Introduction to Text Mining

Organizational

Henning Wachsmuth

https://cs.upb.de/css

Organizational

Meta

- Course number. L.079.05501
- Instructors. Henning Wachsmuth (lectures), Milad Alshomary (tutorials)
- Languages. English, Python

Organizational

Meta

- Course number. L.079.05501
- Instructors. Henning Wachsmuth (lectures), Milad Alshomary (tutorials)
- Languages. English, Python

Tasks

- Six assignments. Bi-weekly; ∼50% written, ∼50% programming.
  First one published on October 17; to be submitted on October 27, 23:59 (UTC+1).
- Exam. Oral (tentatively!!). First round in February.
  Course achievement: 50%+ of all assignment points needed to take the exam.

Organizational

Meta

- Course number. L.079.05501
- Instructors. Henning Wachsmuth (lectures), Milad Alshomary (tutorials)
- Languages. English, Python

Tasks

- Six assignments. Bi-weekly; ∼50% written, ∼50% programming.
  First one published on October 17; to be submitted on October 27, 23:59 (UTC+1).
- Exam. Oral (tentatively!!). First round in February.
  Course achievement: 50%+ of all assignment points needed to take the exam.

Web page (visit frequently!)

- `http://cs.upb.de/css/teaching/courses/text-mining-w19`

Need for consultation?

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

Organizational

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

Four registrations needed

- Module + course. Both until Oct 25, 2019
- Course achievement. Oct 21 – Nov 21, 2019

  Cancellation until January 17, 2020
- Examination. Oct 21 – Nov 21 (phase 1) and Mar 2–6, 2020 (phase 2).

  Cancellation until one week before examination takes place

Organizational

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

Four registrations needed

- Module + course. Both until Oct 25, 2019
- Course achievement. Oct 21 – Nov 21, 2019

  Cancellation until January 17, 2020
- Examination. Oct 21 – Nov 21 (phase 1) and Mar 2–6, 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.

Organizational

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

Four registrations needed

- Module + course. Both until Oct 25, 2019
- Course achievement. Oct 21 – Nov 21, 2019

  Cancellation until January 17, 2020
- Examination. Oct 21 – Nov 21 (phase 1) and Mar 2–6, 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/`

Organizational

Lectures and Tutorials

Dates and locations

- Lectures. Thursday 11–14, as of October 10, in F1.110

  Tentative: No lecture on November 14
- Tutorials. Wednesday 16–18, as of October 16, in F1.110

  First tutorial introduces Python and clarifies the assignment concept.

Organizational

Lectures and Tutorials

Dates and locations

- Lectures. Thursday 11–14, as of October 10, in F1.110
  Tentative: No lecture on November 14
- Tutorials. Wednesday 16–18, as of October 16, in F1.110
  First tutorial introduces Python and clarifies the assignment concept.

Three lecture time options

1. Early. Start 11:00, end 13:30, 15 minutes break
   Mensa-friendly, not campus-friendly, attention-friendly
2. Late. Start 11:15, end 13:45, 15 minutes break
   Not mensa-friendly, campus-friendly, attention-friendly
3. Tough. Start 11:15, end 13:30, no break
   Mensa-friendly, campus-friendly, not attention-friendly

Organizational

Lectures and Tutorials

Dates and locations

- Lectures. Thursday 11–14, as of October 10, in F1.110
  Tentative: No lecture on November 14
- Tutorials. Wednesday 16–18, as of October 16, in F1.110
  First tutorial introduces Python and clarifies the assignment concept.

Three lecture time options

1. Early. Start 11:00, end 13:30, 15 minutes break
   Mensa-friendly, not campus-friendly, attention-friendly
2. Late. Start 11:15, end 13:45, 15 minutes break
   Not mensa-friendly, campus-friendly, attention-friendly
3. Tough. Start 11:15, end 13:30, no break
   Mensa-friendly, campus-friendly, not attention-friendly

Chosen option (based on discussion in the lecture)

- Late. Start 11:15, end 13:45, 15 minutes break

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 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.

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.

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 experiments.
- Machine learning. Fundamental concepts and learning methods.
- Programming. Implementation in Python.

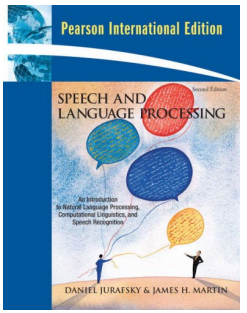
  Python mostly covered in the tutorials only.

Textbooks (Not Mandatory)

Daniel Jurafsky and James H. Martin (2009).

Speech and Language Processing.

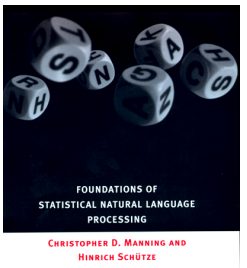
- Oriented towards computational linguistics
- Comprehensive
- Draft of 3rd ed.: http://web.stanford.edu/~jurafsky/slp3



Christoper D. Manning and Hinrich Schütze (1999).

Foundations of Statistical Natural Language Processing.

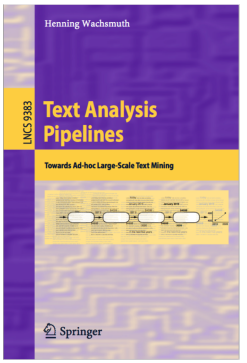
- More oriented towards computer science
- Comprehensive, a bit outdated



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 Methods

V. Text Mining using Grammars

VI. Basics of Machine Learning

VII. Text Mining using Similarities and Clustering

VIII. Text Mining using Classification and Regression

IX. Text Mining using Sequence Labeling

X. Practical Issues