Rozpoznávání nevyžádané pošty na základě charakteristik textu

Petr Uher Jan Ulrych

Matematicko-fyzikální fakulta

3. 1. 2007

Čeho chceme dosáhnout

- Rozpoznávat nevyžádanou poštu
  - charakteristiky textu
  - nevyužívat emailové adresy
- Využít neuronové sítě
- Chyby typu *false-positive* jsou nežádoucí

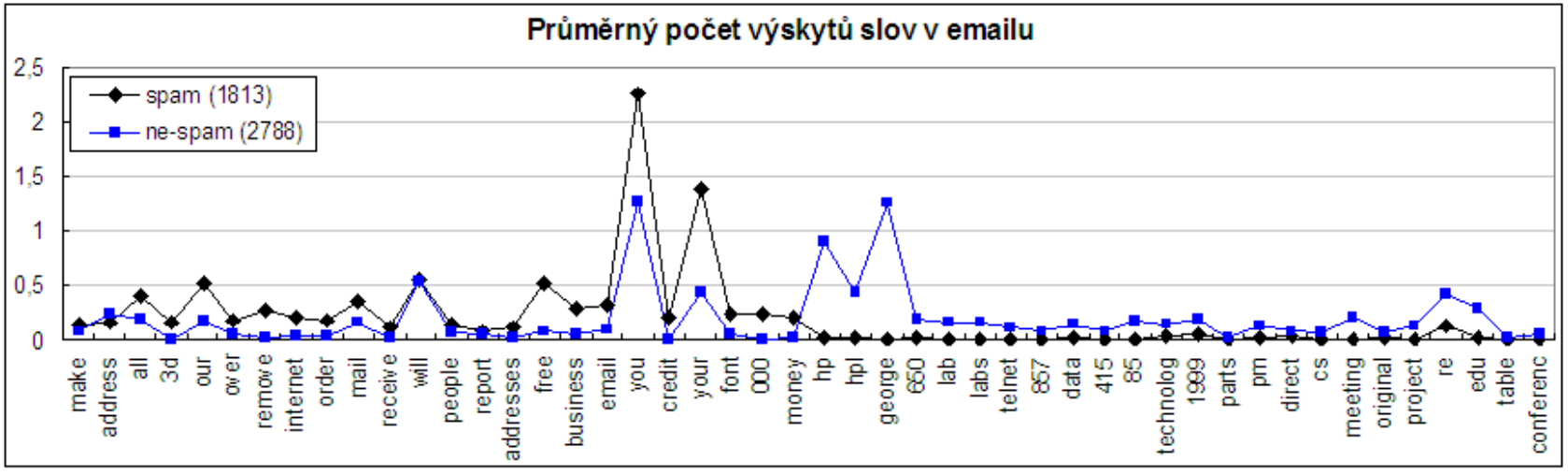
Popis dat

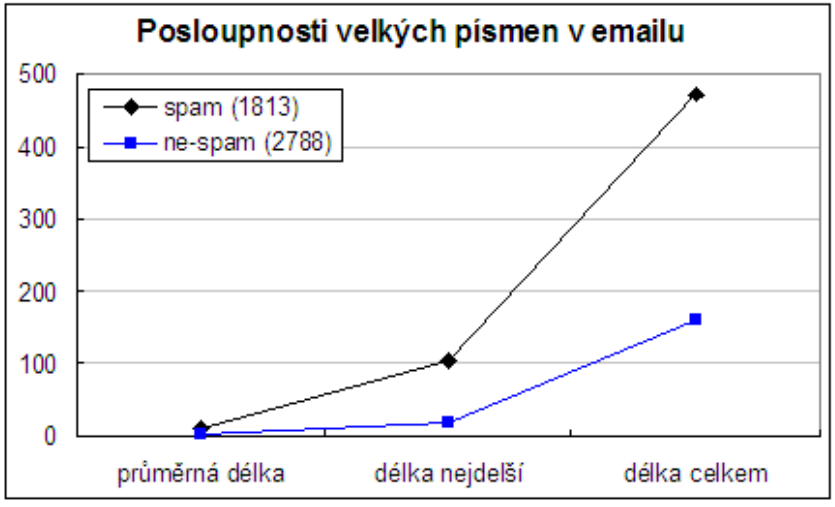
- SPAM E-mail database
  - Spambase Database z UCI Machine Learning
  - http://www.ics.uci.edu/~mlearn/MLSummary.html
- Hewlett-Packard
- Červen–červenec 1999
- Spam
  - reklamy, rychlé získání peněz, řetězové emaily, . . .
  - klasifikováno člověkem
- Hewlett-Packard: rozpoznání, zda je daný email spam nebo ne
  - $\approx 7\%$ chyba klasifikace
  - 20–25% chyba při vyloučení *false-positive*

