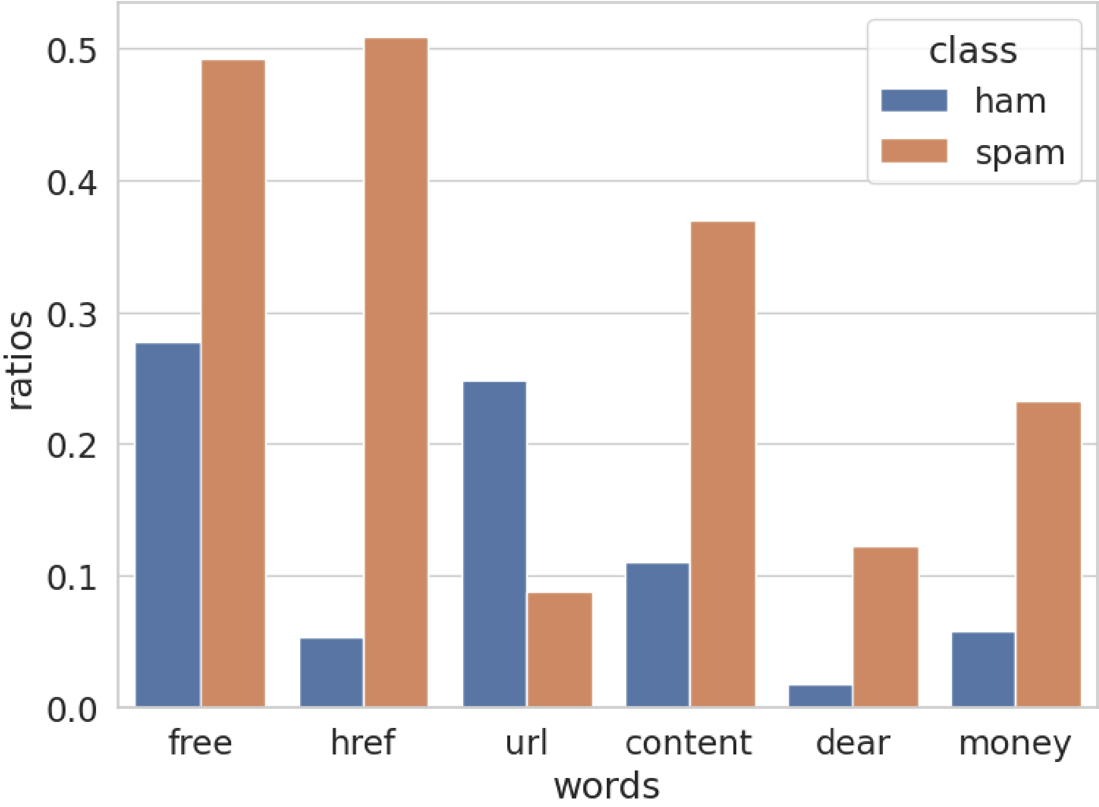
0.1 Question 1

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

We can see that the format used in both emails is different. While the ham email used plain text the spam email was HTML formatted with the tags indicating so

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 = ['free', 'href', 'url', 'content', 'dear', 'money']
         words_pd = pd.DataFrame(words_in_texts(words, train['email']), columns = words)
         words_pd['type'] = train['spam']
         spam_emails = []
         ham_emails = []
         for word in words:
             spam_emails.append(len(words_pd.loc[(words_pd['type'] == 1) & (words_pd[word] == 1)]) / le
             ham_emails.append(len(words_pd.loc[(words_pd['type'] == 0) & (words_pd[word] == 1)]) / len
         df = pd.DataFrame((ham_emails) + (spam_emails), columns=['proportions'])
         df['words'] = words + words
         df['class'] = ['ham'] * len(ham_emails) + ['spam'] * len(spam_emails)
         sns.barplot(x='words', y = 'proportions', hue='class' , data = df)
         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`?

zero_predictor_fp was set to be zero since the zero_predictor always predicts zero since it never predicts any positives so there will be zero false positives. Since zero_predictor_fn never predicts true positives, 1918 was the total amount of spam emails that were being mislabeled as ham emails. As for accuracy, zero_predictor_acc is the proportion of ham emails out of the total number of emails that were truly ham emails. Lastly, zero_predictor_recall is set as zero as true positives never occur.

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 obtained by my_model was 75.76% whilst the zero_predictor obtained 74.47. Thus, since its accuracy is slightly lower than my_model, my_model performed better

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.

my_model may be performing poorly due to the choice of words. There may not be significant differences in the number of times the chosen words appear in spam and ham (ex. url). In other words, for the model to perform better, the chosen words should appear a greater number of times in spam compared to ham and vice-versa.

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 to use the zero predictor classifier since even though my_model has a greater prediction accuracy compared to the zero predictor classifier as it is more beneficial to have false negatives than to have false positives. Thus, for the function of a spam filter, although the zero predictor classifier allows spam emails, it eliminates the risk of filtering out ham emails.