Statistical Natural Language Processing

Course syllabus SS 2018

Course Description

Natural language processing (NLP) is an important part of areas or disciplines that are concerned with language and computation including (computational) linguistics, artificial intelligence and computer science. NLP is crucial both for building practical applications and answering research questions in many academic disciplines.

This course is an undergraduate introduction to (statistical) natural language processing, 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, Python refresher
W2–3	Preliminaries: linear algebra, probability theory
W4	ML: intro, regression
W5	ML: classification, evaluation
W6	ML: sequence learning, unsupervised learning
W7–8	ML: (deep) neural networks
W9	Segmentation/tokenization, n-gram language models
W10	POS tagging, morphology
W11	Parsing
W12	Computational semantics
W13	NLP applications: text classification, machine translation, question answering, …
W14	Wrap up & exam

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

Prerequisites

Successful completion of courses ISCL-BA-06 and ISCL-BA-07 or equivalent coursework or experience is required.

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

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]

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

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

Course work and evaluation

- Assignments (30 % + 5 % bonus)
- Final exam (70 %)
- Full attendance to the course will also be rewarded with a 5 % bonus

In total, the coursework is worth 9 ECTS.[3]

[3] This corresponds to 270 hours of course work (of which only 84 hours are in-class work).

Assignments

There will be seven programming assignments in this course. We will use git version management system through GitHub classroom for distribution and submission of the assignments. Please make sure to obtain a GitHub account, and complete the `Assignment 0` (see the description below).

Each assignment constitute the 5 % of the overall course grade. Late assignments up to one week are graded with a maximum of 2.5 %. The solutions to the assignments will be discussed in the class one week after the deadline. Assignments later than one week will not be accepted.

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

Assignment 0

As a warm-up exercise, you are **required** to complete a short ungraded assignment. To be able to receive the assignment, go to the URL on the printed version of this document distributed during the first session of the course. Please follow the steps described in the `assignment0.pdf` carefully.

If you do not complete this assignment, you will not be able able to receive the other assignments. Deadline for assignment 0 is **April 25, 2018, 10:00**.

Academic conduct

You are encouraged to discuss your assignments and other class works with other class participants, do research on the Internet and use other sources for knowledge and inspiration. However, unless stated/cited clearly and explicitly, all the coursework you submit should be your own work.

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

Practical information

Instructor	Çağrı Çöltekin ⟨`ccoltekin@sfs.uni-tuebingen.de`⟩
Tutor	Verena Blaschke ⟨`verena.blaschke@student.uni-tuebingen.de`⟩
Course hours/location	Mon 12:00–14:00 & Wed 10:00–12:00 & Fri 12:00–14:00, room 0.02
Course web page	`http://sfs.uni-tuebingen.de/~ccoltekin/courses/snlp/`
Office hours	Wed 12:00–14:00 (room 1.09)