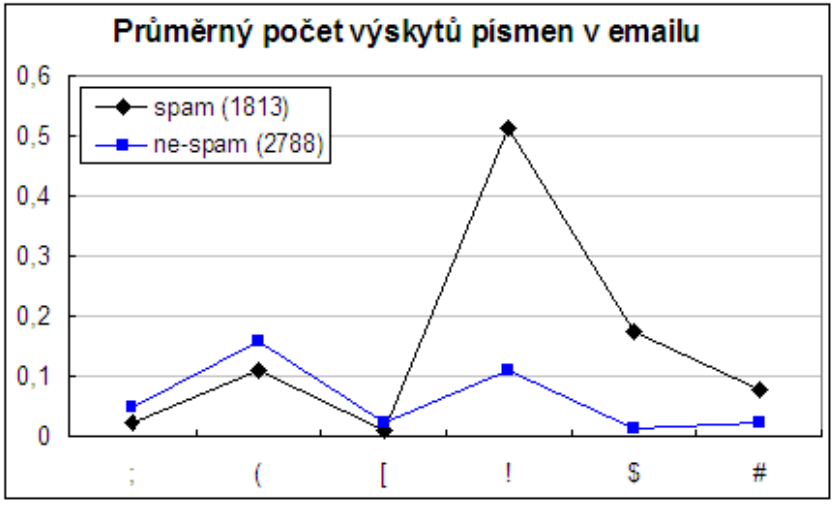
Popis dat

- Rozsah
  - 4601 emailů (1813 Spam = 39,4%)
- Atributy
  - 48 atributů četnosti slov (viz dále)
  - 6 atributů četnosti znaků (;, (, [, !, $, #)
  - průměrná délka posloupnosti velkých písmen
  - délka nejdelší posloupnosti velkých písmen
  - celková délka posloupností velkých písmen
  - 1 atribut klasifikace jako spam (hodnoty $\{0,1\}$)

Základní charakteristiky dat







Architektura sítě

- BP síť architektury 57–15–1 ((lineární), tansig, tansig)
- učení metodou `trainlm`
  - cíl: chyba sítě nižší než 0,01
  - nejvýše 100 epoch učení
- na výstup sítě aplikována skoková funkce

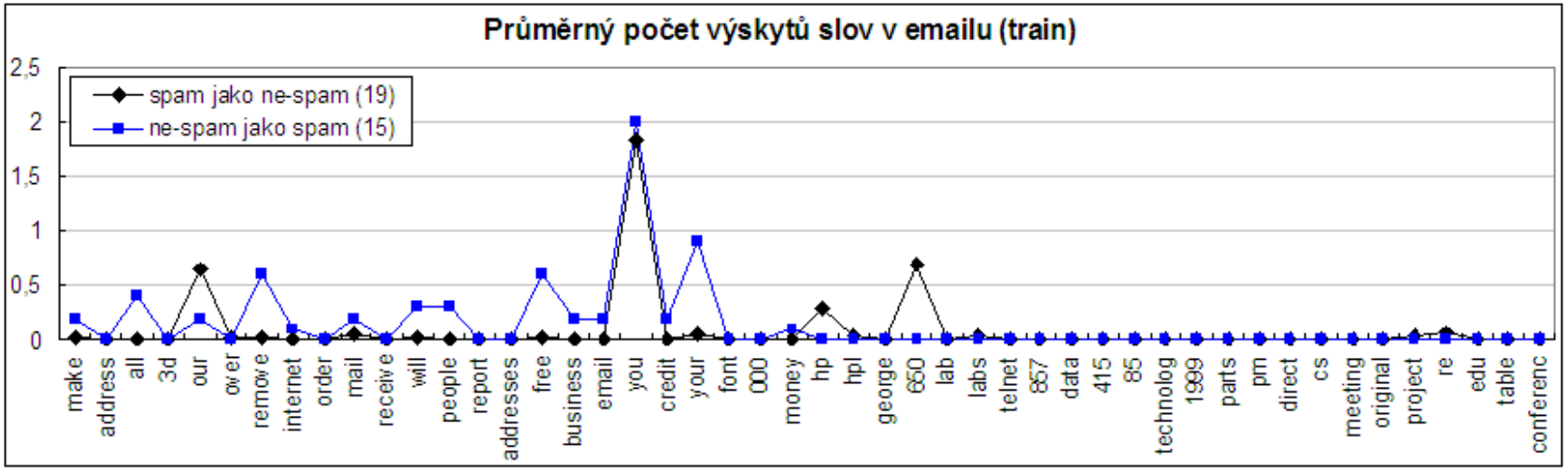
$$s(x) = \begin{cases} 0 & \text{pro } x < 0{,}6; \\ 1 & \text{pro } x \geq 0{,}6. \end{cases}$$

