RulesNotEnglish

Rules Project: Rules Not in English

(part of RulesProjectPlan)

Problem description: SA rules development handles rules aimed at spam in English best, since most SA rules developers that feed the distribution system speak and correspond in English, and the great majority of the testing corpora are based in English. We're not as good at developing, validating, testing, or scoring rules in other languages.

The Problem

- The primary language within the SA development team is English. It's the language they have skill in, as applied to writing rules.
- The primary language of spam and non-spam received by the SA development team is English. It's the language of the emails against which rules can be tested.
- The primary language of emails in mass-check corpora is English. It's the language within which we can develop reasonably reliable scores.

Because of this,

- Quite a few rules are developed which work very well indeed in English, but which cause false positives in other languages, because they haven't been adequately validated or measured against those other languages.
- Too few rules are developed in other languages.
- Even when rules are developed in other languages, such as the Chinese Rules found on CustomRulesets, we are unable within the development team to generate scores in which we have strong confidence.

Potential Solutions

- There are teams and individuals working on rules in other languages, again as shown by th4e Chinese Rules mentioned above. Part of the goal of this project should be to facilitate the implementation of those rules.

- SARE has experienced the English vs Other problem in many of our obfuscation rules, where they hit beautifully on English spam, but have horrible S/O rates for German ham (to pick an example).
  That's why we use 70_sare_name_eng.cf files, to indicate that these rules work well only on systems which expect almost 100% English ham, and little to no ham in other languages.
  I [BobMenschel] have begun to wonder whether it might be worth while having 50_scores.cf for English emails, and then 50_scores_de.cf for German emails, and have SA pick the score appropriately depending upon the language of the email, just as it picks rule descriptions based on the language of the host system.

- LorenWilton suggests: This is why I'd like to see a report-home option in SA that was enabled by default.

We could invent a class of rules that were 'test rules'. They would have nil score and wouldn't report on the mail summary if they hit. But they would show up in the report-home summary is to whether they hit, and whether it was ham or spam.

Then we can make rules that pass initial testing and stick them out for what we believe is good use, or maybe even for pure testing purposes. SA systems around the world would pick up these rules with sa-update, and would report home on the hit stats. If we have a good hitter that sucks in 'de', then we move it to an english-only ruleset, or we have an exclude-de option on the front of the rule or rule grouping. If the sysadmin has set his local language correctly, things should work out correctly.