Mail-Server: Token der Bayes-DB

Frage:

In der Bayes-DB von Spamassassin werden sogenannte "Token" gespeichert. Was genau sind diese Token? Können die Token auch in lesbarer Form angezeigt werden?

Antwort:

Der Bergriff "Token" der bayesischer Filter könnte man mit "Wort" übersetzten. Ist zwar nicht voll identisch, hilft aber der Vorstellung:

Eine Email besteht aus Wörtern.

Der bayesische Filter belegt in der Anlernphase jedes Wort mit einem Wahrscheinlichkeitswert.

In der Analysephase wird eine Email in seine Wörter zerlegt, der hinterlegte Wahrscheinlichkeitswert (wenn vorhanden) abgerufen und dann mit einer Formel auf einen Prozenzsatz zusammen gerechnet.

Dieser ermittelte Wert wird vom SpamAssassin als `BAYES_05`, `BAYES_20`, ..., `BAYES_99` dargestellt.

Die Theorie dahinter ist, dass typische Spam-Wörter wie z.B. "Viagra" in Spam-Mails häufiger auftauchen als in erwünschten Emails.

Und ändern wir "Wort" auf "Token", denn es geht nicht immer nur um Wörter. Sondern auch um Kombinationen von Zeichen (z.B. "`\/iagra`" mit Backlash und Slash als V), Aufrufezeichen und vorallem Wort-Kombinationen.

Anzeigen der Token

Für die Verwaltung der Tokens ist `sa-learn` zuständig. Das "learn" stammt daher, dass dieses Programm in erster Linie in der Anlernphase genutzt wird.

```
#liefert die Magic-Daten:
sa-learn --dump magic

#zeigt alle Tokens an:
sa-learn --dump data
```

Leider zeigt `sa-learn` ab SpamAssassin 3 hier lediglich nur noch (Low-Order-40bit-Sha1-) Hash-Codes an. Dies soll u.a. die Privatsphere der User schützen. Aber in erster Linie dient es einer besseren Performance der Datenbank.

Mail-Server: Token der Bayes-DB

Falls jemand dennoch weiß wie man die realen Tokens anzeigen lassen kann, wäre ich um einen Hinweis dankbar.

Weitere Links:

- Wikipedia: Bayesscher Filter

Eindeutige ID: #1380
huschi
2009-12-14 08:57