Midtvejsopgave

Patrizia Paggio
21/4/2008

Denne opgave handler om part-of-speech (POS) tagging. Den består af flg. tre delopgaver som alle skal løses. Diskussionspunkterne kan evt. gennemgås mundtligt.

(1) Forklar hvad POS-tagging er og hvorfor det er en vigtig proces i sprogteknologiske applikationer.
Giv bl.a. eksempler på forskellige taggede korpora og forskellige tagsæt.

(2) Implementer en POS-tagger i Python. Taggeren skal kombinere forskellige metoder (regulære udtryk, 'n-gram'-tagging) ordnet efter hinanden, og den skal være trænet på et POS-tagget korpus (fx det engelske Brown-korpus).
Evaluer den på det oprindelige trænede korpus stadig ved hjælp af Python.
Beskriv de metoder du har valgt at implementere, og gør rede for deres styrker og svagheder. Diskuter resultaterne.

(3) Indsaml et testkorpus ved at downloade tekst fra flere hjemmesider, fx 2-3 uddrag af litterære værker fra www.thefreelibrary.com.
Skriv Python-kode der tokeniserer teksten og fjerner html-koderne. Anvend den udviklede tagger på dette korpus.
Skriv Python-kode til at undersøge resultaterne ved at udskrive lister af:
- de hyppigste navneord (både ental og flertal)
- de hyppigste verber (i forskellige tider)
- de hyppigste tvetydige ord (dvs. ord der får forskellige tags afhængige af konteksten)

Diskuter disse taggede resultater.
Kommenter anvendelsen af taggeren på testkorpusset i sin helhed.

(4) Kommenter din samlede opgaveløsning, inkl. evt. programmeringsmæssige vanskeligheder og mulige udvidelser.