UNIVERSITY OF GOTHENBURG

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DIT247 Machine learning for natural language processing, 7.5 credits

Maskininlärning för språkteknologi, 7,5 högskolepoäng

Second Cycle

Confirmation

This course syllabus was confirmed by Department of Computer Science and Engineering on 2021-11-15 and was last revised on 2022-11-15 to be valid from 2023-08-28, autumn semester of 2023.

Field of education: Science 100%
Department: Department of Computer Science and Engineering

Position in the educational system

The course can be part of the following programmes: 1) Computer Science, Master's Programme (N2COS), 2) Mathematical Sciences, Master's Programme (N2MAT) and 3) Applied Data Science Master's Programme (N2ADS)

Main field of studies	*Specialization*
Computer Science	A1F, Second cycle, has second-cycle course/s as entry requirements
Data Science	A1F, Second cycle, has second-cycle course/s as entry requirements

Entry requirements

To be eligible to the course, the student should have a Bachelor's degree in any subject.

In addition, the course requires:

- 7.5 credits of courses in programming or equivalent,
- a course including probability and statistics, such as DIT862 Statistical Methods for Data Science or MSG810 Mathematical Statistics and Discrete mathematics,
- a first course in machine learning, such as DIT866 Applied Machine Learning, DIT381 Algorithms for Machine Learning and Inference, or MSA220 Statistical Learning for Big Data.

Applicants must prove knowledge of English: English 6/English B or the equivalent levelof an internationally recognized test, for example TOEFL, IELTS.

Learning outcomes

On successful completion of the course the student will be able to:

Knowledge and understanding

- describe the fundamentals of storing textual data for the world's languages,
- describe the most common types of natural language processing tasks,
- describe the most common types of machine learning models used in modern natural language processing,
- explain how text data can be annotated for a natural language processing task where machine learning techniques are used.

Competence and skills

- apply software libraries using machine learning for common natural processing tasks,
- write the code to implement some machine learning models for natural language processing,
- apply evaluation methods to assess the quality of natural language processing systems.

Judgement and approach

- discuss the advantages and limitations of different machine learning models with respect to a given task in natural language processing,
- reason about what type of data could be useful when training a model for a given natural language processing task,
- select the appropriate evaluation methodology for a natural language processing system and motivate this choice,
- reason about ethical questions pertaining to machine learning based natural language processing systems, such as stereotypes and under-representation.

Course content

The course gives an introduction to machine learning models and architectures used in modern natural language processing (NLP) systems.

Rapid developments in machine learning have revolutionized the field of NLP, including for commerically important applications such as translation, summarization, and information extraction. However, natural language data exhibit a number of peculiarities that make them more challenging to work with than many other types of

data commonly encountered in machine learning: natural language is discrete, structured, and highly ambiguous. It is extremely diverse: not only are there thousands of languages in the world, but in each language there is substantial variation in style and genre. Furthermore, many of the phenomena encountered in language follow long-tail statistical distributions, which makes the production of training data more costly. For these reasons, machine learning architectures for NLP applications tend to be quite different from those used in other fields.

The course covers the following broad areas:

- Working practically with text data, including fundamental tasks such as tokenization and word counting;
- probabilistic models for text, such as topic models;
- overview of the most common types of NLP applications;
- architectures for representation in NLP models, including word embeddings, convolutional and recurrent neural network, and attention models;
- machine learning models for common types of NLP problems, mainly categorization, sequence labeling, structured prediction and generation;
- approaches to transfer learning in NLP.

Sub-courses

1. **Project** *(Projekt),* 7.5 credits
   Grading scale: Pass with distinction (5), Pass with credit (4), Pass (3) and Fail (U)

Form of teaching

Lectures and assignments.

Language of instruction: English

Assessment

The course is examined by mandatory written assignments submitted as written reports, as well as a self-defined project that requires the submission of a written report and an oral presentation. Some of the assignments will be carried out individually and others in groups of normally 2-4 students. The project is conducted by 2-4 students.

A late submission of the assignments or project results in the grade Fail (U), unless special reasons exist. A failed assignment or project will be given the opportunity to submit a new solution on subsequent occasions the course is given.

If a student, who has failed the same examined element on two occasions, wishes to change examiner before the next examination session, such a request is to be submitted to the department in writing and granted unless there are special reasons to the contrary

(Chapter 6, Section 22 of Higher Education Ordinance).

In the event that a course has ceased or undergone major changes, students are to be guaranteed at least three examination sessions (including the ordinary examination session) over a period of at least one year, though at most two years after the course has ceased/been changed. The same applies to work experience and VFU, although this is restricted to just one additional examination session.

Grades

The grading scale comprises: Pass with distinction (5), Pass with credit (4), Pass (3) and Fail (U).
A passing grade for the entire course requires at least a passing grade for all assignments and the project. To be awarded a higher passing grade for the entire course, the student must, in addition, have a higher average on the weighted grades on the assignments and the project.

Course evaluation

The course is evaluated through meetings both during and after the course between teachers and student representatives. Further, an anonymous questionnaire is used to ensure written information. The outcome of the evaluations serves to improve the course by indicating which parts could be added, improved, changed or removed.

The results of and possible changes to the course will be shared with students who participated in the evaluation and students who are starting the course.

Additional information

The course is a joint course together with Chalmers.

Course literature will be announced at the latest 8 weeks prior to the start of the course.

The course replaces the course DIT245, 7.5 credits. The course cannot be included in a degree which contains DIT245. Neither can the course be included in a degree which is based on another degree in which the course DIT245 is included.