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.

Spam had email body embedded in HTML.

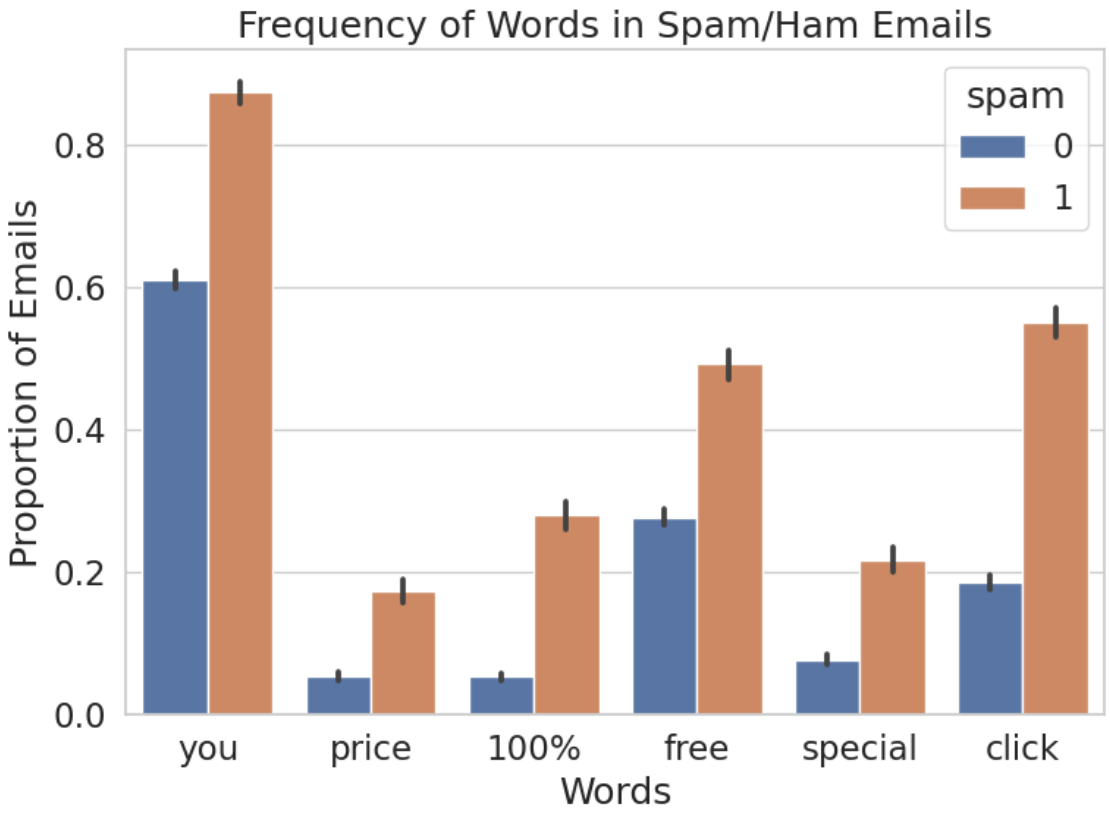
Create your bar chart with the following cell:

```
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 = ["you", "price", "100%", "free", "special", "click"]
         texts = train["email"]

         frequency = pd.DataFrame(words_in_texts(words, texts), columns= words)
         frequency["spam"] = train["spam"]
         new_frequency = frequency.melt("spam")

         sns.barplot(data = new_frequency, x = "variable", y = "value", hue = "spam")
         plt.xlabel('Words')
         plt.ylabel('Proportion of Emails')
         plt.title('Frequency of Words in Spam/Ham Emails')
         plt.tight_layout()
         plt.show()
```



0.2 Question 6c

Explain your results in Question 6a and Question 6b. How did you know what to assign to `zero_predictor_fp`, `zero_predictor_fn`, `zero_predictor_acc`, and `zero_predictor_recall`?

If zero predictor always predicts 0 then zero_predictor_fp would be 0, and zero_predictor_fn should be the number of spam emails since they gets labeled as ham. Zero_predictor_recall is 0 because we get no spam. Formula for zero_predictor_acc is TP + TN/len(emails), no spam, thus we get the percentage of spam = 0 (ham).

0.3 Question 6f

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

My q6e is False, so the number of false positives is not strictly greater than the number of false negatives thus the accuracy of my_model is not greater than the accuracy of the zero predictor.

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.

I choose [“you”, “price”, “100%”, “free”, “special”, “click”], these words can also show up in ham emails too making them less discriminatory.

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.

I would prefer the logistic regression classifier for a spam filter, it has higher prediction accuracy.