CorpusCleaning

Cleaning a Mail Corpus

Here's a few methods used to deal with common forms of corpus pollution – messages in a mail corpus that aren't suitable for use in a MassCheck.

What A Corpus Needs To Look Like

SpamAssassin relies on corpus data to generate optimal scores. This is the policy used by all corpora accepted by the SpamAssassin project (moved here from 'masses/CORPUS_POLICY'):

- **Hand-verified**: all mail must be hand-verified into "spam" and "ham" (non-spam) collections, by its recipient. It may not be solely classified using automated spam-classification algorithms such as SpamAssassin and other spam filters; we need the human decision (although it may be aided by SpamAssassin, of course). Also, we can't use data that's been collected from third-party accounts, since we don't know what the recipient may have signed up for.
- **Reliable source**: Ensure that the mails were classified by a trustworthy source; mails marked as spam by users at your ISP, for example, are not reliable enough for use as a SpamAssassin corpus. (It's pretty well-known that some users will use the "report as spam" button instead of unsubscribing from legit newsletters.)
- **No old spam**: please try to avoid including spam older than 2 months, and ham older than 6 years. (We'll filter it out on the server-side anyway.)
- **Representative mix**: each side of the corpus must contain a representative mix of mail of that type. For ham, that includes commercial ham messages, legitimate business discussions, and verified opt-in mail newsletters. A ham corpus consisting of nothing but reported false-positives will produce bad data (especially for Bayes) – and the same applies for spam. If at all possible, try to include as much of the day-to-day spam you receive, including the "easy stuff".
- **No viruses**: SpamAssassin is a spam filter, not a virus scanner – remove viruses from your corpora. (Phishes, however, fall under "spam", even though some virus scanners mark them as viruses.)
- **No faked bounces**: bounces of viruses or spam sent back to forged or faked from addresses, (so-called blowback or joe-job bounces), these typically have an envelope sender of <> or <MAILER-DAEMON.*>, but please include any valid bounces if you can.
- **No moderation mails**: mailing list moderation administrative messages that contain spam subject lines or excerpts.
- **No spam discussion**: anti-spam or anti-virus mailing lists, especially SpamAssassin, that frequently include spam and virus elements, even though they are technically ham, these often appear to be spam and will skew the results, rewriting the tests to avoid triggering on these messages is not realistic at this time.

Cleaning Out False Positives

To clean a spam corpus of FalsePositives – first, do a mass-check. You will wind up with a 'spam.log' and 'ham.log' file. Run these commands to get a list of the 200 lowest-scoring spams, create a mbox file with just those messages, then open that mbox up in the "mutt" mail client:

```
cd /path/to/your/spamassassin/masses
sort -n -k 2 spam.log | head -200 > id.low
./mboxget < id.low > mbox
mutt -f mbox
```

(you could use another mail client if you want, it's just a std UNIX-format mbox file.)

Now, delete all messages that **really are** spams, and not false positives (or bounces, or virus blowback, or other kinds of undesirable messages). Quit and save the mbox. It now contains only the 'bad' messages.

You can then take that mbox file, grep out the original MassCheck message id strings, and remove those lines from the 'spam.log' file:

```
grep -a X-Mass-Check-Id mbox | sed -e 's/^X-Mass-Check-Id: //' > id.fps
./remove-ids-from-mclog id.fps < spam.log > spam.log.new
mv spam.log.new spam.log
```

You can also remove the offending files, or messages from the source mailboxes, directly. (This is advisable as you'll probably wind up mass-checking them again at some point.) However, this depends on what format you use to store messages; Maildirs, mboxes, etc. etc. (Maildirs are easiest, since you can just delete the files named in the 'id.fps' file.)

Rules that are useful for spotting FPs in the spam corpus:

- ALL_TRUSTED: once a mass-check completes, it's worth grepping the spam.log for ALL_TRUSTED and checking what mails it hits.
- NO_RELAYS: spam which has no relaying Received headers, and therefore did not traverse the internet, is probably not spam.

See also 'Corrupt Messages' below for other stuff to clear out.

Here's a command line to grep a log for a rule name, and generate an mbox of the results, then open it in "mutt":

```
grep 'ALL_TRUSTED' ham.log > grepped.log
./mboxget < grepped.log > mbox
mutt -f mbox
```

Cleaning Out False Negatives

Doing the same operation to clean the ham corpus of FalseNegatives is similar, but reverses a few things... here's the commands to do that:

```
cd /path/to/your/spamassassin/masses
sort -rn -k 2 ham.log | head -200 > id.hi
./mboxget < id.hi > mbox
mutt -f mbox
```

Delete the messages that are good, usable ham, leaving only spams, hams that include bits of spam, virus blowback, bounces, or whatever other undesirable messages you want to get rid of. Quit and save.

```
grep -a X-Mass-Check-Id mbox | sed -e 's/^X-Mass-Check-Id: //' > id.fns
./remove-ids-from-mclog id.fns < ham.log > ham.log.new
mv ham.log.new ham.log
```

Repeat, if necessary...

