Week 10, Project4, Description

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/>

Here are two short videos that you may find helpful.

The first video shows how to unzip the provided files.

<https://youtu.be/6IzhRaSePKU>

The second video provides a short overview of predictive classifiers.

<https://youtu.be/5ikDo4SrLNQ>

For more adventurous students, you are welcome (encouraged!) to come up with a different set of documents (including scraped web pages!?) that have already been classified (e.g. tagged), then analyze these documents to predict how new documents should be classified.

This assignment is due end of day on Sunday,  November 4.  You may work in a small team if you want.  We will look at some of your solutions in our meetup on Wednesday, November 7.