a false negative for every sample that is actually postive. Thus, the number of false negatives is equal to the number of spam emails. The accuracy and recal are calculated using these values as well as the number of datapoints (accuracy).

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 of "my_model" (0.757) roughly equal to the zero predictor (0.744). These are comperable accuracy stats, meaning I would look to other measures like precision, recall, etc to judge model performance.

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 distributions of these words in spam and ham emails are overlapping, meaning that they are not relaibale indicators of the class. This is beacuse they are somewhat common/neutral words.

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 zero predictor is not a spam filter, functionally speaking. It would let every spam email into the inbox because it has a recall of 0, meaning it never identifies any spam emails to be filtered out. However, I would still perfer to use it over "my_model," because "my_model" has an fpr of 0.04, which means that essentially 1 out of every 25 ham emails will be filtered, which could have real consequences.