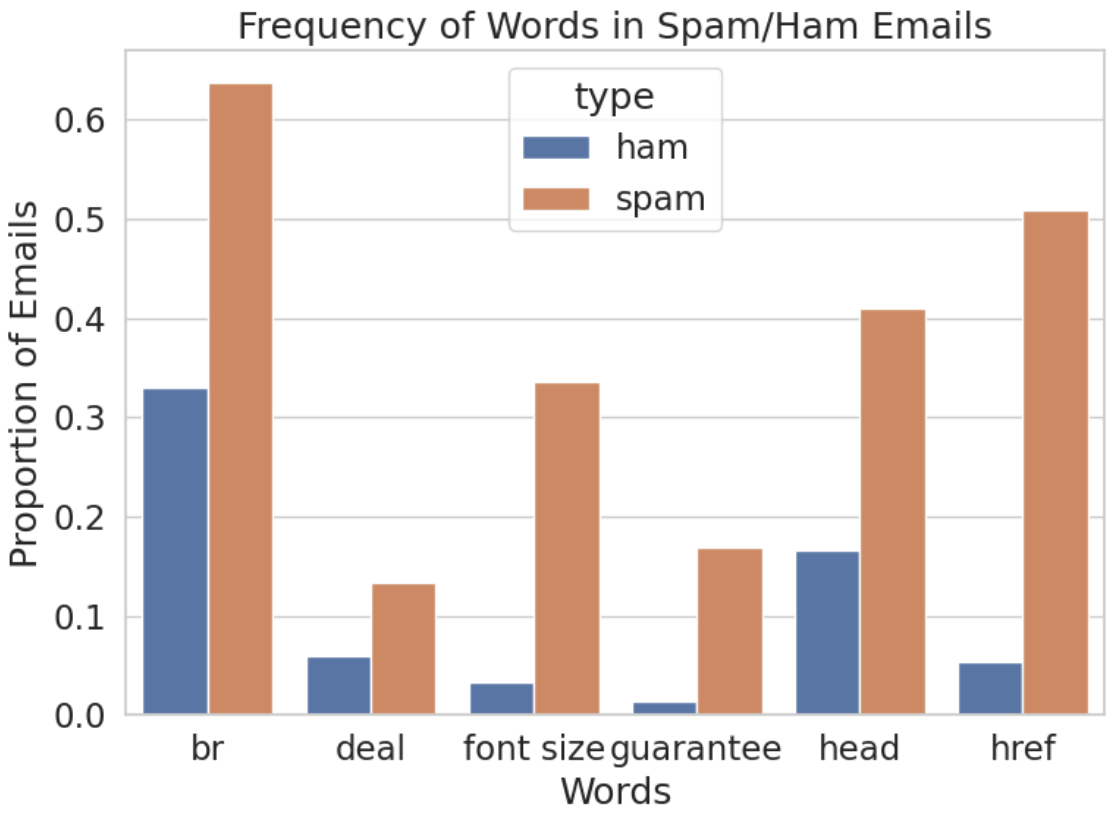
0.1 Question 1

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

In the example above, the spam email has '<head>', '<body>' '<html>' and '<font size>' while the normal email doesn't.

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

         # Create a new dataframe to plot
         train['type'] = train['spam'].map({0:'ham', 1:'spam'})
         list_words = ['head', 'font size', 'br','href', 'guarantee', 'deal']
         for word in list_words:
             train[word] = train['email'].str.contains(word).astype(int)

         list_words.insert(0,'type')
         new_df = train[list_words].melt('type').groupby(['type','variable']).mean().reset_index()

         # Plot
         plt.figure(figsize=(8,6))

         sns.barplot(data=new_df, 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

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.

FP = 0 because there is no predicted spam email. FN = total number of spam emails because they are all incorrectly flagged as non-spam emails. Accuracy is the same as the proportion of actual spam emails in the dataset, and recall is 0 because there is no predicted spam emails in the classifier model.

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 negatives (FN = 1699) than false positives (FP = 122) when using the logistic regression classifier from Question 5.

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?

They are pretty the same. Their difference is negligible. (Our logistic regression classifier got 75.76% while the predicting 0 for every email got 74.47%)

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.

The words used for this classifier are 'drug', 'bank', 'prescription', 'memo', and 'private.' Although these word can appear in spam emails, they can also appear in non-spam emails. For example, while 'drug' and 'prescription' can appear in a spam email trying to trick the readers to click and follow, they can also appear in legitimate emails from doctor or medical clinics. In other words, those words are not good to use as attributes that predict spam from ham emails.

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 prefer the first one that uses a set of words for a spam filter because while it does not do any better job than the latter classifier, it has a high precision rate (64%), meaning among predicted spam emails, 65% are accurately labeled as spams.