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
- Prompt Engineering Mini-Project ($\approx$ week 7)
  - Explore zero-shot capabilities of Large Language Models
- Group project ($\approx$ week 9, until semester end)
  - Apply learnings to a research problem

Course Deliverables

What You'll Need to Do

1. Active participation
2. Course project implementation
3. Project exposé & work plan (1-2 pages)
4. Mid-term presentation (5min)
5. Final report ($\geq$4 pages double column + references)

Course Projects

What to Expect

- $\approx \frac{1}{2}$ semester, small groups (2-3 people)
  Workload: 10 ECTS (Leipzig) or 6 ECTS (Weimar)

- 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!

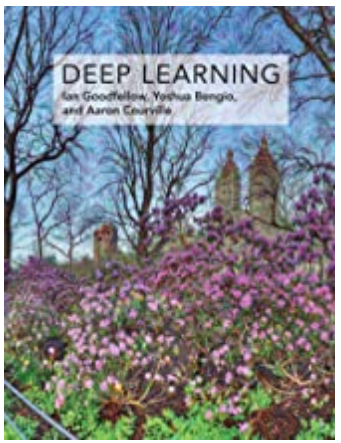
Course Prerequisites

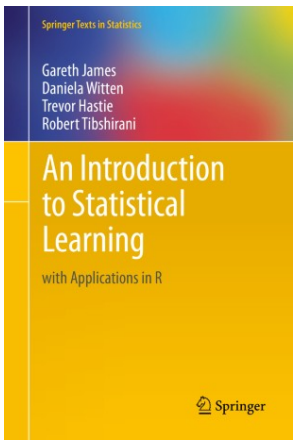
What we Hope you Already Know...

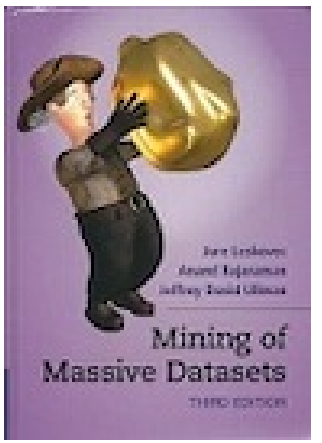
- Good Python skills (or expert in another language & willing to self-teach)
- Prior exposure to machine learning basics
- Comfortable working with Linux, on the command line
- Comfortable using commandline tools like SSH, git, tmux/screen
- Basic understanding of algorithms, file systems, networking, ...

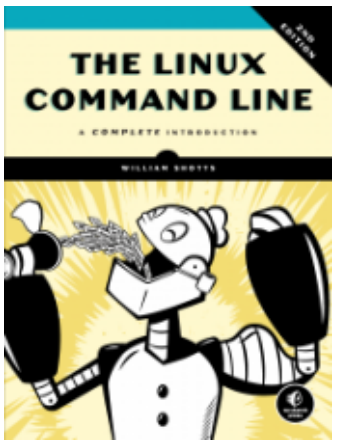
Course Prerequisites

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







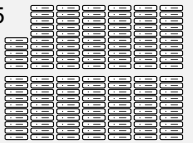
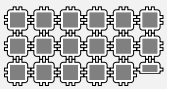
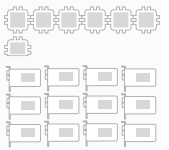
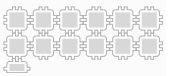
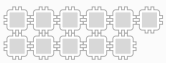
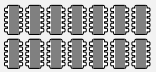


[deeplearningbook.org] [statlearning.com] [mmds.org] [linuxcommand.org] [ralsina.gitlab.io/boxes-book]

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

Course Facilities

Compute Clusters

	α-web [2009]	β-web [2015]	γ-web [2016 + 2021]	δ-web [2018]	ε-web [2020]
Nodes	44 	135 	9 	78 	55 
Disk [PB]	0.2 	4.1 	0.08 	12 	0.1 
Cores	176  ≅ 3.2 TFLOPs	1,740  ≅ 67.4 TFLOPs	672 + 227,328  ≅ 8 PFLOPs	1,248  ≅ 119.8 TFLOPs	1,100  ≅ 44 TFLOPs
RAM [TB]	0.8 	28 	7.5 	10 	7 

Course Facilities

Compute Clusters

	α-web [2009]	β-web [2015]	γ-web [2016 + 2021]	δ-web [2018]	ε-web [2020]
Nodes	44	135	9	78	55
Disk [PB]	0.2	4.1	0.08	12	0.1
Cores	176 ≅ 3.2 TFLOPs	1,740 ≅ 67.4 TFLOPs	672 + 227,328 ≅ 8 PFLOPs	1,248 ≅ 119.8 TFLOPs	1,100 ≅ 44 TFLOPs
RAM [TB]	0.8	28	7.5	10	7

Course Facilities

Compute Clusters

	α-web [2009]	β-web [2015]	γ-web [2016 + 2021]	δ-web [2018]	ε-web [2020]
Nodes	44	135	9	78	55
Disk [PB]	0.2	4.1	0.08	12	0.1
Cores	176	1,740	672 + 227,328	1,248	1,100
	≅ 3.2 TFLOPs	≅ 67.4 TFLOPs	≅ 8 PFLOPs	≅ 119.8 TFLOPs	≅ 44 TFLOPs
RAM [TB]	0.8	28	7.5	10	7