0.1 Question 1a

Generate your visualization in the cell below.

In [12]:
```
train
```

Out[12]:

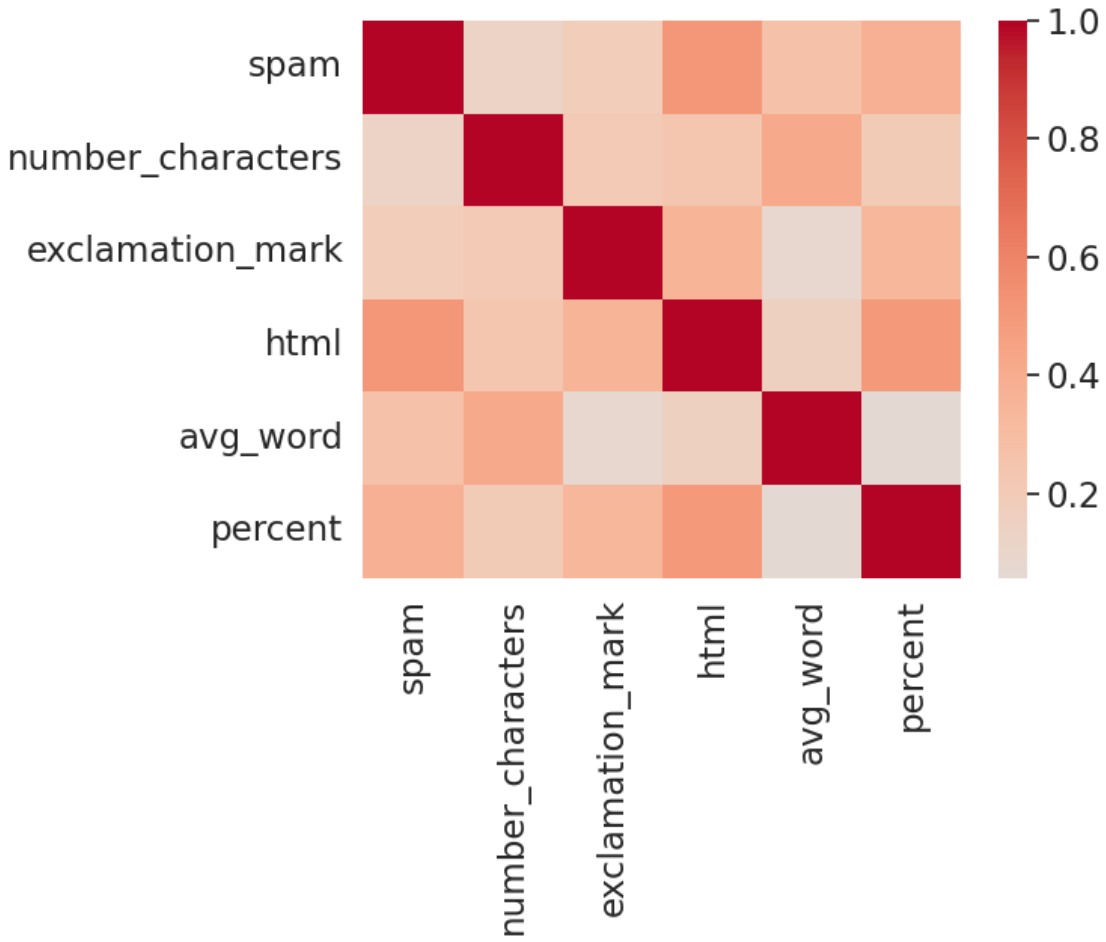
	id	subject	email	spam
0	7657	Subject: Patch to enable/disable log\n	while i was playing with the past issues, it a…	0
1	6911	Subject: When an engineer flaps his wings\n	url: http://diveintomark.org/archives/2002/10/…	0
2	6074	Subject: Re: [Razor-users] razor plugins for m…	no, please post a link!\n \n fox\n ----- origi…	0
3	4376	Subject: NYTimes.com Article: Stop Those Press…	this article from nytimes.com \n has been sent…	0
4	5766	Subject: What's facing FBI's new CIO? (Tech Up…	<html>\n <head>\n <title>tech update today</ti…	0
…	…	…	…	…
7508	5734	Subject: [Spambayes] understanding high false …	>>>>> "tp" == tim peters <tim.one@comcast.net>…	0
7509	5191	Subject: Reach millions on the internet!!\n	\n dear consumers, increase your business sale…	1
7510	5390	Subject: Facts about sex.\n	\n forwarded-by: flower\n \n did you know that…	0
7511	860	Subject: Re: Zoot apt/openssh & new DVD playin…	on tue, oct 08, 2002 at 04:36:13pm +0200, matt…	0
7512	7270	Subject: Re: Internet radio - example from a c…	chris haun wrote:\n > \n > we would need someo…	0

[7513 rows x 4 columns]

In [13]:
```
train['email_len'] = train['email'].str.len()
train['number_characters'] = train['email'].apply(lambda x: sum(i.isdigit() for i in x))
train['exclamation_mark'] = train['email'].str.count('!')
train['html'] = train['email'].str.contains('<html', case=False).astype(int)
train['reply_chain_depth'] = train['email'].str.count('>')
train['exclam_per_char'] = train['exclamation_mark'] / train['email_len'] * 1000
train['avg_word'] = train['email'].apply(lambda x: np.mean([len(i) for i in x.split()]) if x.sp
train['percent'] = train['email'].str.contains('%').astype(int)
features = ['spam','number_characters', 'exclamation_mark', 'html', 'avg_word', 'percent']
correlation = train[features].corr()
```

```
correlation
sns.heatmap(correlation, cmap = 'coolwarm', center = 0)
```

Out[13]: <Axes: >



0.2 Question 1b

In two to three sentences, describe what you plotted and its implications with respect to your features.

