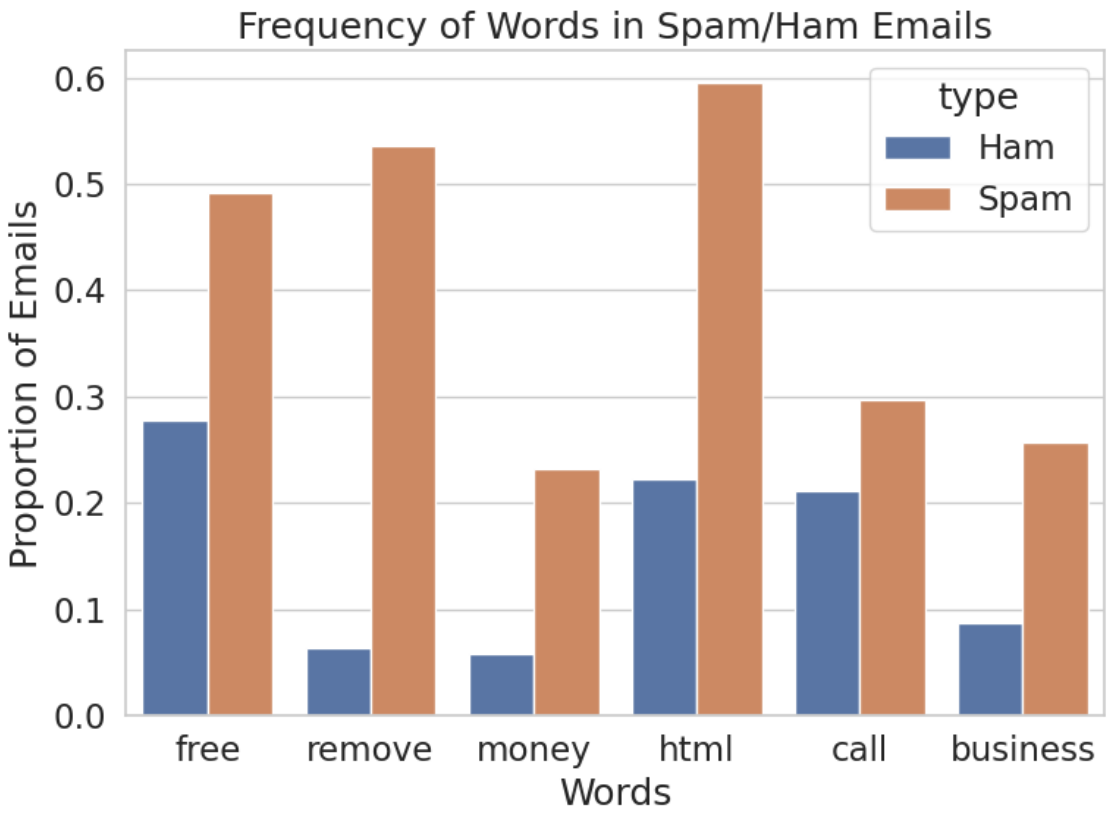
0.1 Question 1

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

One noticeable difference between the two emails is the use of highly promotional and persuasive language in the spam email. The spam message includes exaggerated financial promises and urgent calls to action. Meanwhile, ham emails use neutral, informational language without pressure or unrealistic claims. This pattern of urgent, sales-driven phrasing can be a strong indicator for identifying spam emails.

Create your bar chart in the following cell:

```
In [99]: train = train.reset_index(drop=True) # We must do this in order to preserve the ordering of em
         plt.figure(figsize=(8,6))

         words = ['free', 'remove', 'money', 'html', 'call', 'business']

         indicator_array = words_in_texts(words, train['email'])
         word_df = pd.DataFrame(indicator_array, columns=words)

         word_df['type'] = train['spam'].map({0: 'Ham', 1: 'Spam'})

         prop_df = word_df.groupby('type')[words].mean().reset_index()
         prop_long = prop_df.melt(id_vars='type', var_name='variable', value_name='value')

         sns.barplot(data=prop_long, x='variable', y='value', hue='type')
         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 `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 zero predictor always predicts 0, meaning it always labels emails as ham. Because of this, it can never produce a false positive, where the prediction is spam when the email is ham, so the zero_predictor_fp must be 0. However, every actual spam email will be incorrectly labeled as ham, which is why zero_predictor_fn equals the total number of spam messages in Y_train. Its accuracy is simply the proportion of ham emails in the dataset, since those are the only ones it gets correct, and its recall is 0 because it never correctly identifies any spam emails.

0.3 Question 6f

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

A logistic regression spam classifier should ideally avoid false negatives because missing a spam email means it reaches the user's inbox, which can be more harmful than incorrectly filtering out a legitimate email. In contrast, a false positive removes a ham email, which is inconvenient but generally less risky. Comparing false positives and false negatives helps us understand whether the model is more "conservative" or "aggressive" in identifying spam. Since both errors matter differently depending on the context, evaluating their balance gives insight into how well the classifier aligns with the goals of a safe and reliable email filter.

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 logistic regression classifier performs poorly because the five words chosen in Question 4, such as "drug," "bank," "prescription," "memo," and 'private", are extremely rare in the training dataset. When features appear in only a tiny fraction of emails, they provide almost no meaningful signal for distinguishing spam from ham, causing the model to default to predicting the majority class. As a result, the classifier fails to learn patterns that generalize and ends up with zero true positives and zero true negatives. In short, the model is ineffective because it is trying to learn from features that are too sparse to carry useful information.

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 not prefer either the logistic regression classifier or the zero predictor for a spam filter, but between the two, the logistic regression model is still the better option. Although it performed poorly with the limited word features we used, it can achieve non-zero recall and identify at least some spam once given more informative features. In contrast, the zero predictor has a recall of 0, meaning it never detects any spam at all, which makes it unusable for filtering harmful messages. Therefore, even though the current logistic model is weak, it has more potential to improve and provides more value than a classifier that always predicts ham.