Course syllabus

Språkteknologi
Language Technology

EDAN20, 7,5 credits, A (Second Cycle)

Valid for: 2023/24
Faculty: Faculty of Engineering, LTH
Decided by: PLED C/D
Date of Decision: 2023-04-18

General Information

Elective for: C4-pv, D4-pv, D4-mai, E4-bg, F4, F4-pv, F4-mai, Pi4-pv, Pi4-bam, MMSR2
Language of instruction: The course will be given in English

Aim

In the past 15 years, language technology has considerably matured driven by the massive increase of textual and spoken data and the need to process them automatically. Although there are few systems entirely dedicated to language processing, there are now scores of applications that are to some extent "language-enabled" and embed language processing techniques such as spelling and grammar checkers, information retrieval and extraction, or spoken dialogue systems. This makes the field form a new requirement for the CS engineers.

The course introduces theories used in language technology. It attempts to cover the whole field from character encoding and statistical language models to semantics and conversational agents, going through syntax and parsing. It focuses on proven techniques as well as significant industrial or laboratory applications.

Learning outcomes

Knowledge and understanding
For a passing grade the student must

- Understand the field of language technology and major applications using them
- Know the most important techniques, fundamental algorithms, and most common architectures used in applications

- Create and implement language processing algorithms. Write, interpret, evaluate, and improve them during the programming laboratories.

Competences and skills
For a passing grade the student must

- Understand and develop annotation schemes, create and process structured documents
- Understand and write regular expressions and use them in languages like Python
- Understand and use machine--learning algorithms and statistical techniques
- Develop and evaluate algorithms in major fields of language technology: language models, partial parsing, dependency parsing, and semantic parsing using real data.

Judgement and approach
For a passing grade the student must

- Show curiosity, creativity, and problem solving aptitudes
- Show an understanding of industrial and research issues in language technology

Contents

- An overview of language technology: disciplines, applications, and examples
- Corpus and word processing: regular expressions, automata, an introduction to Python, concordances, tokenization, counting words, collocations
- Morphology and part-of-speech tagging: word morphology, transducers, part-of-speech tagging,
- Phrase-structure grammars: constituents, trees, DCG rules, unification.
- Partial parsing: multiword detection, noun group and verb group extraction, information extraction, evaluation
- Syntax: formalisms, constituency and dependency, functions, parsing, statistical parsing, dependency parsing.
- Semantics: formal semantics, lambda-calculus, lexical semantics, predicate--argument structures, frame semantics, semantic parsing.
- Discourse and dialogue: reference and coreference, discourse and rhetoric, discourse relations, parsing discourse relations, dialogue automata, speech acts, multimodality.

Examination details

Grading scale: TH - (U,3,4,5) - (Fail, Three, Four, Five)
Assessment: Compulsory course items: Assignments and possibly an examination. The coursework assignments are carried out in teams of two students, but can also be carried out individually. The first laboratory session will be dedicated to a hands-on approach to the programming tools used in the course. The assignments will then consist of six programming problems and individual reports. Passing all the assignments will consist in passing the course with a mark of 3. Optionally, the students will be able to set an examination and improve their mark to 4 or 5.

The examiner, in consultation with Disability Support Services, may deviate from the regular form of examination in order to provide a permanently disabled student with a form of examination equivalent to that of a student without a disability.

Parts
Code: 0113. **Name:** Statistical Techniques for Text Analysis.
Credits: 3,5. **Grading scale:** UG. **Assessment:** To qualify for a passing grade the laboratory work must be completed. **Contents:** Laboratory work.
Code: 0213. **Name:** Syntactic and Semantic Processing of Text.

Credits: 4. **Grading scale:** UG. **Assessment:** To qualify for a passing grade the laboratory work must be completed. **Contents:** Laboratory work.

Code: 0313. **Name:** Written Examination.

Credits: 0. **Grading scale:** TH. **Assessment:** Passing the course with a mark of 3 will consist in passing all the assignments. Optionally, the students will be able to take the written examination and improve their mark to 4 or 5. **Contents:** Optional written examination.

Admission

Admission requirements:

- EDAA01 Programming - Second Course or EDAA30 Programming in Java - Second Course or FRTF25 Introduction to Machine Learning, Systems and Control

The number of participants is limited to: No
The course overlaps following course/s: EDA171

Reading list

- Pierre Nugues: Language Processing with Perl and Prolog, Theories, Implementation, and Application. Springer Verlag, 2014, ISBN: 978-3-642-41464-0.

Contact and other information

Course coordinator: Professor Pierre Nugues, Pierre.Nugues@cs.lth.se
Course homepage: http://cs.lth.se/edan20