Above, I plotted a heatmap to identify how features relate with spam emails. Features that appeared darker on the spam row, are more likely to contribute to distinguishing between spam and non-spam emails. On my heat map, the highest contributor was if an html was present.

1 Question 4

Describe the process of improving your model. You should use at least 2-3 sentences each to address the following questions:

1. How did you find better features for your model?
2. What did you try that worked or didn't work?
3. What was surprising in your search for good features?

To find better features in my model, I would look at how high the correlation was to spam. What worked well was using features that were normalized because it reduced bias. In my search for good features, I found it surprising that an email with words that had a lot of capital letters was not a huge indicator of whether emails were spam or not.

2 Question 5: ROC Curve

In most cases, we won't be able to get 0 false positives and 0 false negatives, so we have to compromise. For example, in the case of cancer screenings, false negatives are comparatively worse than false positives — a false negative means that a patient might not discover that they have cancer until it's too late. In contrast, a patient can receive another screening for a false positive.

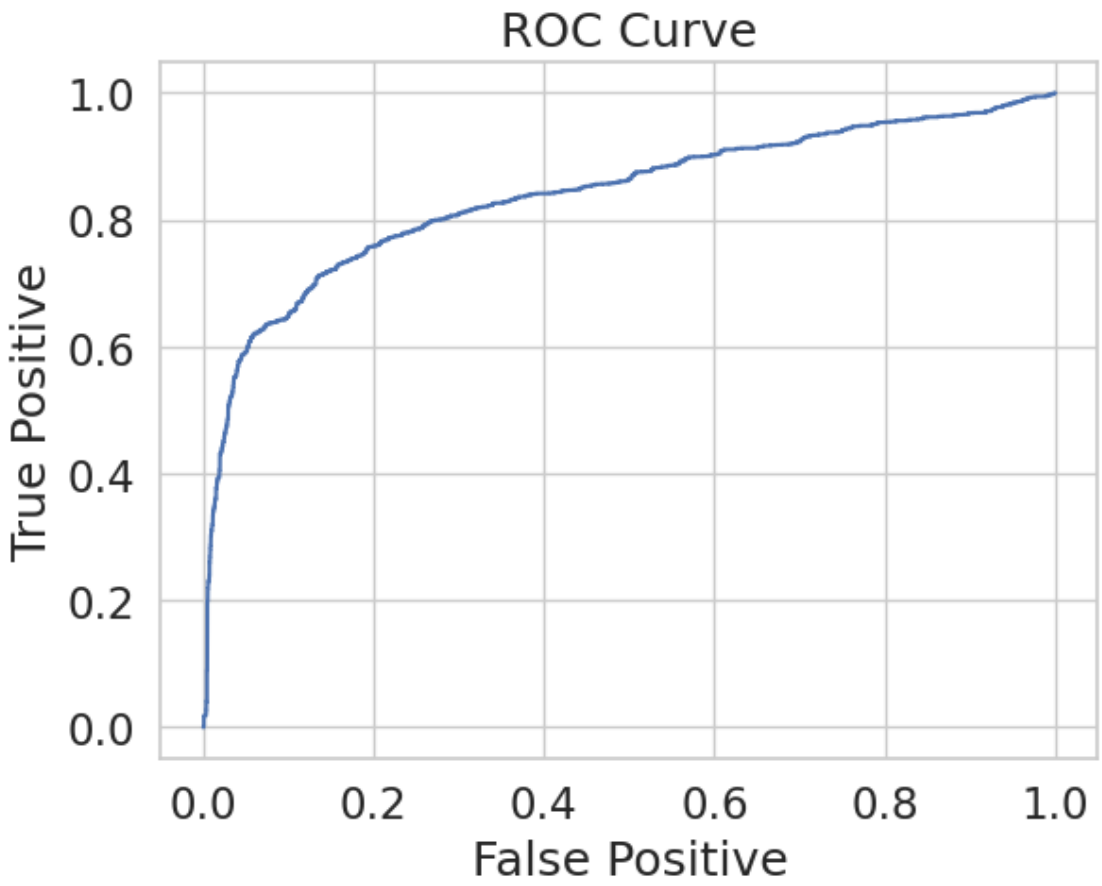
Recall that logistic regression calculates the probability that an example belongs to a particular class. To classify an example, we say that an email is spam if our classifier gives it ≥ 0.5 probability of being spam. However, **we can adjust that cutoff threshold**. We can say that an email is spam only if our classifier gives it ≥ 0.7 probability of being spam, for example. This is how we can trade off false positives and false negatives.

