Introduction to Text Mining

Organizational

Henning Wachsmuth

https://cs.upb.de/css

Organizational

Course

- Lectures. Henning Wachsmuth
- Tutorials. Milad Alshomary, Maximilian Spliethöver
- Languages. English, Python

Information

- Public. `https://cs.upb.de/css/teaching/courses/text-mining-w20`
- Internal. `https://paul.upb.de` and `https://panda.upb.de`

  → L.079.05501 Introduction to Text Mining (in English)

Dates (latest video/slide upload, live chat)

- Lectures. Thursday 14–17 c.t., as of October 29, online
- Tutorials. Wednesday 14–16 c.t., as of November 4, online

  First tutorial introduces Python and clarifies the assignment concept.

Need for consultation?

- Set up appointment with me via e-mail (henningw@upb.de).

Organizational

Web Resources of this Course

Course web page

- General. Detailed course information, general announcements
- Lectures. Slides

PAUL

- General. Standard course information
- Registration. Module, course, course achievement, exam

PANDA

- General. All announcements, asynchronous Q&A (forum)
- Lectures and tutorials. Videos, slides
- Assignments. Sheets, group forums and submissions, results

BigBlueButton (BBB)

- Lectures and tutorials. Weekly synchronous Q&A (text/audio chat)

  Links provided in PANDA.

Organizational

How to Complete the Course (information from the student advisory service)

Four registrations needed

- Module + course. Both until Nov 13, 2020
- Course achievement. Nov 9 – Dec 3, 2020

  Cancellation until Jan 29, 2020
- Examination. Nov 9 – Dec 3 (phase 1) and Mar 1–5, 2020 (phase 2).

  Cancellation until one week before examination takes place

How to register

- All registrations are done in PAUL, with two clicks ("Register", "Submit").
- Register for everything you see in PAUL for this module or course.

  All relevant information is available in PAUL — somewhere.

Notice

- Regularly check the e-mail address that PAUL sends its messages to.
- If anything looks suspicious in PAUL, contact the examination office.
- For advice, contact study-cs@mail.upb.de or see office hours: `https://cs.upb.de/studium/beratung-und-unterstuetzung/fachberatung/`

Course

Overall goal

- Learn major skills needed to approach typical text mining tasks.

Contents

- Several linguistic and statistical text analysis techniques.
- Several text mining tasks and applications.
- Needed basics of linguistics, empirical methods, and machine learning.

Competences

- Understanding of theory and practice of text mining.
- Design and implementation of text mining approaches for given tasks.
- Scientific experiments and evaluations on large amounts of data.

Course

Basics this Course Builds upon

Required basics

- Models and algorithms. Concepts and methods from first semesters.
- Languages. Understanding of natural and formal languages.
- Math. Basic probability theory and linear algebra.
- Programming. Some experience with software development.

Covered basics

- Linguistics. Fundamental language concepts and phenomena.
- Statistics. Concepts and methods related to empirical methods.
- Machine learning. Fundamental learning concepts and methods.
- Programming. Implementation in Python.

  Python mostly covered in the tutorials only.

Course

Your Tasks

Course achievement (“Studienleistung”)

- 6 assignment sheets, bi-weekly (∼50% written, ∼50% programming).

  First sheet published on Nov 5; to be submitted by Nov 15, 23:59 (UTC+1).
- Group submissions of up to 3 people strongly recommended.
- Notice. 50%+ of all assignment points needed to take the exam.

Exams

- Oral, ∼30 minutes, questions on all lecture parts, in English.

  A list of example questions will be provided early enough.
- First exam dates tentatively second half of February.

  Details follow in some weeks.

Differences for 4-ECTS students

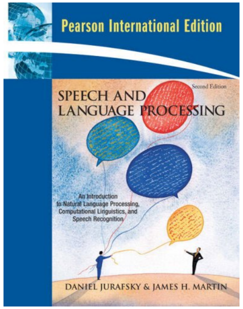
- Exam will not include last lecture part and another part of your choice.
- Still, 50%+ of all assignment points needed.

Textbooks (Not Mandatory)

Daniel Jurafsky and James H. Martin (2009).

Speech and Language Processing.

- Oriented towards computational linguistics
- Comprehensive
- Draft 3rd ed.: `http://web.stanford.edu/~jurafsky/slp3`



Pearson International Edition

Second Edition

SPEECH AND LANGUAGE PROCESSING

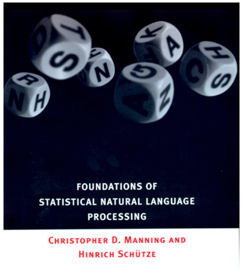
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition

DANIEL JURAFSKY & JAMES H. MARTIN

Christoper D. Manning and Hinrich Schütze (1999).

Foundations of Statistical Natural Language Processing.

- More oriented towards computer science
- Comprehensive, a bit outdated



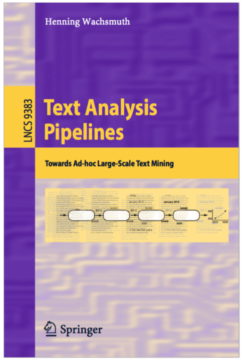
FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING

CHRISTOPHER D. MANNING AND HINRICH SCHÜTZE

Henning Wachsmuth (2015).

Text Analysis Pipelines.

- Rather oriented towards computer science
- Focused on advanced text mining techniques
- Book preprint: `http://www.arguana.com/publications/wachsmuth15a-springer-preprint.pdf`



Henning Wachsmuth

LNCS 9383

Text Analysis Pipelines

Towards Ad-hoc Large-Scale Text Mining

Springer

Outline of the Course

I. Overview

II. Basics of Linguistics

III. Text Mining using Rules

IV. Basics of Empirical Methods

V. Text Mining using Grammars

VI. Basics of Machine Learning

VII. Text Mining using Unsupervised Learning

VIII. Text Mining using Supervised Learning

IX. Practical Issues