AntiSPAM pomocí Machine Learning v Pythonu

Získávání pravidel pro rozhodování, zda-li je daná e-mailová zpráva SPAM či HAM není vždy jednoduché a vyžaduje zdlouhavou a empiricky podloženou simulaci, sytém výpočtu skóre apod. Tímto zůusobem k probematice přistupují tradiční rešení, jako například Apache SpamAssassin[1].

Pro účely tohoto projektu je naopak mnohem výhodnější implementovat řešení jako klasifikační úlohu pomocí Machine Learning. Tento přístup je mnohem méně náročný na vlastní definice pravidel a při správné trénovací množině dat dokáže zajistit lepší výsledky. Jediné riziko je možný výskyt false-positive.

Naive Bayes Classifier

Nejvíce přímočaré řešení pro klasifikaci e-mailů je použít naivní Bayesovský klasifikátor nad textem každé zprávy. Pro každou zprávu tedy vytvoříme seznam vlastností, který odpovídá počtu výskytu jednotlivých lemmat v jeho obsahu a předmětu. Zároveň pro větší odolnost z těchto slov odstraňuje interpunkci a speciální znaky. Slova, která považujeme za příliš časté v češtině či angličtině taktéž nezohledňujeme. Seznam takových *stop words* získáme například zde[2].

Po natrénování klasifikátoru na korpusu je nutné jej uložit. Pro lepší přenositelnost se program exportuje komprimovaný ve formátu GZIP.

Testovací a trénovací korpus

Pro natrénování klasifikátoru byl použit korpus SPAMů a HAMů získaný z těchto zdrojů:

- CSDMC2010 od CSMining[3]
- Enron email dataset[4]
- TREC 2007 od NIST.gov[5]

Celkem se jedná o **88 847** zpráv, z toho:

HAMů	**34 871**
SPAMů	**53 976**

Bohužel je velmi problematické najít použitelný dataset pro e-maily v češtině. Předpokládám tedy velkou chybovost pro ne-anglické e-maily. Snahy dodat vlastní set SPAMů v češtině nebyly moc úspěšné, jelikož e-mailové služby SPAM starší 30 dní odstraňují. Navíc je třeba taká zmínit, že v testovacím korpusu je bohužel malé procento e-mailů nezpracovatelných pomocí `eml_parser`, jedná se o desítky až nižší stovky zpráv.

Trénování a následná kontrola byla provedena na těchto datech v poměru 4:1 (*trénování : testování*).

Použité externí knihovny

- `eml_parser` – pro čtení EML souborů
- `bs4` – za účelem parsování HTML obsahu
- `nltk` – Machine Learning

[1] http://spamassassin.apache.org/
[2] https://code.google.com/archive/p/stop-words/
[3] http://csmining.org/index.php/spam-email-datasets-.html
[4] http://www2.aueb.gr/users/ion/data/enron-spam/
[5] http://trec.nist.gov/data/spam.html