Filtrování spamu (např. na FI)

Marek González, 342188

Součástí spamové kontroly na FI MUNI je program *dSpam* zajímavý právě z hlediska *machine learning*. Jedná se o program disponující algoritmy strojového učení s učitelem. Konkrétně jde o varianty bayesovksé sítě.

Sběr učící množiny emailů pro klasifikátory se na FI MUNI děje prostřednictvím schránky fungující jako volavka. Adresa této schránky je pro roboty na internetu dobře dostupná. Producenti spamu na ni pak zasílají spam, který slouží jako pozitivní data při fázi učení. Jako negativní data slouží emaily zaslané na adresu *notspam@fi.muni.cz.*[1]

Detekce spamu začíná fázi tokenizace zprávy. dSpam umožňuje zvolit mezi čtyřmi metodami tokenizace. Nejjednodušší metoda uvažuje za token každé slovo. Tedy pro zprávu *"Heute Abend war ich mit meiner Freundin im Kino und habe viel gelacht"* vznikne třináct tokenů. O něco složitější metoda považuje za token dvě slova jdoucí po sobě. Pro předchozí příklad by tedy vzniklo tokenů dvanáct. Třetí a čtvrtou metodou je pravoúhlý řídký bigram a řídké binární polynomiální hashování. Tyto metody berou jako složený token kombinaci slov, které zapadnou do posuvného okénka o určité velikosti.

Protože naivní bayesovský klasifikátor dokáží spameři zmást pouhým vložením většího množství slov typickými i pro ham, aniž by zastínili smysl spamu, program umožňuje kromě naivního Bayesova klasifikátoru použít i specializovanější algoritmy, popřípadě i jejich kombinaci. Tyto algoritmy jsou: Grahamů bayesovký klasifikátor, Burtonův bayesovksý klasifikátor a Fisher-Robinsonův algoritmus.

Klasifikace je založena na Bayesově teorému. Nejprve je pro každý token vypočítána pravděpodobnost, s jakou se vyskytuje ve spamu a s jakou v hamu. Poté je vypočítána pravděpodobnost, zda je zpráva spamem, na základě Bayesova teorému.[2]

Autoři dSpamu prohlašují, že průměrná přesnost se pohybuje mezi 99.5% a 99.95% a že vůbec nejlepší dosažená přesnost byla 99.991 %. Nicméně nezávislé testy výše zmíněna čísla nepotvrdily. [3]

Zdroje

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

[2] http://wiki.linuxwall.info/doku.php/en:ressources:dossiers:dspam#method_of_detection%C2%A0

[3] http://en.wikipedia.org/wiki/DSPAM