Notebook

April 22, 2020

0.0.1 Question 1c

Discuss one thing you notice that is different between the two emails that might relate to the identification of spam.

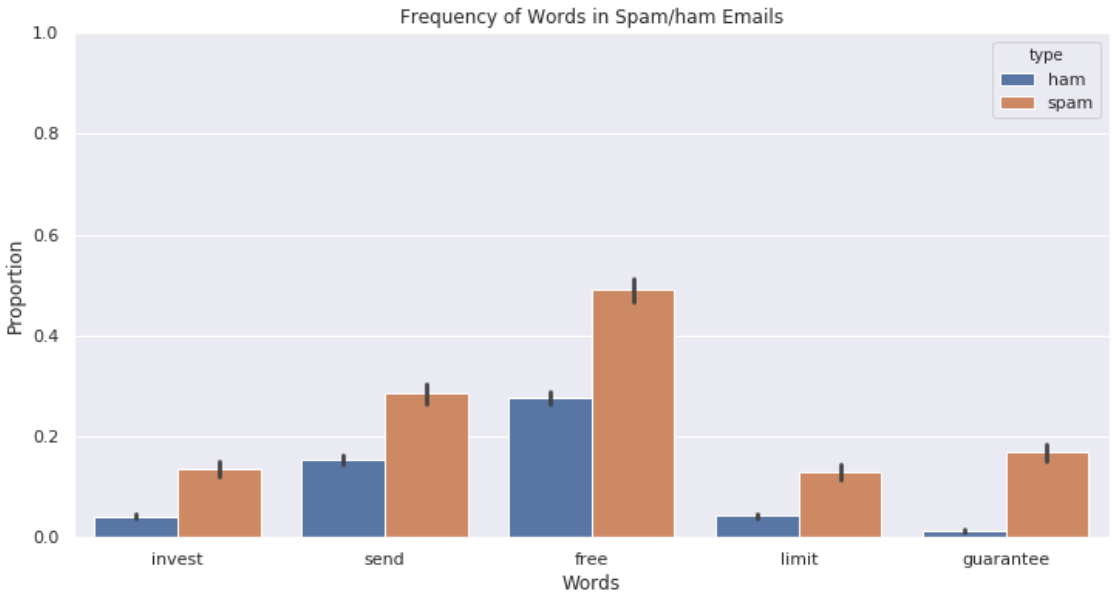
In the case of spam mails, the html format was often used, and specific words tend to appear like ‘free’, ‘money’.

0.0.2 Question 3a

Create a bar chart like the one above comparing the proportion of spam and ham emails containing certain words. Choose a set of words that are different from the ones above, but also have different proportions for the two classes. Make sure to only consider emails from `train`.

In [12]:
```
train=train.reset_index(drop=True) # We must do this in order to preserve the ordering of emai
words_3a = ['invest', 'send', 'free', 'limit', 'guarantee'];
list_email = train['email'].tolist();
list_spam = train['spam'].tolist();
each_contain = words_in_texts(words_3a, list_email);
word_1 = [x[0] for x in each_contain];
word_2 = [x[1] for x in each_contain];
word_3 = [x[2] for x in each_contain];
word_4 = [x[3] for x in each_contain];
word_5 = [x[4] for x in each_contain];
list_spam = ['spam' if x == 1 else x for x in list_spam];
list_spam = ['ham' if x == 0 else x for x in list_spam];
new_df = pd.DataFrame({
    'invest': word_1,
    'send': word_2,
    'free': word_3,
    'limit': word_4,
    'guarantee': word_5,
    'type': list_spam
});
new_df = new_df.melt("type");
sns.set(rc={'figure.figsize':(12,6)});
graph = sns.barplot(x = 'variable', y = 'value', hue = 'type', data = new_df);
graph.set_ylim(0,1);
plt.xlabel("Words");
plt.ylabel("Proportion");
plt.title("Frequency of Words in Spam/ham Emails");
plt.show
```

Out[12]:
```
<function matplotlib.pyplot.show(*args, **kw)>
```



0.0.3 Question 3b

Create a *class conditional density plot* like the one above (using `sns.distplot`), comparing the distribution of the length of spam emails to the distribution of the length of ham emails in the training set. Set the x-axis limit from 0 to 50000.

```
In [13]: list_email = train['email'].tolist();
         list_spam = train['spam'].tolist();
         list_length_email = [len(x) for x in list_email];
         new_df2 = pd.DataFrame({
             'length': list_length_email,
             'type': list_spam
         });
         spam_length = new_df2[new_df2["type"] == 1]["length"];
         ham_length = new_df2[new_df2["type"] == 0]["length"];
         sns_plot = sns.distplot(ham_length, label = "Ham", hist = False);
         sns_plot = sns.distplot(spam_length, label = "Spam", hist = False);
         sns_plot.set_xlim(0, 50000);
         plt.xlabel("Lengh of email body");
         plt.ylabel("Distribution")
         plt.show()
```