# Project 4 - Document Classification (90 points)

## Collaboration (extra credit)

Use GitHub as a group, share code and project documentation.

## Data Collection (5 points)

Use a corpus of labeled spam and ham (non-spam) e-mails

## Data Storage (10 points)

Manually unzip the data (5 points)

Automatically unzip the data (5 points)

## Project Code (70 points)

Predict the class of new documents withheld from the example corpus. (40 points) or

Come up with a different set of documents (including scraped web pages!?) (60 points)

Use the dictionary of common words (10 points)

Separate the message header from the message body (5 points)

Analyze these documents to predict how new documents should be classified (algorithm)(10 points)

## Presentation (10 points)

## Extra Credit (1 point each)

Try out statistics and data models

Start early, ask many questions, actively post on the provided discussion forum, etc.