

UsingOverlap

Using 'overlap'

In the [SpamAssassin](#) "masses" directory, there's a tool called 'overlap', which is used to determine how much the rules in the ruleset overlap with each other.

For example, let's say I have a log file in *spam.log*, and want to examine how much the rules that start with `_T_DRUG_` overlap with each other. I run overlap like so:

```
./overlap spam.log > ov
pcregrep "\sT_DRUG.*,T_DRUG" ov | sort -r +1 -n
```

Which in this case produces this output:

87	1.000	0.690	T_DRUGS_SLEEP_EREC, T_DRUGS_SLEEP
87	1.000	0.084	T_DRUGS_SLEEP_EREC, T_DRUGS_ERECTILE
703	1.000	0.679	T_DRUGS_ERECTILE_OBFU, T_DRUGS_ERECTILE
328	1.000	0.715	T_DRUGS_ANXIETY_EREC, T_DRUGS_ANXIETY
328	1.000	0.317	T_DRUGS_ANXIETY_EREC, T_DRUGS_ERECTILE
315	1.000	0.887	T_DRUGS_DIET_EREC, T_DRUGS_DIET
315	1.000	0.304	T_DRUGS_DIET_EREC, T_DRUGS_ERECTILE
311	1.000	0.523	T_DRUGS_PAIN_EREC, T_DRUGS_PAIN
311	1.000	0.300	T_DRUGS_PAIN_EREC, T_DRUGS_ERECTILE
289	1.000	0.630	T_DRUGS_ANXIETY_OBFU, T_DRUGS_ANXIETY

Explanation of the columns: the first number is how many mails hit both rules; the second, how much of the hits for the first rule also hit the second; the third, how much of the hits for the second rule also hit the first.

So in the case of this line:

87	1.000	0.690	T_DRUGS_SLEEP_EREC, T_DRUGS_SLEEP
----	-------	-------	-----------------------------------

87 mails hit both rules; all of the mails that hit `T_DRUGS_SLEEP_EREC` also hit `T_DRUGS_SLEEP`; and 69% of the mails that hit `T_DRUGS_SLEEP` also hit `T_DRUGS_SLEEP_EREC`.

Overlap is very useful, if you believe that some rules are all hitting the same spam messages.