HandClassifiedCorpora

Like a Bayesian learning system, SpamAssassin's GeneticAlgorithm requires a corpus of hand-classified mail. Our guidelines are (quoting and expanding on "masses/CORPUS_POLICY"):

- hand-verified as "spam" and "ham" (non-spam) piles – *not* just classified using existing spam-classification algorithms, such as SpamAssassin itself. Note that it's fine to use SpamAssassin to pre-filter them into the right piles, just make sure you scan the results "by hand" afterwards to verify that SpamAssassin made the correct diagnosis in each case.
- containing a representative mix of ham mail – that includes commercial-sounding-but-not-spam messages, legitimate business discussions (which may include talk of "sales", "marketing", "offers", bankruptcies, mortgages, etc), or verified opt-in mail newsletters. This is a *very* important point! Your ham corpus should contain as much ham as is possible, as close to ALL valid emails received by everybody as is possible, with only the exceptions noted here. ("as is possible" recognizes that for privacy and confidentiality reasons some ham cannot be stored anywhere but its destination email folder.)
- containing no old spam mail. Older spam uses different tricks and terminology, which will impact SpamAssassin's accuracy when it's filtering "live", new mail. Please try not to scan spam older than 6 months. For this purpose it may be useful to categorize your spam by month, and to regularly delete those files with the older spam.
- containing a representative mix of spam mail. If you bounce high-scoring spam, or have collections of only user submissions of missed spams or false positive hams, this will unbalance the corpus; it's better to scan collections of *all* spam received at a set of email accounts, instead of a subset.
- cleaned of viruses, bounce mails from broken virus and spam filters, and forwarded spam messages. These will skew the results.
- and finally, cleaned of discussion of spam or virus messages or signatures (such as SpamAssassin-talk or bugtraq mailing list messages). Even though they are ham, these often contain snippets of code that incorrectly trigger tests, and again will skew the results. (Rewriting the tests to avoid triggering on SpamAssassin-talk messages is not realistic, unfortunately.)
- *(if you're mass-checking for a RescoreMassCheck* the corpora must contain *both* ham and spam. If it contains only one, the accuracy figures reported for the Bayes rules will be invalid.

Once you run MassCheck, see the instructions in CorpusCleaning for details of how to verify that the top scorers are not accidental spam that got through.

(Aside: yes, it's "corpora". See PluralOfCorpus)

Mail to NOT Include in Ham or Spam

- Mailing lists
  Do not include discussion mailing lists in either your ham or spam corpus. Mailing lists tend to be too similar in content, and all mail is sent by the same mail server. Furthermore it can unduly bias the results if multiple masscheck participants are subscribed to the same mailing lists. Try to stick to mail directly to you. Generally low-traffic announce-only mailing lists are OK.
- Spam Sent via Legitimate Services (Facebook, Livejournal, etc.)
  Occasionally you receive spam text posted to your account on services like LiveJournal or Facebook. DO NOT include such mail in either ham or spam folder. Just delete it. Why? We don't want to count these as spam, causing false marks against highly safe whitelist rules like USER_IN_DEF_DKIM_WL. They do not count as ham either, because spam URL's or spam text would throw off the statistics if they show up in the ham folder. Simply delete them.

Minor things that are nice to have

- eliminate duplicates – there should be one and only one copy of any single email, whether spam or ham. (This isn't a hard and fast rule, as it can be very time-consuming. Just remove dups where they all arrive at the same time, in sequence, if possible, but don't really worry about it too much.)