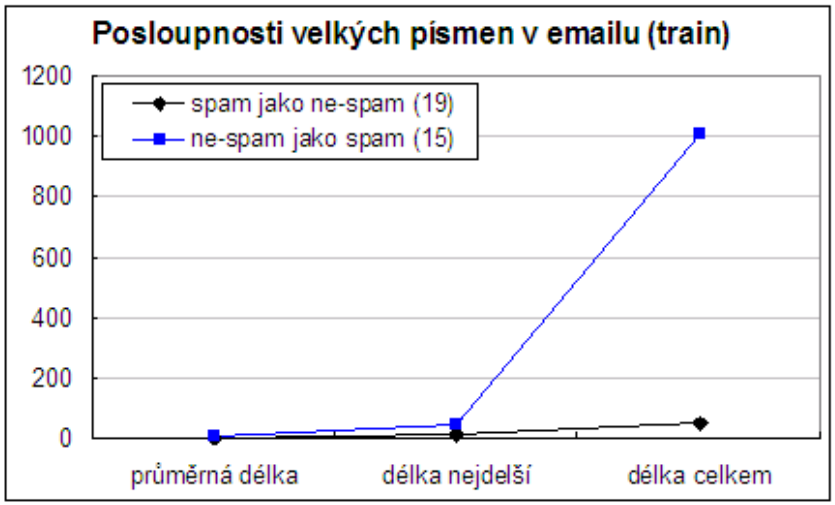
- trénovací množina
  - velikost: 2000
  - výběr: náhodný s rovnoměrným rozdělením

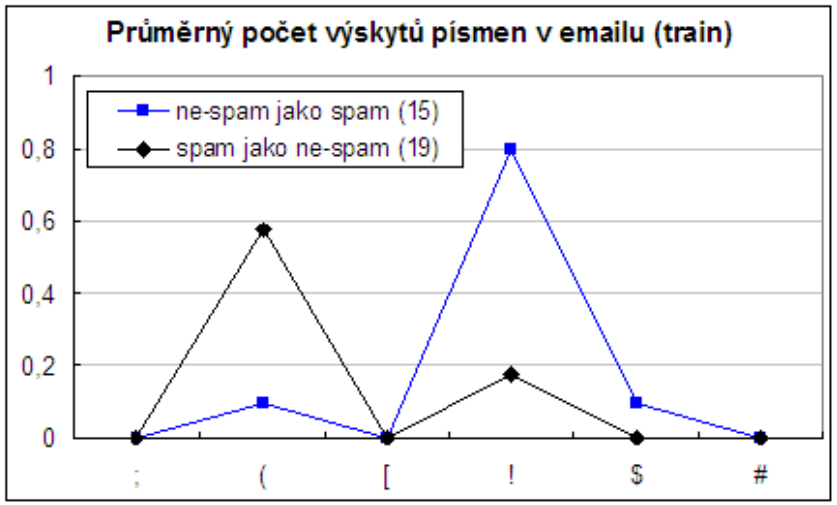
Výsledky

- síť se naučila rozpoznávat nevyžádané emaily . . .
- . . . s chybou $\approx 1\%$ na trénovací množině
- . . . s chybou $\approx 7{,}5\%$ na testovací množině
- nepodařilo se vyloučit chyby typu *false-positive*
- nepodařilo se snížit chybu na testovací množině

