It can be useful to be able to classify new "test" documents using already classified "training" documents. A common example is using a corpus of labeled spam and ham (non-spam) e-mails to predict whether or not a new document is spam.

For this project, you can start with a spam/ham dataset, then predict the class of new documents (either withheld from the training dataset or from another source such as your own spam folder).   One example corpus: **https://spamassassin.apache.org/old/publiccorpus/ (Links to an external site.)**

You may work alone or in a group on this project. You're welcome to use any tools or approach that you like. Due before our next meetup.