RescoreSet01Details

Rescore Mass-checks for Set 0 and Set 1

The set 0 and 1 mass-check runs for 3.0.0 are now finished! This page is obsolete

Here's the procedure you'll need to follow, if you wish to submit data for the rescoring run next time around:

First, send mail to <submit.at.spamassassin.org>, and ask for a rescore-submission account if you haven't already got one.

It's helpful, but not required, to have some or all of the helper applications installed:

- the Mail::SPF::Query module
- the Net::DNS module
- Razor
- DCC
- Pyzor

If you're running nightly mass-checks, please feel free to disable them when running the rescore mass-check runs. Also, please note that the nightly submission accounts will work for rescore submissions as well.

Then run these commands:

```
wget http://spamassassin.apache.org/released/Mail-SpamAssassin-3.0.0-pre2.tar.gz
tar xvfz Mail-SpamAssassin-3.0.0-pre2.tar.gz
cd Mail-SpamAssassin-3.0.0
perl Makefile.PL < /dev/null; make

cd masses
mkdir spamassassin
rm spamassassin/bayes*
echo "use_bayes 0" > spamassassin/user_prefs
echo "use_auto_whitelist 0" >> spamassassin/user_prefs
rm ham.log spam.log

./mass-check --net -j 4 --restart=400 --after=1041397200 --all <targets>
```

`<targets>` is the list of directories, mboxes, etc., like
`spam:dir:~/Mail/spam`. See the comments at the top of "mass-check" for details.

This takes *ages* to run. `-j 4` controls the number of processes to use; 4 should be OK for a single-processor machine, since most of the time they'll be waiting for network results to arrive. If you have adequate RAM and don't mind the load, you can use `-j 6` or `-j 8`. There's not much benefit in going higher than `-j 8`.

The `--after=1041397200` option tells mass-check to ignore messages older than 18 months ago (in this case January 1 2003). This is useful if your corpus has older messages intermingled with your newer messages.

If you have an unusual network layout, you may need to specify
`trusted_networks` and/or `internal_networks` in the `spamassassin/user_prefs` file. But SA should be able to infer it in most cases. If you get less than a 10% or 15% spam hit rate for RCVD_IN_XBL, then you might need to use these configuration parameters.

Once it finishes:

```
USER="[whatever your username is]"
RSYNC_PASSWORD="[whatever your password is]"
export RSYNC_PASSWORD

rsync -CPcvuzb ham.log $USER@rsync.spamassassin.org::submit/ham-nobayes-net-$USER.log
rsync -CPcvuzb spam.log $USER@rsync.spamassassin.org::submit/spam-nobayes-net-$USER.log
```

That's it! Then we do the bayes+nonet and bayes+net runs later on.

The results for this run will need to be in by Monday July 19th. If you're still running then, submit what you have so far and beg for more time. 😉