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 attribute I noticed is that the spam email contains a lot of raw HTML tags like

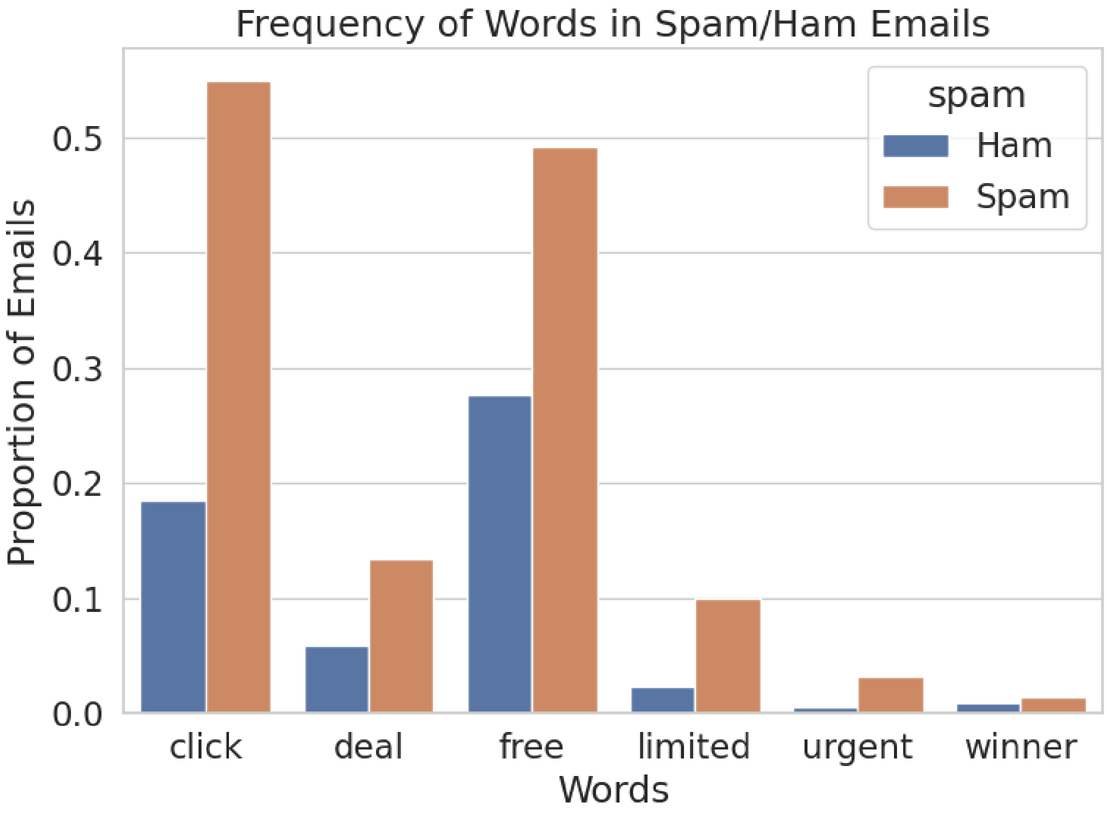
,

, and

, while the ham email does not. The presence of HTML formatting tags can help uniquely identify spam emails because spam messages often use HTML to hide malicious links and display misleading content.

Create your bar chart in the following cell:

```
In [41]: # Step 1: Pick new words
         interesting_words = ['free', 'winner', 'urgent', 'deal', 'limited', 'click']

         # Step 2: Check if those words appear
         X_train_words = words_in_texts(interesting_words, train['email'])

         # Step 3: Put into DataFrame
         X_train_df = pd.DataFrame(X_train_words, columns=interesting_words)
         X_train_df['spam'] = train['spam'].values

         # Step 4: Melt
         df_melted = X_train_df.melt(id_vars=['spam'], var_name='word', value_name='presence')

         # Step 5: Group and compute proportions
         word_counts = df_melted.groupby(['word', 'spam'])['presence'].mean().reset_index()
         word_counts['spam'] = word_counts['spam'].map({1: 'Spam', 0: 'Ham'})

         # Step 6: Plot
         plt.figure(figsize=(8,6))
         sns.barplot(x='word', y='presence', hue='spam', data=word_counts)
         plt.title('Frequency of Words in Spam/Ham Emails')
         plt.ylabel('Proportion of Emails')
         plt.xlabel('Words')
         plt.tight_layout()
         plt.show()
```


Frequency of Words in Spam/Ham Emails
spam
Ham
Spam
Proportion of Emails
0.0
0.1
0.2
0.3
0.4
0.5
click
deal
free
limited
urgent
winner
Words

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 (ham) for every email. • For zero_predictor_fp, there are no false positives because the model never predicts spam (1), so zero_predictor_fp = 0. • For zero_predictor_fn, every time an email is actually spam (1), it gets missed and predicted as ham (0). Therefore, zero_predictor_fn is the total number of spam emails in the training set, which was 1918. • For zero_predictor_acc, we calculate the proportion of correct predictions. The model is only correct when an email is actually ham (0), so we used np.mean(Y_train == 0) to get the accuracy, which was about 0.745. • For zero_predictor_recall, recall measures how many spam emails were correctly identified. Since the zero predictor never predicts spam at all, the number of true positives is 0, and the recall is 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 accuracy of the logistic regression classifier (my_model) is slightly higher than the accuracy of the zero predictor. Specifically: • The logistic regression classifier had a training accuracy of approximately 0.7576. • The zero predictor had a training accuracy of approximately 0.7447.

This means that the logistic regression model performs slightly better by correctly classifying more emails overall, especially because it can correctly identify some spam emails, whereas the zero predictor always predicts ham and misses all spam.

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 (my_model) may be performing poorly because the five word features we used (‘drug’, ‘bank’, ‘prescription’, ‘memo’, ‘private’) are not very common across the email set. Many emails, both spam and ham, might not contain any of these words at all. As a result, the model often has very little information to distinguish between spam and ham, leading to poor performance, low recall (missing a lot of spam), and overall limited predictive power.

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) for a spam filter instead of the zero predictor. Even though the logistic regression model does not have very high recall (only about 11%), it still correctly identifies some spam emails, while the zero predictor catches none at all (zero recall). Detecting at least some spam is critical for a spam filter, and the logistic regression model has a slightly higher accuracy and non-zero recall compared to the zero predictor.