Rules that are useful for spotting FNs (or spam discussions!) in the ham corpus:

- BAYES_99: once a mass-check completes, it's worth grepping the ham.log for BAYES_99 and checking what mails it hits.
- any of the other top-listed rules in the HitFrequencies report, especially network tests such as the SURBL rules

See also 'Corrupt Messages' below for other stuff to clear out.

Saving a list of verified non-spams

To make corpus cleaning easier next time, you can save a list of emails that scored high that weren't spam, to automatically skip. When viewing emails as above, they have a "X-Mass-Check-Id:" header which lists the file they came from, which you can use to remove any email that was actually spam from the id.hi file. Then grab the file names out of id.hi with awk, and make a backup copy:

```
awk {'print $3'} < id.hi > ~/sa/id.hi.good
```

Next time, run:

```
sort -rn -k 2 ham.log | fgrep -vf ~/sa/id.hi.good | head -n 200 > id.hi
./mboxget < id.hi > mbox
mutt -f mbox
```

Corrupt Messages

Occasionally, these will crop up – some MUAs have a tendency to mess up mail messages or folders, making them unsuitable for use with MassCheck. SpamAssassin includes a few rules that can help identify corrupt messages.

- MISSING_HEADERS: if a message doesn't have all the normal headers, such as From, To, and Subject, this will fire. Be sure to hand-verify any ham and spam messages that hit this to ensure that they're formatted correctly (in RFC-2822 format).
- NO_HEADERS_MESSAGE (or a combo of MISSING_HEADERS,MISSING_DATE,MISSING_SUBJECT in versions < 3.2.0): generally means you've got message without most of the important RFC-822 headers (often errors generated by MUAs/MDAs).
- EMPTY_MESSAGE: generally zero-length files, esp if accompanied by NO_RECEIVED.
- MISSING_HB_SEP: This is another danger sign, typically indicating that a header line has had a newline inserted incorrectly somehow, or an mbox "From" line has been inserted between RFC-822 headers.
- ANY_BOUNCE_MESSAGE: this indicates that the mail was a bounce message, a C/R challenge, or a "virus warning" from a broken scanner. These should be removed from both the ham and spam corpora, in general.

Other Corpus Cleaning Methods

DSPAM

DSPAM is well known standalone bayesian tool, you can crosscheck your corpus fast and easy with it.

It doesn't seem to be maintained anymore, here is probably the best version: https://github.com/ensc/dspam (download the master). If you are not comfortable compiling things, then you need to find some package.

Example how to build and install it simply in your home directory:

```
unzip master.zip && cd dspam-master
# autoconf/automake/gcc stuff obviously needed
./autogen.sh
./configure --prefix=$HOME/dspam --disable-trusted-user-security --disable-syslog
make && make install
```

This assumes your corpus is in Maildir format (file per message).

Make sure PATH includes $HOME/dspam/bin if installed there.

You can experiment with different learning methods. It's probably best to feed all manually verified messages first with --source=corpus. It's not an exact science, so mixing methods might come up with different FPs/FNs.

Learn the corpus (method 1):

```
# Always clear old data first
rm -rf $HOME/dspam/var
# This will learn the folders with --source=error
dspam_train $LOGNAME /path/to/spam /path/to/ham
```

Learn the corpus (method 2):

```
# Clear old data unless you are learning some additional corpus
rm -rf $HOME/dspam/var
# Feed your folders with --source=corpus
find /path/to/spam -type f | while read -r f; do
  dspam --user $LOGNAME --source=corpus --class=spam < "$f"
done
find /path/to/ham -type f | while read -r f; do
  dspam --user $LOGNAME --source=corpus --class=innocent < "$f"
done
```

Check the corpus:

```
/bin/bash
find /path/to/spam -type f | while read -r f; do
  RESULT=$(dspam --user $LOGNAME --classify < "$f")
  # Tune confidence >= 0.6 check if needed
  if [[ "$RESULT" =~ (result=\"Innocent\".*confidence=(1|0\.[6-9].)) ]]; then
    echo "$f ${BASH_REMATCH[1]}"
  fi
done
find /path/to/ham -type f | while read -r f; do
  RESULT=$(dspam --user $LOGNAME --classify < "$f")
  # Tune confidence >= 0.6 check if needed
  if [[ "$RESULT" =~ (result=\"Spam\".*confidence=(1|0\.[6-9].)) ]]; then
    echo "$f ${BASH_REMATCH[1]}"
  fi
done
```

It will output list of messages to check. Move to correct folder if indeed in wrong place.

```
/path/to/spam/message123 result="Innocent"; class="Innocent"; probability=0.0000; confidence=0.73
/path/to/ham/message234 result="Spam"; class="Spam"; probability=0.0005; confidence=0.61
```

If you move stuff around a lot, do a new learn and check.

If it keeps reporting some messages wrong, you can script some whitelist method to ignore certain files etc.