Now it's time to construct and preprocess the corpus.

So we'll start by loading the tm package with library(tm).

And now that we'll have done that, we'll construct a variable called corpus using the corpus and vector source functions and passing in all the emails in our data set, which is emails$email.

So now that we've constructed the corpus, we can output the first email in the corpus.

We'll start out by calling the strwrap function to get it on multiple lines, and then we can select the first element in the corpus using the double square bracket notation and selecting element 1.

And we can see that this is exactly the same email that we saw originally, the email about the working paper.

So now we're ready to preprocess the corpus using the tm map function.

So first, we'll convert the corpus to lowercase using tm map and the two lower function.

So we'll have corpus = tm_map(corpus, tolower).

And then we'll do the exact same thing except removing punctuation, so we'll have corpus = tm_map(corpus, removePunctuation).

We'll remove the stop words with remove words function and we'll pass along the stop words of the English language as the words we want to remove.

And lastly, we're going to stem the document.

So corpus = tm_map(corpus, stemDocument).

And now that we've gone through those four preprocessing steps, we can take a second look at the first email in the corpus.

So again, call strwrap(corpusstrwrap(corpus{[1]).

And now it looks quite a bit different.

We can come up to the top here.

It's a lot harder to read now that we removed all the stop words and punctuation and word stems, but now the emails in this corpus are ready for our machine learning algorithms.