

Text Mining

Programme course

6 credits

Text Mining

TDDE16

Valid from: 2020 Spring semester

Determined by

Board of Studies for Computer Science
and Media Technology

Date determined

2019-09-23

Main field of study

Information Technology, Computer Science and Engineering, Computer Science

Course level

Second cycle

Advancement level

A1X

Course offered for

- Master's Programme in Computer Science
- Computer Science and Engineering, M Sc in Engineering
- Information Technology, M Sc in Engineering
- Computer Science and Software Engineering, M Sc in Engineering

Entry requirements

Note: Admission requirements for non-programme students usually also include admission requirements for the programme and threshold requirements for progression within the programme, or corresponding.

Prerequisites

Mathematical analysis; Linear Algebra; Probability and Statistics; Machine Learning; Basic programming.

Intended learning outcomes

The overall aim of the course is to provide an introduction to quantitative analysis of text, with special focus on applying machine learning methods to text documents. The student will learn all the main steps when working with text: efficient extraction of text, natural language processing of the text in a form suitable for statistical machine learning methods which are subsequently used for, among other things, text prediction.

After completing the course the student should be able to:

1. use basic methods for information extraction and retrieval of textual data
2. apply text processing techniques to prepare documents for statistical modelling
3. apply relevant machine learning models for analyzing textual data and correctly interpreting the results
4. use machine learning models for text prediction
5. evaluate the performance of machine learning models for textual data

Course content

Introduction and overview of quantitative text analysis and its applications. Information extraction. Web crawling. Information retrieval. Tf-idf. Vector space models. Text preprocessing. Bag of words. N-grams. Sparsity and smoothing for text. Document classification. Sentiment analysis. Model evaluation. Topic models.

Teaching and working methods

The course consists of lectures, computer laboratory work and an individual project. The lectures introduce concepts and theories that students then use in problem solving at the computer labs and in the project work.

Examination

PRA1	Project	3 credits	U, 3, 4, 5
LAB1	Laboratory exercises	3 credits	U, G

UPG1 consists of computer exercises that tests the students' ability to translate theoretical knowledge into practical problem solving in machine learning.

UPG2 is an individual project where the student solves a real-world problem involving text. The project is documented and evaluated by a written project report.

Grades

Four-grade scale, LiU, U, 3, 4, 5

Other information

Supplementary courses

Natural Language Processing

About teaching and examination language

The teaching language is presented in the Overview tab for each course. The examination language relates to the teaching language as follows:

- If teaching language is Swedish, the course as a whole or in large parts, is taught in Swedish. Please note that although teaching language is Swedish, parts of the course could be given in English. Examination language is Swedish.
- If teaching language is Swedish/English, the course as a whole will be taught in English if students without prior knowledge of the Swedish language participate. Examination language is Swedish or English (depending on teaching language).
- If teaching language is English, the course as a whole is taught in English. Examination language is English.

Other

The course is conducted in a manner where both men's and women's experience and knowledge are made visible and developed.

The planning and implementation of a course should correspond to the course syllabus. The course evaluation should therefore be conducted with the course syllabus as a starting point.

Department

Institutionen för datavetenskap

Director of Studies or equivalent

Ann-Charlotte Hallberg

Examiner

Marco Kuhlmann

Course website and other links

<http://www.ida.liu.se/~TDDE16/>

Education components

Preliminary scheduled hours: 28 h

Recommended self-study hours: 132 h

Common rules

Course syllabus

A syllabus must be established for each course. The syllabus specifies the aim and contents of the course, and the prior knowledge that a student must have in order to be able to benefit from the course.

Timetabling

Courses are timetabled after a decision has been made for this course concerning its assignment to a timetable module.

Interrupting a course

The vice-chancellor's decision concerning regulations for registration, deregistration and reporting results (Dnr LiU-2015-01241) states that interruptions in study are to be recorded in Ladok. Thus, all students who do not participate in a course for which they have registered must record the interruption, such that the registration on the course can be removed. Deregistration from a course is carried out using a web-based form: <https://www.lith.liu.se/for-studenter/kurskomplettering?l=en>.

Cancelled courses

Courses with few participants (fewer than 10) may be cancelled or organised in a manner that differs from that stated in the course syllabus. The Dean is to deliberate and decide whether a course is to be cancelled or changed from the course syllabus.

Guidelines relating to examinations and examiners

For details, see Guidelines for education and examination for first-cycle and second-cycle education at Linköping University, <http://styrdokument.liu.se/Regelsamling/VisaBeslut/917592>.

An examiner must be employed as a teacher at LiU according to the LiU Regulations for Appointments (<https://styrdokument.liu.se/Regelsamling/VisaBeslut/622784>). For courses in second-cycle, the following teachers can be appointed as examiner: Professor (including Adjunct and Visiting Professor), Associate Professor (including Adjunct), Senior Lecturer (including Adjunct and Visiting Senior Lecturer), Research Fellow, or Postdoc. For courses in first-cycle, Assistant Lecturer (including Adjunct and Visiting Assistant Lecturer) can also be appointed as examiner in addition to those listed for second-cycle courses. In exceptional cases, a Part-time Lecturer can also be appointed as an examiner at both first- and second cycle, see Delegation of authority for the Board of Faculty of Science and Engineering.

Forms of examination

Examination

Written and oral examinations are held at least three times a year: once immediately after the end of the course, once in August, and once (usually) in one of the re-examination periods. Examinations held at other times are to follow a decision of the board of studies.

Principles for examination scheduling for courses that follow the study periods:

- courses given in VT1 are examined for the first time in March, with re-examination in June and August
- courses given in VT2 are examined for the first time in May, with re-examination in August and October
- courses given in HT1 are examined for the first time in October, with re-examination in January and August
- courses given in HT2 are examined for the first time in January, with re-examination in March and in August.

