SocNightlyMassCheck

Nightly Mass-Check for Normal People

Nightly mass check is how we check rules in the SandBox and adjust to the ever changing behavior of spammers. But there are many problems in the current process which limits its current effectiveness.

Problems

- 1) It is only really usable on Linux/Unix hosts, or maybe Cygwin with a bunch of effort, although not easily with any Windows clients. Even if you are a Linux user, it is *TOO HARD* to setup and get it to work properly.
- 2) As a result, most if not all people sorting corpora and participating in nightly mass checks and corpus scoring are somehow related to a single demographic: computer hackers.
- 3) We don't have participants who are representative of normal users. These are people that use computers for reasons other than their work or hobby.
- 4) We don't have various languages represented in the corpus, especially Asian languages.
- 5) Normal users are less likely to understand the sorting rules and require training. Maybe their results would be less trusted.

Proposal

We need a way to make nightly mass check easily accessible to normal users. They need easy to use software to do mass checks and submit results. They must be properly trained on the sorting rules. Our project then needs some way of tracking the level of trust of these growing number of submitters.

I envision that generally hackers that care about spamassassin will urge their non-hacker friends to use this software as part of their daily e-mail. It is easy to convince people about the social benefit and how you can volunteer some time to help the rest of the world.

I think it would be sufficient to have a few dozen participants from different demographics, regions and languages in order to improve spamassassin. After this more accessible mass check software and supporting project infrastructure is ready, we could do a call for volunteers where our community can go out and find people in these varying demographics to participate. Our existing community of hackers and sysadmins can train individuals in corpus sorting and get them started.

You may think "this is crazy, why Windows?" The reason is this system *MUST* be easily accessible to normal users.

jm: I am still not convinced on this particular point. Note that a lot of "normal" users run MacOS X!

Requirements

- 1) MUST be able to run on Windows, where most normal users are.
- 2) MUST be able to read local Outlook, Outlook Express, and Thunderbird mail folders for ham and spam.
- 3) MUST be able to submit results to the spamassassin project.
- 4) MUST be very easy to setup and use, point and click. No editing of text files.
- 5) MUST document an easy to understand guide for normal users to learn the corpus sorting rules.
- 6) MUST devise some kind of improved accounting system that allows different levels of trust in submitted nightly mass check logs.

Possible Implementation Details

- 1) Implementing this really wouldn't be all that hard because you would use existing components like perl, spamassassin, ssh, and rsync.
- 2) You would need to tie them all together using some toolkit to make a frontend like gtk or qt that works in Cygwin. From the frontend you choose mail client folders to read for HAM and SPAM, and schedule the nightly mass check time.
- 3) You might need some kind of service or applet visible from the systray in order to do the scheduled nightly mass check.
- 4) Use an "InstallShield" type click-thru installer, which seems to be the standard on Windows. They shouldn't need to make *ANY* choices to download and configure Cygwin. It should just dump everything needed into a single folder that contains this mass check bundle. (use NSIS ?)
- 5) The use of existing FOSS components and a cross-platform toolkit like gtk or qt would allow this to build on Linux too.