Big Data and Language Technologies
Contents

I. Course Organization

II. Introduction
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 (≈ week 7)
  - Explore zero-shot capabilities of Large Language Models

- Group project (≈ 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 (≥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]
## Compute Clusters

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nodes</strong></td>
<td>44</td>
<td>135</td>
<td>9</td>
<td>78</td>
<td>55</td>
</tr>
<tr>
<td><strong>Disk [PB]</strong></td>
<td>0.2</td>
<td>4.1</td>
<td>0.08</td>
<td>12</td>
<td>0.1</td>
</tr>
<tr>
<td><strong>Cores</strong></td>
<td>176</td>
<td>1,740</td>
<td>672 + 227,328</td>
<td>1,248</td>
<td>1,100</td>
</tr>
<tr>
<td><strong>RAM [TB]</strong></td>
<td>0.8</td>
<td>28</td>
<td>7.5</td>
<td>10</td>
<td>7</td>
</tr>
</tbody>
</table>

- α-web [2009]:
  - Nodes: 44
  - RAM [TB]: 0.8
  - Cores: 176
  - Disk [PB]: 0.2
  - Equivalent: $\cong 3.2$ TFLOPs

- β-web [2015]:
  - Nodes: 135
  - RAM [TB]: 28
  - Cores: 1,740
  - Disk [PB]: 4.1
  - Equivalent: $\cong 67.4$ TFLOPs

- γ-web [2016 + 2021]:
  - Nodes: 9
  - RAM [TB]: 7.5
  - Cores: 672
  - Disk [PB]: 0.08
  - Equivalent: $\cong 8$ PFLOPs

- δ-web [2018]:
  - Nodes: 78
  - RAM [TB]: 10
  - Cores: 1,248
  - Disk [PB]: 12
  - Equivalent: $\cong 119.8$ TFLOPs

- ε-web [2020]:
  - Nodes: 55
  - RAM [TB]: 7
  - Cores: 1,100
  - Disk [PB]: 0.1
  - Equivalent: $\cong 44$ TFLOPs
## Course Facilities

### Compute Clusters

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
<td>44</td>
<td>135</td>
<td>9</td>
<td>78</td>
<td>55</td>
</tr>
<tr>
<td>Disk [PB]</td>
<td>0.2</td>
<td>4.1</td>
<td>0.08</td>
<td>12</td>
<td>0.1</td>
</tr>
<tr>
<td>Cores</td>
<td>176</td>
<td>1,740</td>
<td>672 + 227,328</td>
<td>1,248</td>
<td>1,100</td>
</tr>
<tr>
<td>RAM [TB]</td>
<td>0.8</td>
<td>28</td>
<td>7.5</td>
<td>10</td>
<td>7</td>
</tr>
</tbody>
</table>

\[\begin{align*}
\text{α-web [2009]} & \quad 44 \quad \text{Nodes} & \quad 0.2 \quad \text{Disk [PB]} & \quad 176 \quad \text{Cores} & \quad 0.8 \quad \text{RAM [TB]} \\
\text{β-web [2015]} & \quad 135 \quad \text{Nodes} & \quad 4.1 \quad \text{Disk [PB]} & \quad 1,740 \quad \text{Cores} & \quad 28 \quad \text{RAM [TB]} \\
\text{γ-web [2016 + 2021]} & \quad 9 \quad \text{Nodes} & \quad 0.08 \quad \text{Disk [PB]} & \quad 672 \quad \text{Cores} & \quad 7.5 \quad \text{RAM [TB]} \\
\text{δ-web [2018]} & \quad 78 \quad \text{Nodes} & \quad 12 \quad \text{Disk [PB]} & \quad 1,248 \quad \text{Cores} & \quad 10 \quad \text{RAM [TB]} \\
\text{ε-web [2020]} & \quad 55 \quad \text{Nodes} & \quad 0.1 \quad \text{Disk [PB]} & \quad 1,100 \quad \text{Cores} & \quad 7 \quad \text{RAM [TB]} \\
\end{align*}\]
## Course Facilities

### Compute Clusters

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Nodes</th>
<th>RAM [TB]</th>
<th>Cores</th>
<th>Disk [PB]</th>
<th>FLOPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>α-web [2009]</td>
<td>44</td>
<td>0.8</td>
<td>176</td>
<td>0.2</td>
<td>≅ 3.2 TFLOPs</td>
</tr>
<tr>
<td>β-web [2015]</td>
<td>135</td>
<td>4.1</td>
<td>1,740</td>
<td>4.1</td>
<td>≅ 67.4 TFLOPs</td>
</tr>
<tr>
<td>γ-web [2016 + 2021]</td>
<td>9</td>
<td>0.08</td>
<td>672</td>
<td>0.08</td>
<td>≅ 8 PFLOPs</td>
</tr>
<tr>
<td>δ-web [2018]</td>
<td>78</td>
<td>12</td>
<td>1,248</td>
<td>12</td>
<td>≅ 119.8 TFLOPs</td>
</tr>
<tr>
<td>ε-web [2020]</td>
<td>55</td>
<td>0.1</td>
<td>1,100</td>
<td>0.1</td>
<td>≅ 44 TFLOPs</td>
</tr>
</tbody>
</table>