Big Data and Language Technologies

webis.de

Contents

Objectives

- ❑ understand and explain the basic concepts of current machine learning models for language processing, understanding, and generation

- ❑ gain insights into the tool landscape for big data and AI-based language technologies

- ❑ work on a small research problem in language technology

- ❑ practice scientific work, writing & presentation

- ❑ get hands-on experience with cutting-edge tools

Course Organization

❑ Lectures

 – Understand theoretical foundations

❑ Labs (starting today)

 – Learn implementation skills, focus on deep learning with Python

❑ Group project (starting in a few weeks)

 – Apply learnings to a research problem

Course Deliverables

1. Active participation

2. Course project implementation

3. Project exposé & work plan (1-2 pages)

4. Mid-term presentation (5min)

5. Final report (4 pages double column + references)

Course Projects
What to Expect

- $\approx \frac{1}{2}$ semester, small groups (2-3 people)

- Focus on practical realization

- Some topic ideas (details to follow)
  - Large-scale web data analytics pipelines
  - Website classification & template induction
  - Large Language Model benchmarking OR constrained generation OR fine-tuning
  - Language usage analysis
  - Text reuse detection
  - Source code retrieval OR malware detection

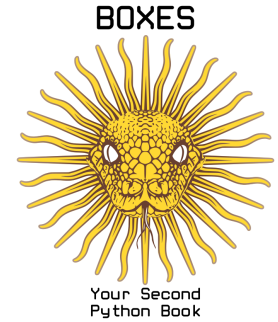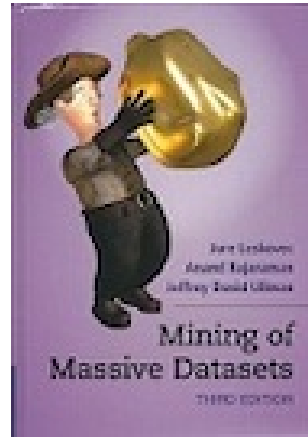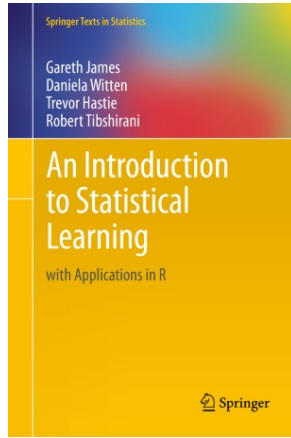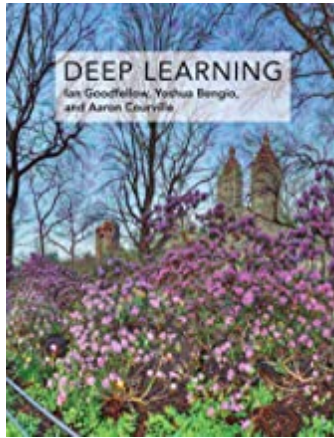- ...OR propose your own idea!

Course Prerequisites
What we Hope you Already Know. . .

❑ Good Python experience (or expert in another language and willing to self-teach)

❑ Prior exposure to machine learning basics

❑ Comfortable working with Linux, on the command line

❑ Comfortable using commandline tools like SSH, git, tmux/screen

❑ Basic understanding of algorithms, file systems, networking, . . . in that context

Course Prerequisites
. . . But if you Don't, Start Here

[deeplearningbook.org] [statlearning.com] [mmds.org] [linuxcommand.org] [ralsina.gitlab.io/boxes-book]

[neuralnetworksanddeeplearning.com] [webis.de/lecturenotes.html#machine-learning]