Datasets used in CASI

Spam data Table 8.3, and Chapter 16

Email spam data. 4601 email messages sent to "George" at HP-Labs.
He labeled 1813 of these as *spam*, with the remainder being good email (*ham*).

The goal is to build a customized spam filter for George.

The feature set tracks 57 of the most commonly used, non-trivial words in the corpus, using a bag-of-words model.
Recorded for each email message is the relative frequency of each of these words and tokens.
Included as well are three different recordings of capitalized letters.

These are a publicly available database, available from the UC Irvine data repository:
archive.ics.uci.edu/ml/datasets/Spambase More details about the data can be found there.

Our data matrix has 59 columns:
spam Logical variable, TRUE is spam, FALSE is ham (good email).
testid Logical variable. An optional split into train (FALSE) and test (TRUE) data (as used in, for example "Elements of Statistical Learning").
The remainder of the columns are features used to build a prediction model.

SPAM.csv