The examination schedule is based on the structure of timetable modules, but there may be deviations from this, mainly in the case of courses that are studied and examined for several programmes and in lower grades (i.e. 1 and 2).

Examinations for courses that the board of studies has decided are to be held in alternate years are held three times during the school year in which the course is given according to the principles stated above.

Examinations for courses that are cancelled or rescheduled such that they are not given in one or several years are held three times during the year that immediately follows the course, with examination scheduling that corresponds to the scheduling that was in force before the course was cancelled or rescheduled.

When a course is given for the last time, the regular examination and two re-examinations will be offered. Thereafter, examinations are phased out by offering three examinations during the following academic year at the same times as the examinations in any substitute course. If there is no substitute course, three examinations will be offered during re-examination periods during the following academic year. Other examination times are decided by the board of studies. In all cases above, the examination is also offered one more time during the academic year after the following, unless the board of studies decides otherwise.

If a course is given during several periods of the year (for programmes, or on different occasions for different programmes) the board or boards of studies determine together the scheduling and frequency of re-examination occasions.

Registration for examination

In order to take an examination, a student must register in advance at the Student Portal during the registration period, which opens 30 days before the date of the examination and closes 10 days before it. Candidates are informed of the location of the examination by email, four days in advance. Students who have not

registered for an examination run the risk of being refused admittance to the examination, if space is not available.

Symbols used in the examination registration system:

** denotes that the examination is being given for the penultimate time.

* denotes that the examination is being given for the last time.

Code of conduct for students during examinations

Details are given in a decision in the university's rule book:
<http://styrdokument.liu.se/Regelsamling/VisaBeslut/622682>.

Retakes for higher grade

Students at the Institute of Technology at LiU have the right to retake written examinations and computer-based examinations in an attempt to achieve a higher grade. This is valid for all examination components with code "TEN" and "DAT". The same right may not be exercised for other examination components, unless otherwise specified in the course syllabus.

A retake is not possible on courses that are included in an issued degree diploma.

Retakes of other forms of examination

Regulations concerning retakes of other forms of examination than written examinations and computer-based examinations are given in the LiU guidelines for examinations and examiners,
<http://styrdokument.liu.se/Regelsamling/VisaBeslut/917592>.

Plagiarism

For examinations that involve the writing of reports, in cases in which it can be assumed that the student has had access to other sources (such as during project work, writing essays, etc.), the material submitted must be prepared in accordance with principles for acceptable practice when referring to sources (references or quotations for which the source is specified) when the text, images, ideas, data, etc. of other people are used. It is also to be made clear whether the author has reused his or her own text, images, ideas, data, etc. from previous examinations, such as degree projects, project reports, etc. (this is sometimes known as "self-plagiarism").

A failure to specify such sources may be regarded as attempted deception during examination.

Attempts to cheat

In the event of a suspected attempt by a student to cheat during an examination, or when study performance is to be assessed as specified in Chapter 10 of the Higher Education Ordinance, the examiner is to report this to the disciplinary board of the university. Possible consequences for the student are suspension from study and a formal warning. More information is available at <https://www.student.liu.se/studenttjanster/lagar-regler-rattigheter?l=en>.

Grades

The grades that are preferably to be used are Fail (U), Pass (3), Pass not without distinction (4) and Pass with distinction (5).

1. Grades U, 3, 4, 5 are to be awarded for courses that have written examinations.
2. Grades Fail (U) and Pass (G) may be awarded for courses with a large degree of practical components such as laboratory work, project work and group work.
3. Grades Fail (U) and Pass (G) are to be used for degree projects and other independent work.

Examination components

1. Grades U, 3, 4, 5 are to be awarded for written examinations (TEN).
2. Examination components for which the grades Fail (U) and Pass (G) may be awarded are laboratory work (LAB), project work (PRA), preparatory written examination (KTR), oral examination (MUN), computer-based examination (DAT), home assignment (HEM), and assignment (UPG).
3. Students receive grades either Fail (U) or Pass (G) for other examination components in which the examination criteria are satisfied principally through active attendance such as other examination (ANN), tutorial group (BAS) or examination item (MOM).
4. Grades Fail (U) and Pass (G) are to be used for the examination components Opposition (OPPO) and Attendance at thesis presentation (AUSK) (i.e. part of the degree project).

For mandatory components, the following applies: If special circumstances prevail, and if it is possible with consideration of the nature of the compulsory component, the examiner may decide to replace the compulsory component with another equivalent component. (In accordance with the LiU Guidelines for education and examination for first-cycle and second-cycle education at Linköping University, <http://styrdokument.liu.se/Regelsamling/VisaBeslut/917592>).

For written examinations, the following applies: If the LiU coordinator for students with disabilities has granted a student the right to an adapted examination for a written examination in an examination hall, the student has the right to it. If the coordinator has instead recommended for the student an adapted examination or alternative form of examination, the examiner may grant this if the examiner assesses that it is possible, based on consideration of the course objectives. (In accordance with the LiU Guidelines for education and examination for first-cycle and second-cycle education at Linköping University, <http://styrdokument.liu.se/Regelsamling/VisaBeslut/917592>).

The examination results for a student are reported at the relevant department.

Regulations (apply to LiU in its entirety)

The university is a government agency whose operations are regulated by

legislation and ordinances, which include the Higher Education Act and the Higher Education Ordinance. In addition to legislation and ordinances, operations are subject to several policy documents. The Linköping University rule book collects currently valid decisions of a regulatory nature taken by the university board, the vice-chancellor and faculty/department boards.

LiU's rule book for education at first-cycle and second-cycle levels is available at http://stydokument.liu.se/Regelsamling/Innehall/Utbildning_pa_grund-_och_avancerad_niva.