STAN49, Statistics: Analysis of Textual Data, 7.5 credits

Statistik: Analys av textdata, 7,5 högskolepoäng

Second Cycle / Avancerad nivå

Details of approval

The syllabus was approved by The Board of the Department of Statistics on 2021-11-29 to be valid from 2022-08-29, autumn semester 2022.

General Information

Second cycle level course in statistics. The course may be included in Master's degree in statistics. The course may also be taken as a single subject course or within other Master’s programmes at Lund University.

Language of instruction: English

Main field of studies	*Depth of study relative to the degree requirements*
Statistics	A1F, Second cycle, has second-cycle course/s as entry requirements

Learning outcomes

Knowledge and understanding

For a passing grade, the student shall

- be able to describe various techniques for preprocessing textual data and to explain when and why these should be utilised,
- be able to account for different ways of representing text data as vectors,
- be able to explain different techniques for classification and clustering of text data,
- be able to explain different techniques for topic modelling and sentiment analysis, and
- be able to explain different techniques for information extraction and text summarisation.

Competence and skills

For a passing grade, the student shall

- be able to apply techniques to solve various textual data analysis problems,
- independently be able to identify and formulate an issue related to textual data and be able to present a solution to this, and
- in writing be able to report clearly and discuss his/her findings of various textual data analysis problems.

Judgement and approach

For a passing grade, the student shall

- be able to make assessments of model choices based on the issue and available data and computational capacity.

Course content

The course provides an introduction to statistical analysis of text. The following topics are covered:

- Preprocessing of textual data
- Text representation
- Text classification
- Text clustering
- Topic modelling
- Sentiment analysis
- Text summarisation

During the course, both methods based on classic statistical approaches (including Bayesian models) and modern approaches such as deep learning and recurrent neural networks will be presented.

Course design

The course is designed as a series of lectures, computer lab sessions and seminars. Peer reviewing of other students' assignments is an important and compulsory part of the course.

Assessment

The examination consists of quizzes, assignments, and peer reviewing of assignments. The final grade is determined as a weighted sum of the results on the quizzes (33 %) and the assignments (67 %).

The University views plagiarism very seriously, and will take disciplinary actions against students for any kind of attempted malpractice in examinations and assessments. Plagiarism is considered to be a very serious academic offence. The penalty that maybe imposed for this, and other unfair practice in examinations or assessments, includes suspension from the University.

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.

Subcourses that are part of this course can be found in an appendix at the end of this document.

Grades

Marking scale: Fail, E, D, C, B, A.

A (Excellent) 85-100 points/percent. A distinguished result that is excellent with regard to theoretical depth, practical relevance, analytical ability and independent thought.

B (Very good) 75-84 points/percent. A very good result with regard to theoretical depth, practical relevance, analytical ability and independent thought.

C (Good) 65-74 points/percent. The result is of a good standard with regard to theoretical depth, practical relevance, analytical ability and independent thought and lives up to expectations.

D (Satisfactory) 55-64 points/percent. The result is of a satisfactory standard with regard to theoretical depth, practical relevance, analytical ability and independent thought.

E (Sufficient) 50-54 points/percent. The result satisfies the minimum requirements with regard to theoretical depth, practical relevance, analytical ability and independent thought, but not more.

F (Fail) 0-49 points/percent. The result does not meet the minimum requirements with regard to theoretical depth, practical relevance, analytical ability and independent thought.

To pass the course, the students must have been awarded the grade of E or higher.

Entry requirements

STAN48 Statistics: Advanced Statistical Programming and STAN52 Statistics: Advanced Machine Learning, or the equivalent.

Subcourses in STAN49, Statistics: Analysis of Textual Data

Applies from V23

2301 Quizzes, 2,5 hp
Grading scale: Fail, Pass
2302 Assignments, 5,0 hp
Grading scale: Fail, Pass