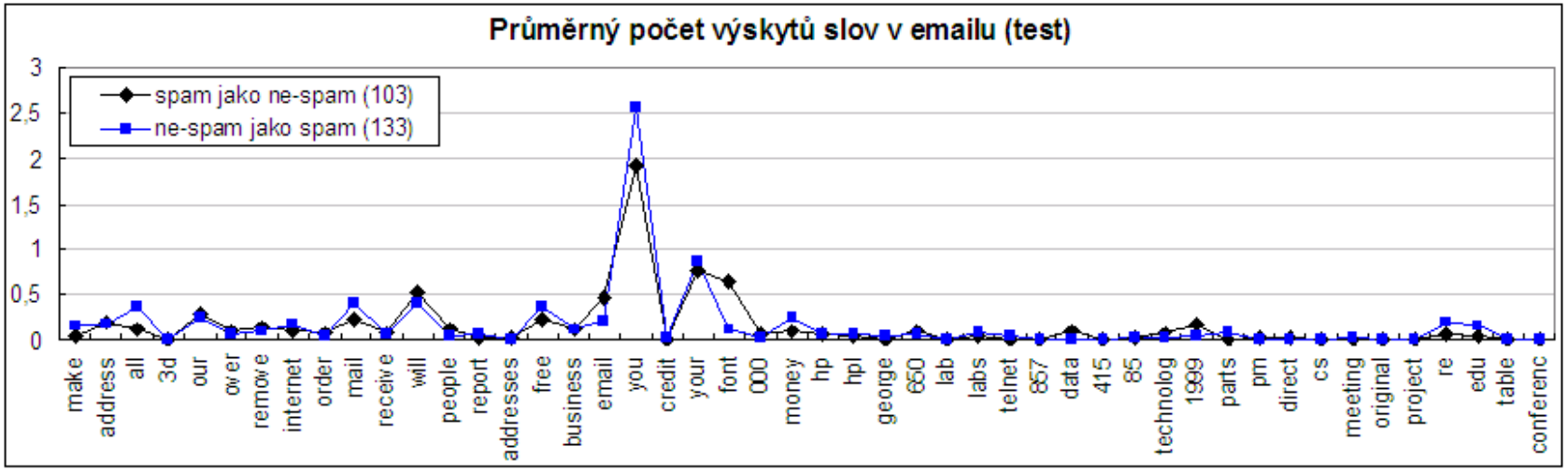
Chybně rozpoznané emaily

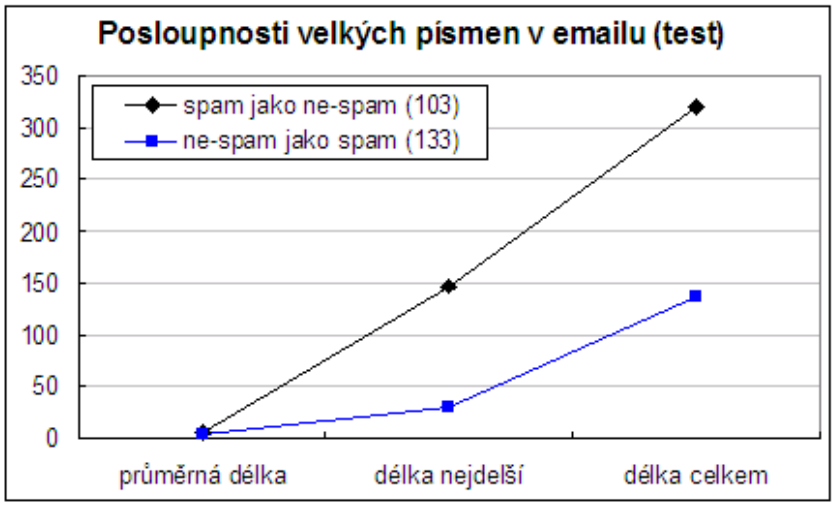


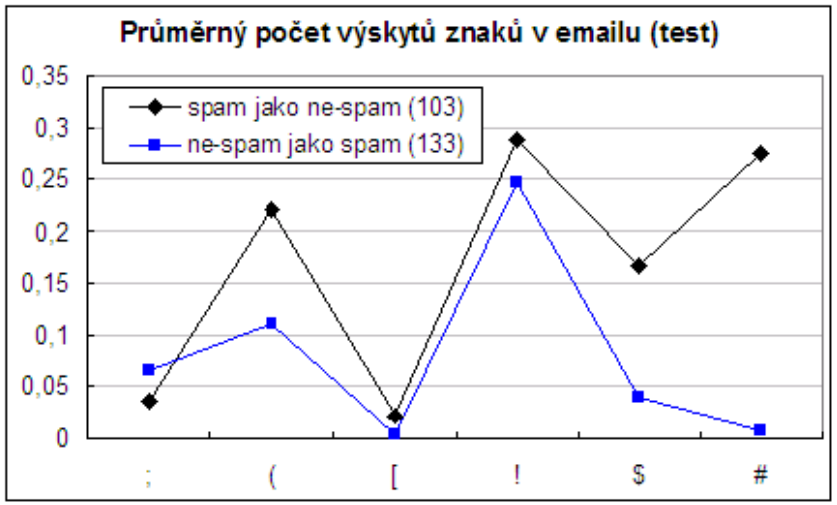




Chybně rozpoznané emaily







Poznámky k výsledkům

- Trénovací data
  - obecná slova (you, your)
  - chyba: společná slova pro spam i ne-spam
  - zlepšení: větší množství slov
- Na architektuře sítě příliš nezáleží
  - více vrstev
  - více neuronů ve skrytých vrstvách
- Předzpracování dat nemá významný vliv na velikost chyby
  - normalizace min–max
  - normalizace podle střední hodnoty a směrodatné odchylky
  - PCA

Shrnutí

- Neuronovou síť lze použít jako antispamový filtr.
- Nutný výběr reprezentativních dat.