IB047

Automatické značkování

Pavel Rychlý

pary@fi.muni.cz

25. dubna 2013

Automatické značkování

- vstup text
- výstup text + morfologické značky, příp. základní tvary
- různé přístupy
  - pravidlové
  - statistické
- trénování na označkovaných datech
- vyhodnocení na nazávislých datech

Vyhodnocení značkování

- precision – přesnost

$$precision = \frac{tp}{tp + fp}$$

- recall – pokrytí

$$recall = \frac{tp}{tp + fn}$$

- accuracy – úspěšnost

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

Statistické značkování

- pravděpodobnosti značek, slov, ...
- volíme nejpravděpodobnější značku
- odhad pravděpodobností z trénovacích dat

Vyhlazování pravděpodobností

- (ne-)nulová pravděpodobnost pro neviděné jevy
- snížení posti pro časté jevy, určení posti pro neviděné jevy
- Good-Turing
  $N = \sum_{r=1}^{max} rN_r$
  $p_0 = N_1/N$
  $p_r = \frac{(r+1)S(N_{r+1})}{rS(N_r)}$