Stanford POS Tagger

Lubomír Sedlář

4. prosince 2012

Obecně

- `http://nlp.stanford.edu/software/tagger.shtml`
- tagger napsaný v Javě (v 1.6+)
- licence GNU GPL v2 nebo pozdější
- dvě varianty – základní (43 MB), plná (163 MB)
- značky z Penn Treebank

Implementační detaily

- Maximum entropy classifier
- slovo se klasifikuje na základě levého kontextu délky 3
- heuristiky pro zlepšení přesnosti
  - podstatná jména s velkými písmeny
  - pro slovesa až 8 předchozích slov
  - pro částice doménová znalost

Výsledky

- Angličtina: správnost 97,28 %, na neviděných slovech 90,46 %
- Rychlost až 15 tisíc slov za sekundu (na „2008 nothing-special Intel server“)
- Arabština 96,5 %
- Čínština 93–94 %
- Němčina 96,9 %