# *Statistical Natural Language Processing*

*Course syllabus, summer 2021*

## *Course Description*

This course is an undergraduate introduction to (statistical) natural language processing (NLP), aiming to expose students to a large variety of topics in NLP. In the first part of the course, we will go through a number of established and 'traditional' machine learning methods, as well as some popular and 'new' ones. The second part of the course introduces common tasks, methods and applications of NLP.

This is a practical, fast-paced, broad introduction to the field. Fluency in programming and ability to learn new programming languages and/or environments will be assumed.

The course language is English.

Tentative schedule

| | |
|---|---|
| W1 | Introduction |
| W2–3 | Preliminaries: linear algebra, probability theory |
| W3 | ML Intro, regression |
| W4–5 | Classification, evaluation |
| W6 | Neural networks (basics) |
| W7 | Unsupervised learning |
| W8 | Dense vector representations |
| W9–10 | Sequence learning |
| W11 | Language models |
| W12 | Tokenization, POS tagging, morphology, text classification |
| W13 | Statistical parsing |
| W14 | Wrap up & exam |

Visit the course web page for a more detailed and up-to-date version of the schedule.

## *Prerequisites*

The students should be fluent in programming, either able to program in Python, or capable of learning by themselves in a short time. Some familiarity with (computational) linguistics is also assumed.

For the ISCL students, the above requirements are covered in courses ISCL-BA-06 and ISCL-BA-07.

## *Recommended literature*

Daniel Jurafsky and James H. Martin (2009) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Pearson Prentice Hall, second edition.[1]

Trevor Hastie, Robert Tibshirani, and Jerome Friedman (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer-Verlag, second edition.[2]

## *Course work and evaluation*

In total, the coursework is worth 9 ECTS. Your grade will be determined based on 7 graded programming assignments (60 %) and a written final exam (40 %) at the end of the course. You can also get up to 5 % bonus by completing weekly quizzes administered through Moodle.[3]

If you are a master's student, you can either take the course as a 6ECTS 'Proseminar' for the regular coursework, or as a 9ECTS 'Hauptseminar' with an additional project and associated term paper.

Each assignment constitute 10 % of the overall course grade. Only 6 best assignment scores will contribute to your final score. Late assignments up to one week are graded with a maximum of 5 % (of the total grade). Assignments later than one week will not be accepted.

[1] Chapters from 3rd edition draft are available at `http://web.stanford.edu/~jurafsky/slp3/`.

[2] An updated version of the complete book is available at `http://web.stanford.edu/~hastie/ElemStatLearn/`.

[3] The following grade scale will be used to determine your final grade.

| Percent | Local | ECTS |
|---|---|---|
| > 96 | 1.0 | A |
| 93–96 | 1.3 | A |
| 89–92 | 1.7 | B |
| 85–88 | 2.0 | B |
| 81–84 | 2.3 | C |
| 77–80 | 2.7 | C |
| 73–76 | 3.0 | C |
| 69–72 | 3.3 | D |
| 65–68 | 3.7 | E |
| 60–64 | 4.0 | E |
| < 60 | 5.0 | F |

You are encouraged work on the assignments in pairs, but you are *not allowed* to pair with the same participant twice.

A retake of the final exam is possible, only if you failed the course, but it is still possible to get a passing overall grade by obtaining a higher grade on the exam.

If you take the course as a 'Hauptseminar', half of your grade will be determined by the assignments, while the other half will be based on a term project/paper.

## *Online course environment*

Due to the COVID-19 pandemic, the course is completely held online. We will make use of the utilities offered by Moodle, but more importantly we will use git version management system through GitHub classroom environment for distribution and submission of the assignments. You are required to register the course space on Moodle.[4] The information on online lectures will be posted in this course space. *Please also make sure to obtain a GitHub account, and complete the 'beginning of semester survey' on Moodle.*[5]

## *Academic conduct*

You are encouraged to discuss your assignments and other class work with others, do research on the Internet and use other sources for knowledge and inspiration. However, unless stated/cited explicitly, all the coursework you submit should be your own work. You are required to cite any source you have used. If you 'borrow' code that is crucial for the solution of an assignment, you will lose points. Not indicating the source of external code is plagiarism.

Plagiarism or any other form of academic misconduct will not be treated lightly.

## *Practical information*

| | |
|---|---|
| Instructor | Çağrı Çöltekin ⟨`ccoltekin@sfs.uni-tuebingen.de`⟩ |
| Office hours | Mon 14:00–15:00 (online) |
| Tutors | Anna-Katharina Dick ⟨`anna-katharina.dick@student.uni-tuebingen.de`⟩<br>Jingwen Li ⟨`jingwen.li@student.uni-tuebingen.de`⟩ |
| Time | Mon 12:00–14:00 & Wed 12:00–14:00 & Fri 12:00–14:00 |
| Location | online |
| Course web page | `http://sfs.uni-tuebingen.de/~ccoltekin/courses/snlp/` |

[4] `https://moodle.zdv.uni-tuebingen.de/course/view.php?id=1645`.

[5] `https://moodle.zdv.uni-tuebingen.de/mod/feedback/view.php?id=78334`.

Updated: April 18, 2021