Sprogteknologi I
Undervisningsplan
Forårssemester 2009

Version 1

Patrizia Paggio

25/1/2009

6.feb:
Lektion 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.
NLTK-bogen: kap. 1 + kap.3, s. 8–11.

Øvelser:
Test af NLTK-installation og øvelser med tekstprocessering (strings, lists, frequency distribution).

13.feb:
Lektion 2. Regulære udtryk og tilstandsmaskiner

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

Læsestof:
J&M: s.21–52.
NLTK-bogen: kap.3, s.1–7, 12–30.

Øvelser:
Regulære udtryk i Python.
Fra NLTK-bogen, kap. 2: nr. 1, 2, 4.
Øvelse med at matche danske vokaler.

20.feb:
Lektion 4. Ord og korpora

Teoretisk præsentation:
At arbejde med korpora i NLTK.

Læsestof:
NLTK-bogen: kap. 2 (minus sektion 2.5).

Øvelser:
Korpusanalyse med NLTK.
Fra NLTK-bogen, kap.2: et udvalg blandt nr. 3, 9, 12, 21, 22, 41.

27.feb:
Lektion 3. Mere om ord og korpora

Teoretisk præsentation:
Ord, tokens, typer, tokenisering
Ordfrekvens, Zipfs lov

Læsestof:
NLTK-bogen: kap. 3.

Øvelser:
Fra NLTK-bogen, kap.2: nr. 49(a).
Fra NLTK-bogen, kap.3: nr. 13, 14, 24.

6.mar:
Lektion 5. POS tagging

Teoretisk præsentation:
Ordklasser
Syntaktisk opmærkning
Gold standard og evaluering

Læsestof:
J&M: s.287-319.
NLTK-bogen: kapitel 4 (s.1–25).

Øvelser:
At analysere tagget tekst med NLTK, fx fra NLTK-bogen kap. 4: nr. 3, 6, 11, 12.
Vi vil også anvende nogle af de gennemgåede programmer på det danske PAROLE-korpus.

13.mar:
Lektion 6. Mere om PoS tagging

Teoretisk præsentation:
Transformationsbaseret PoS-tagging
CST-taggeren

Læsestof:
NLTK-bogen: kapitel 4 (s.25–30); artikel om Brill's tagger

Øvelser:
Fra NLTK-bogen, kap.4: nr. 20.

Arrangement kl.13-15:
Christiane Fellbaum taler om WordNet i Lingvistkredsen (www.lingvistkredsen.dk/). Detaljer følger senere.

20.mar:
Lektion 7. Statistiske sprogmodeller og n-grams

Teoretisk præsentation:
Hvad er en statistisk sprogmodel
Unsmoothed n-grams
Smoothing, backoff, entropi
Tekstklassifikation

Læsestof:
J&M: s.191-229.
Alternativt: NLTK-bogen, kap.5 (klassifikation)

Øvelser:
Øvelser om bigrammer, kollokationer, ordassociationer, evt. tekstklassifikation.
Gennemgang af evt. problemer og tidligere øvelser.

27.mar
Lektion 8. Morfologi og lemmatisering

Teoretisk præsentation:
Bøjnings- og afledningsmorfologi
Two-level morphology
Stemming og lemmatisering

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

Øvelser:
Stemming i Python.
Fra NLTK-bogen kap.3: nr. 25, 33.
Vi vil også arbejde med CST's on-line lemmatiser.

3.apr:
Lektion 9. Syntaks og grammatikker

Teoretisk præsentation:
Ord og syntagmer
Chunking

Grammatikker og syntakstræer

Læsestof:
J&M: s.323–352.
NLTK-bogen: kapitel 7.

Øvelser:
Fra NLTK-bogen kap.7: nr. 4, 5, 6, 11 (tilpasset PAROLE), 18 (ny).

Udlevering af midtvejsopgave

17.apr:
Lektion 10. Parsing med CFG's

Gennemgang af midtvejsopgave

Teoretisk præsentation:
Top-down og bottom-up parsing
Chart parsing
Finite state parsing

Læsestof:
J&M: s.357–391.
NLTK-bogen: kapitel 8.

Øvelser:
Fra NLTK-bogen kap.8:
nr. 14, 15, 16 (grammatikken testes med både top-down og shift-reduce parsere).
Plus øvelse fra appendix om Chart Parsing.

24.apr:
Lektion 11. Probabilistisk parsing og unifikationsbaseret parsing

Teoretisk præsentation:
PCFG's og probabilistisk parsing
Human parsing
Trækstrukturer og unifikation

Læsestof:
NLTK-bogen: enten kapitel 9 eller appendix om Probabilistisk Parsing.
J&M: s.395–428.

Øvelser:
Vælges enten fra NLTK-bogen kap.9 (en af opgaverne kan tilpasses til at kunne parse et antal danske sætninger), eller fra Appendix om Probabilistic Parsing (noget om at inducere grammatikproduktioner fra en træbank, om at parse nogle engelske sætninger med viterbi-parseren og om at producere en "trace").

1.maj:
Lektion 12. Semantik

Gennemgang af pensumopgivelser

Teoretisk præsentation:
Semantisk kompositionalitet og formel semantik
Leksikalsk semantik
Wordnet, inkl. anvendelser fx til informationssøgning

Læsestof:
J&M: s.543-626, 590-626.
NLTK-bogen: kap.2, sektion 2.5.

Øvelser:
Wordnet øvelser fra NLTK (fx 2.52, ny), 'Word association'-øvelse med WordNet.

22.maj:
Lektion 13. Spørgetime

Gennemgang af eksamensform og -regler
Gennemgang af eksamensopgave
Spørsgmål til eksamen

Eksamen

Pensumopgivelser: Fristen er ikke fastlagt endnu. Sidste år var det 13. maj.
Opgaveformulering: Fristen er ikke fastlagt endnu. Sidste år var det 16. maj, dvs. 2 uger inden fristen for opgaveafleveringen.
Opgaveaflevering: Fristen er ikke fastlagt endnu. Sidste år var det 6. juni inden kl.15.

Litteratur

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

- The NLTK book (http://www.nltk.org/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.
- Jørg Asmussen (2006) At måle og veje korpusser. I: A. Braasch et al. *Sprogteknologi i dansk perspektiv*, Reitzels Forlag 2006.
- 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.