SoughtRules

The "sought" ruleset

Our spamtrap network collects multiple hundreds of megabytes of spam per day. Wouldn't it be great if there was a way to feed that directly into a script to automatically extract rules?

This is now possible, and the results are the "sought.cf" ruleset – an automatically-generated ruleset which seeks good rules directly from the SpamAssassin spamtraps, updated every 4 hours.

Update: this is no longer active, and should not be used.

Gory Details

If you're curious, here is a technical explanation of the algorithm used, and here is an examination of their efficiency against our test corpora. Here are instructions on how it was used.