Sprogteknologi I
Undervisningsplan
Forårssemester 2008

Patrizia Paggio

27/9/2007

1 Introduktion til sprogteknologi

Teoretisk præsentation:

Hvad er sprogteknologi
Hvorfor er det svært at processere sprog
Eksempler på applikationer
Kursusgennemgang

Læsestof:

J&M: s.1-18.

Øvelser:

Test af NLTK-installation og øvelse med tekstprocessering (fra NLTK-bogen, kapitel 2, s.1-26):

s. 8: 1, 3
s.15: 9
s.19: 1, 2
s.26: 5

2 Regulære udtryk og 'finite state automata'

Teoretisk præsentation:

Regulære udtryk: hvad de er og hvad de bruges til (e.g. navnegenkendelse)
Regulære udtryk i Pyhton
Finite state automata

Læsestof:

J&M: s.21-52.
NLTK-bogen: kap.2, s.26-31.

Øvelser:

Regulære udtryk i Python (fra NLTK-bogen, kapitel 2):

s.31: 1, 2
s.32: 6

3 Ord og korpora

Teoretisk præsentation:

Ord, tokens, typer, tokenisering
Ordfrekvens, Zipfs lov
Dokumentindeksering

Læsestof:

NLTK-bogen: kap. 3.

Øvelser:

Tokenisering og korpusanalyse med NLTK (NLTK bog, kap.3).

s. 8: 5, 6
s.13: 5
s.22: 5, 9.a

4 PoS tagging

Teoretisk præsentation:

Ordklasser
Syntaktisk opmærkning
Gold standard og evaluering

Læsestof:

J&M: s.287-319.
NLTK-bogen: kapitel 4 (s.1-15).

Øvelser:

At analysere tagget tekst med NLTK, fx fra NLTK-bogen kap. 4:

s. 9: 2, 3
s.12: 5, 6

(Look also at CST on-line tools, and apply some of the programs from the exercises to the Danish PAROLE corpus)

5 Statistiske sprogmodeller og n-grams

Teoretisk præsentation:

Hvad er en statistisk sprogmodel
Unsmoothed n-grams
Smoothing, backoff, entropi

Læsestof:

J&M: s.191-229.
NLTK-bogen: kapitel 4 (s.16-24); kapitel 7 (s.18-19).

Øvelser:

Fra NLTK-bogen kap.3-4:

kap.3	s.22:	6, 7
kap.4	s.21:	5, 6

Også øvelse hvor man genererer ord-associationer ud fra bigrams og kollokationer (se chapter3/exCollocAssoc.py): metoden kan senere sammenlignes med brug af WordNet til samme formål.

6 Morfologi og lemmatisering

Teoretisk præsentation:

Bøjnings- og afledningsmorfologi
Two-level morphology
Stemming og lemmatisering

Læsestof:

J&M: s.57-90.
Jongejan (2006).

Øvelser:

The Soundex algorithm and stemming in Python. Fra NLTK-bogen kap.2-3:

kap.2	s.32:	7
kap.3	s.13:	3

Plus en øvelse hvor man bruger CST-lemmatiseringsprogrammet.

7 Syntaks og grammatikker

Teoretisk præsentation:

Ord og syntagmer
Chunking
Grammatikker og syntakstræer

Læsestof:

J&M: s.323–352.
NLTK-bogen: kapitel 7 (s.1-11, 15-16), kapitel 8 (s.7-9).

Øvelser:

Fra NLTK-bogen kap.7-8:

kap.7	s.8:	3-5
kap.7	s.17:	3
kap.8	s.12:	9, 14

8 Parsing med CFG's

Teoretisk præsentation:

Top-down og bottom-up parsing
Chart parsing
Finite state parsing

Læsestof:

J&M: s.357–391. NLTK-bogen: kapitel 8 (s.13-15).

Øvelser:

Fra NLTK-bogen kap.8:

s.20:	3, 4
s.21:	5 (test grammatikken med både top-down og shift-reduce parsere)

9 Probabilistisk parsing og unifikationsbaseret parsing

Teoretisk præsentation:

PCFG's og probabilistisk parsing
Human parsing
Trækstrukturer og unifikation

Læsestof:

NLTK-bogen: kapitel 9 (s.21-25, 29).
J&M: s.395–428.

Øvelser:

Fra NLTK-bogen kap.9:

s.19 tilpas Listing 2 og 3 til at kunne trace kørsler med bottom-up og top-down strategier, og modificer grammatikken til at kunne parse et antal danske sætninger.

s.29 tilpas programmet på siden til at inducere grammatikproduktioner fra en større portion af træbanken og til at parse nogle engelske sætninger med viterbi-parseren. Lav en "trace" for at se hvad der sker.

10 Semantik

Teoretisk præsentation:

Semantisk kompositionalitet og formel semantik
Leksikalsk semantik
Wordnet

Læsestof:

J&M: s.543-626, 590-626.

Øvelser:

Wordnet øvelser fra NLTK, Word association øvelse med WordNet

11 En avanceret sprogteknologisk applikation

Enten I: "Semantisk entydiggørelse og tekstkategorisering" eller II: "Informationssøgning".

Teoretisk præsentation I:

Leksikalsk tvetydighed
Semantisk entydiggørelse vha. maskinindlæring
En anvendelse: tekstkategorisering

Læsestof I:

J&M: s.632-646, 658-660.

Øvelser I:

Måske en øvelse om tekstkategorisering fra J&M: den forudsætter kendskab til naive Bayes, som bogen introducerer ifm. speech. Måske kan jeg gøre det ifm. probabilistisk parsing.

Teoretisk præsentation II:

Termer og indeksering
Boolsk og vektorbaseret søgning
Evaluering i IR

Læsestof II:

J&M: s.646-658.

Øvelser II:

Se kurset i Informationsøgning.

12 Spørgetime

Litteratur

- The NLTK book (http://nltk.sourceforge.net/index.php/Book). Enheder per side: ca. 3000 (Teknisk test).
- J&M: Jurafsky, Daniel and James Martin (2000) *Speech and Language Processing*. Prentice-Hall. Enheder per side: ca. 3000.

NB: En normalside for tekniske tekster (fx. NLTK-bogen) er 1550 enheder inkl. mellemrum. For almindelige tekster er det 2400 enheder.

Andre relevante tekster:

- Ann Copestake: lectures in NLP (http://www.cl.cam.ac.uk/users/aac).
- Dorte Haltrup Hansen (2006) Sprogteknologiske værktøjer til tekst- og informationshåndtering. I: A. Braasch et al. *Sprogteknologi i dansk perspektiv*, Reitzels Forlag 2006.
- Bart Jongejan (2006) CST's lemmatiser for dansk. I: A. Braasch et al. *Sprogteknologi i dansk perspektiv*, Reitzels Forlag 2006.