The Receiver Operating Characteristic (ROC) curve shows this trade-off for each possible cutoff probability. In the cell below, plot an ROC curve for your final classifier (the one you use to make predictions for Gradescope) on the training data. Lecture 23 may be helpful.

Hint: You'll want to use the `.predict_proba` method (documentation) for your classifier instead of `.predict` to get probabilities instead of binary predictions.

```
In [22]: prob = model.predict_proba(X_train)[:, 1]
         true_val = train['spam']
         f, t, thresh = roc_curve(true_val, prob)
         plt.plot(f, t)
         plt.xlabel('False Positive')
         plt.ylabel('True Positive')
         plt.title('ROC Curve')
```

```
Out[22]: Text(0.5, 1.0, 'ROC Curve')
```


ROC Curve
True Positive
False Positive
0.0
0.2
0.4
0.6
0.8
1.0

2.0.1 Question 6a

Pick at least **one** of the emails provided above to comment on. How would you classify the email (e.g., spam or ham), and does this align with the classification provided in the training data? What could be a reason someone would disagree with *your* classification of the email? In 2-3 sentences, explain your perspective and potential reasons for disagreement.

Based on example 3, I would classify the email as spam because the email is marketing telecoms and provides a link to it's website. Although, in the training data, it is classified as ham. Someone may disagree with my classifaction of the email if they believe that the company marketing the telecoms is a reliable source.

2.0.2 Question 6b

As data scientists, we sometimes take the data to be a fixed "ground truth," establishing the "correct" classification of emails. However, as you might have seen above, some emails can be ambiguous; people may disagree about whether an email is actually spam or ham. How does the ambiguity in our labeled data (spam or ham) affect our understanding of the model's predictions and the way we measure/evaluate our model's performance?

The ambiguity in our labeled data affects our understanding of the model's predictions and the way we measure our model's performance because emails deemed spam by one person could be important or helpful information to another. The model relies on a persons opinions on what to classify as spam vs. ham which can include personal bias. So, the "ground truth" although correct in one situation could be incorrect in another.

Part ii Please provide below the index of the email that you flipped classes (`email_idx`). Additionally, in 2-3 sentences, explain why you think the feature you chose to remove changed how your email was classified.

The email idx is 7507. When removing bank it changed the initial classification to spam because it had strong correlation to spam emails. Removing bank reduced the model's confidence that an email was spam, explaining why it is classified as ham now.

Part i In this context, do you think you could easily find a feature that could change an email's classification as you did in part a)? Why or why not?

Working with a larger model would make it difficult to easily find a feature that could change an email's classification because the influence of a single feature is not as strong. So, if one feature was to be removed from the model, the classification would not flip instantly.

Part ii Would you expect this new model to be more or less interpretable than `simple_model`?

Note: A model is considered interpretable if you can easily understand the reasoning behind its predictions and classifications. For example, the model we saw in part a), `simple_model`, is considered interpretable as we can identify which features contribute to an email's classification.

This new model would be less interpretable than simple_model because there are more features within the model. Since there are 1000 features within the model, it would not be as easy to interpret which features contribute to an email's classification because it is harder to understand how each feature influences the prediction.

2.0.3 Question 7c

Now, imagine you're a data scientist at Meta, developing a text classification model to decide whether to remove certain posts / comments on Facebook. In particular, you're primarily working on moderating the following categories of content: * Hate speech * Misinformation * Violence and incitement

Pick one of these types of content to focus on (or if you have another type you'd like to focus on, feel free to comment on that!). What content would fall under the category you've chosen? Refer to Facebook's Community Standards, which outline what is and isn't allowed on Facebook.

The content that would fall under hate speech is information about someones race, religion, gender, nationality, e.t.c. Post's that negatively target these identities are against Facebook's Community Standards. For example, posts that include racials slurs or ones that promote violence against certain groups of people are against Facebook's guidelines. Facebook states that hate speech is not allowed because people are more willing to use their voice when they are not attacked on the basis of who they are.

2.0.4 Question 7d

What are the stakes of misclassifying a post in the context of a social media platform? Comment on what a false positive and false negative means for the category of content you've chosen (hate speech, misinformation, or violence and incitement).

Misclassifying a post in the context of a social media platform can impact users and a platform's reputation. If a post is false positive then it means a post was incorrectly flagged as hate speech. This could unfairly have a post removed and silence those who spoke out. If a post is false negative then it means a post was incorrectly deemed non-hate speech. This allows people to target groups of people which could prevent them from speaking out. This will promote cyber bullying.

2.0.5 Question 7e

As a data scientist, why might having an interpretable model be useful when moderating content online?

Having an interpretable model is useful when moderating content online because it will be easier to find out why a post was flagged. If a model is wrongly taken down, having an interpretable model makes it easier to find out why the post was flagged by looking at certain words or phrases that triggered the model's classification. Overall, a interpretable model will help us understand what features contribute to a models decision making and it will make it easier to refine the model when false positives occur.