Language Processing II
Teaching plan
Spring semester 2014
Part I: shallow text processing

Version 3

Patrizia Paggio and Costanza Navarretta

11/3/2014

19 Feb:
Lesson 3. Getting started

Lecture:
Finite state automata and regular expressions.
Practical:
Regular expressions in Python.
Searching and tokenizing text.
Lists and strings.
Readings:
J&M: chapter 2 on Regular Expressions and Automata
NLTK book: chapter 3.

26 Feb:
Lesson 4. Text corpora

Lecture:
Text corpora and annotation.
Frequency and Zipf's law.
The NLTK text corpora.
Practical:
Exercises with the NLTK corpus.
Readings:
M. Baroni (2008). Distributions in text. In Anke Ldeling and Merja Kyt (eds.), Corpus Linguistics: An International Handbook. Berlin: Mouton de Gruyter.
Biber, D. and Conrad, S. (2001). Quantitative corpus- based research: Much more than bean counting. TESOL Quarterly 35, 331-6.
NLTK book: chapter 2.1-3.

5 March:
Lesson 5. PoS-tagging

Lecture:
Part-of-Speech classes and tagging.
Tagging methods.
God standards and evaluation.
Readings:
J&M: chaper 5 on POS Tagging.
NLTK book: chapter 5

12 March:
Lesson 6. PoS-tagging

Practical:
PoS-tagging in NLTK on data in English and other languages.
Readings:
NLTK book: chapter 5

19 March:
Lesson 7. Syntactic structure

Lecture:
Words and phrases.
Chunking.
Advantages and disadvantages of shallow analysis methods.
Full syntactic and dependency parsing
Practical:
Chunking with NLTK on data in English and other languages.
Readings:
article to be announced.
NLTK book: chapters 7.2-4.

27 March
Lesson 8. Text classification

Note: this lesson is Thursday 15-17!
Lecture:
Text classification.
Presentation of project task.
Practical:
Project task work.
Readings:
article to be announced.
NLTK book: chapter 6.

2 April
Lesson cancelled

Discussion of project results

9 April
Lesson 9. Text classification

Discussion of project results

Literature

A normal page for technical text (e.g. the NLTK book) consists of 1550 characters including spaces. For non-techincal text the count is 2400 characters.

- The NLTK book (http://www.nltk.org/book). Characters per page: appr. 3000.
- J&M: Jurafsky, Daniel and James Martin (2000) *Speech and Language Processing*. Prentice-Hall. Characters per page: appr. 3000.