Spam filtering - a succesful solution (e.g. at FI)

Marek González

dSpam

- součást antispamové ochrany na FI
- statistický bayesovský filtr
- automatické, nepřetržité učení

Učení

- učení s učitelem
- pozitivní data (spamy) přes honeypot nebo na spam@fi.muni.cz
- negativní data (hamy) na notspam@fi.muni.cz

Tokenizace

metody:

1) word
2) chain
3) orthogonal sparse bigram
4) sparse binary polynomial hashing

word – co slovo to token

```
TOKEN: 'Heute'
TOKEN: 'Abend'
TOKEN: 'war'
TOKEN: 'ich'
[...]
```

chain – token 2 slova po sobě

```
TOKEN: 'Heute+Abend'
TOKEN: 'Abend+war'
TOKEN: 'war+ich'
TOKEN: 'ich+mit'
[...]
```

"Heute Abend war ich mit meiner Freundin im Kino und habe viel gelacht"

OSB – pro každé slovo +/- 4 sliding window

```
TOKEN: 'Heute+#+#+#+mit'
TOKEN: 'Abend+#+#+mit'
TOKEN: 'war+#+mit'
TOKEN: 'ich+mit'
TOKEN: 'Abend+#+#+#+meiner'
TOKEN: 'war+#+#+meiner'
[...]
```

SPBH – podobné jako OSB, uvažuje ale i slova nejen na hranici sliding window. Navíc tokeny mají váhy podle počtu slov, ze kterých jsou složeny.

```
TOKEN: 'mit'
TOKEN: 'ich+mit'
TOKEN: 'war+#+mit'
TOKEN: 'war+ich+mit'
TOKEN: 'Abend+#+#+mit'
TOKEN: 'Abend+#+ich+mit'
```

Klasifikace

- 4 klasifikátory Naive Bayesian,Graham-Bayesian, Burton-Bayesian,Fisher-Robinson's Chi-Square Algorithm , lze kombinovat.
- Klasifikátor neanalyzuje všechny tokeny. Pouze X nejvýznamnějších.
- Neuvažuje se vícenásobný výskyt tokenů.
- Záleží na použitém klasifikátoru.

Zpráva: “Hi! Buy Viagra.”

Slovník:

Token	Počet ve spamu (s)	Počet v hamu (h)	p(spam) = s/(s+h)
Hi	25	62	0.29
Buy	157	87	0.64
Viagra	231	11	0.95

Výpočet:

S = 0.29 * 0.64 * 0.95 = 0.176

H = (1-0.29) * (1-0.64) * (1-0.95) = 0.71 * 0.36 * 0.05 = 0.0127

Pvalue = S / (S + H) = 0,176 / (0,176 + 0.0127) = 0.93

Zdroje

- http://wiki.linuxwall.info/doku.php/en:ressources:dossiers:dspam#method_of_detection

- http://www.fi.muni.cz/tech/unix/spamy-a-viry.xhtml