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.

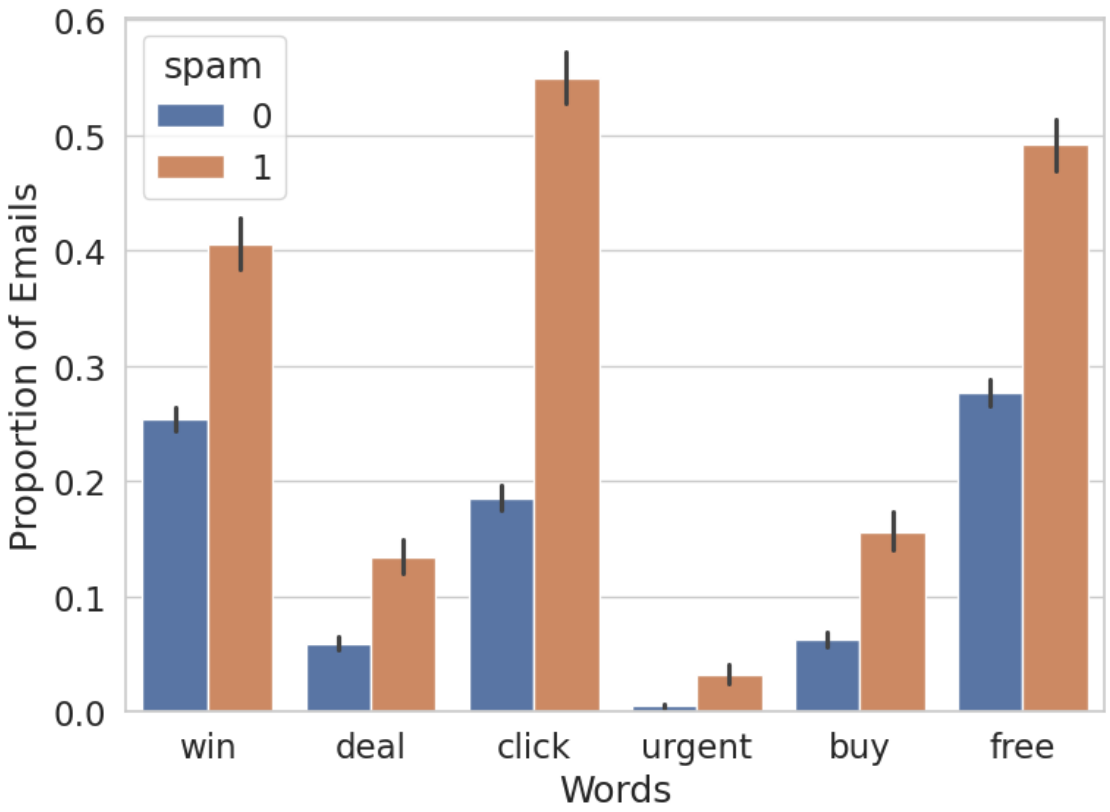
The difference between the two emails, which can be identify as a spam email is that oftern the sender is unknown, which the system of the e-mail shows as unknown person.

Create your bar chart in the following cell:

```
In [32]: train = train.reset_index(drop=True) # We must do this in order to preserve the ordering of em
         plt.figure(figsize=(8,6))
         spam_words = ['win', 'deal', 'click', 'urgent', 'buy', 'free']
         by_word = words_in_texts(spam_words, train['email'])
         by_word = pd.DataFrame(by_word)
         by_word['spam'] = train['spam']
         by_word = by_word.melt('spam')
         sns.barplot(x = by_word['variable'], y = by_word['value'], hue = by_word['spam']).set(xticklab
         plt.xlabel('Words')
         plt.ylabel('Proportion of Emails')

         plt.tight_layout()
         plt.show()
```

```
/tmp/ipykernel_107/1482277965.py:8: UserWarning: FixedFormatter should only be used together with Fixedl
  sns.barplot(x = by_word['variable'], y = by_word['value'], hue = by_word['spam']).set(xticklabels=spar
```



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`?

zero_predictor_fp is zero. Since the zero predictor never predicts 1, which cannot produce any false positives that I stated as 0. zero_predictor_fn can state as Y_train, since the model always predicts 0, every actual spam email (label = 1) becomes a false negative. To input this we can count how many 1s are in Y_train zero_predictor_acc can be calculated by True negative by the number of emails. zero_predictor_recall since recall can be calculated by True positive divide by True positive + False Positive, it shows that zero_predictor_recall will be 0

0.3 Question 6f

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

The result shows that the logistic regression is 0.59 and the accuracy of the zero predictor is 0.74 that from the result, it shows that the since the accuracy of the zero predictor is 74% it shows that the zero predictor is more accurate.

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 may be performing poorly because the word features chosen in Question 4 are either too common across both spam and emails or not sufficiently correlated with the spam label. This result the model to not effectively distinguish between the two classes based on these features, leading to low predictive accuracy and poor generalization.

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 logistic regression classifier my_model over the zero predictor. Although the zero predictor achieves higher accuracy by predicting ham, the predictor works by recalling the 0 and sometimes never identifies spam. In contrast, the logistic regression classifer may have lower accuracy, but it achieves non-zero recall, meaning it can correctly detect at least some spam emails. Thus, logistic regression is a more useful and practical choice in this context.