Introduction to Text Mining

Organizational

Henning Wachsmuth

https://cs.upb.de/css

Organizational

Meta

- Course number. L.079.05534
- Modules. Human machine interaction, Computer science 2
- Instructors. Henning Wachsmuth (lectures), Milad Alshomary (tutorials)
- Languages. English, Python

Tasks

- Six assignments. Bi-weekly; ~50% programming, ~50% written.
  First one published on October 18; to be submitted on October 28, 23:59 (UTC+1).
- Exam. Oral. First round tentatively in February.
  50%+ of all assignment points needed to take the exam. General registration on PAUL.

Latest information

- Web page. `http://cs.upb.de/css/teaching/courses/text-mining-w18`
- PAUL. `http://paul.upb.de`

Organizational

Lectures and Tutorials

Dates and locations

- Lectures. Thursday 11–14, as of October 11, in O2

  No lecture on November 1 (holiday)

- Tutorials. Monday 11–13, as of October 15, in H3

  First tutorial introduces Python and clarifies the assignment concept.

Three lecture time options

1. Early bird. Start 11:00, end 13:30, 15 minutes break

   Mensa-friendly, not Fü-friendly, attention-friendly

2. Starvin' Marvin. Start 11:15, end 13:45, 15 minutes break

   Not Mensa-friendly, Fü-friendly, attention-friendly

3. Workhorse. Start at 11:15, end at 13:30, no break

   Mensa-friendly, Fü-friendly, not attention-friendly

Chosen option (based on discussion in the lecture)

- Early bird. As of October 18, the lecture will start at 11:00 s.t.

Goals of the course

Overall

- 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 research, 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.

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.
- Development. Some experience with software development in any programming language.

Covered basics

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

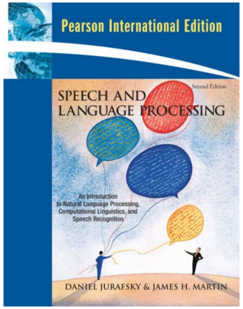
  Python mostly covered in the tutorials only.

Textbooks (Not Mandatory)

Daniel Jurafsky and James H. Martin (2009).

Speech and Language Processing.

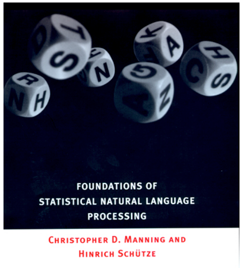
- Oriented towards computational linguistics
- Comprehensive
- Draft of 3rd ed.: `http://web.stanford.edu/~jurafsky/slp3`


Pearson International Edition
SPEECH AND LANGUAGE PROCESSING
Second Edition
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.

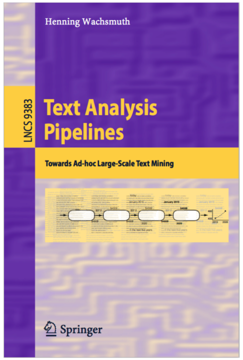
- More oriented towards computer science
- Comprehensive, a bit outdated


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
- Thesis version: `http://www.arguana.com/publications/wachsmuth15c-lncs.pdf`



Outline of the Course

I. Overview

II. Basics of Linguistics

III. Text Mining using Rules

IV. Basics of Empirical Research

V. Text Mining using Grammars

VI. Basics of Machine Learning

VII. Text Mining using Clustering

VIII. Text Mining using Classification and Regression

IX. Practical Issues

X. Text Mining using Sequence Labeling