Seminar CS715

Large-Scale Data Integration

Data Integration using Deep Learning
Hallo

- **Prof. Dr. Christian Bizer**
- Professor for Information Systems V
- **Research Interests:**
  - Web Data Integration
  - Data and Web Mining
  - Data Web Technologies
- **Room:** B6 - B1.15
- **eMail:** christian.bizer@uni-mannheim.de
- **Consultation:** Wednesday, 13:30-14:30
Hallo

- M. Sc. Wi-Inf. Ralph Peeters
- Graduate Research Associate
- Research Interests:
  - Entity Matching using Deep Learning
  - Product Data Integration
- Room: B6, 26, C 1.04
- eMail: ralph.peeters@uni-mannheim.de
Hallo

- M. Sc. Wi-Inf. Alexander Brinkmann
- Graduate Research Associate
- Research Interests:
  - Data Search using Deep Learning
  - Product Data Categorization
- Room: B6, 26, C 1.04
- eMail: alexander.brinkmann@uni-mannheim.de
Hallo

- **M. Sc. Wi-Inf. Keti Korini**
- Graduate Research Associate
- Research Interests:
  - Table Annotation using Deep Learning
  - Schema Matching
- Room: B6, 26, C 1.03
- eMail: kkorini@uni-mannheim.de
You and Your Experience

- A Short Round of Introductions
  - What are you studying?
  - Which DWS courses did you attend?
  - What kind of experience do you have with Data Science/Data Engineering projects?

- Participants
  - Vu, Tuananh
  - Jansen, Tim
  - Der, Reng
  - Elzamarany, Ahmed
  - Pradeep Kumar, Sreehari
  - Zong, Xu
Agenda of Today‘s Kickoff Meeting

1. Seminar organization
2. Seminar topics
3. How to structure your seminar paper / presentation?
4. Your Questions
1. Organization
Learning Targets

- Writing a seminar thesis as an exercise for your master thesis
- Understanding and presenting state-of-the-art scientific work
- Searching and citing scientific papers / journal articles
- How to structure your thesis and presentation
- How to argue, how to explain, how to write!
- How to write a nicely formatted paper using LaTeX
<table>
<thead>
<tr>
<th>Date</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Wednesday, 14.09.2022</strong></td>
<td>Kick-off meeting and topic assignment</td>
</tr>
<tr>
<td>(10:00-11:30)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Read papers about your topic and search for additional literature</td>
</tr>
<tr>
<td></td>
<td>Prepare outline and argumentation line for the presentation</td>
</tr>
<tr>
<td><strong>Until 7.10.2022</strong></td>
<td>Meet with your mentor to discuss your presentation</td>
</tr>
<tr>
<td></td>
<td>Prepare draft of your presentation</td>
</tr>
<tr>
<td><strong>Until 28.10.2022</strong></td>
<td>Send draft presentation to your mentor</td>
</tr>
<tr>
<td></td>
<td>Finalize your presentation</td>
</tr>
<tr>
<td><strong>Friday, 18.11.2022</strong></td>
<td>Presentation and discussion of your topic</td>
</tr>
<tr>
<td>(10:00-12:30)</td>
<td>(30 % of your final grade)</td>
</tr>
<tr>
<td></td>
<td>Write seminar thesis</td>
</tr>
<tr>
<td><strong>Sunday, 22.01.2023</strong></td>
<td>Submission of your seminar thesis (70 % of your final grade)</td>
</tr>
</tbody>
</table>
Formal Requirements

- Presentation
  - 12 minutes + 8 minutes discussion
  - should be 100% understandable for all participants

- Written report (paper)
  - 12-15 pages single column
    - including abstract and appendixes
    - not including bibliography
    - every additional page reduces your grade by 0.3
  - written in English language
  - use latex template of Springer Computer Science Proceedings

- Final grade
  - 70% written report
  - 30% presentation
Which template to use?

http://www.springer.com/de/it-informatik/lncs/conference-proceedings-guidelines
2. Seminar Topics
Motivation and Overview

Symbolic Representations $\rightarrow$ Distributed Representations

Traditional Machine Learning $\rightarrow$ Neural Nets

1. Entity Matching using Deep Learning

- Student: Elzamarany, Ahmed
- Mentor: Ralph Peeters


- More references and benchmarks: [Papers with Code: Entity Resolution](https://paperswithcode.com/task/entity-resolution)
Topics

2. Entity Matching using Contrastive Learning

- Student: Pradeep Kumar, Sreehari
- Mentor: Ralph Peeters

3. Active Learning for Entity Matching

- Student: Der, Reng
- Mentor: Ralph Peeters


- More references and benchmark results: [Papers with Code: MusicBrainz20K](https://paperswithcode.com)
4. Representation Learning for Tabular Data

- Student: Vu, Tuananh

- Mentor: Keti Korini


Topics

5. Column Type Annotation in Tabular Data

- Student: Jansen, Tim
- Mentor: Keti Korini


- More references and benchmarks: Papers with Code: Column Type Annotation
6. Information Extraction for E-Commerce Product Data

- Student: Zong, Xu
- Mentor: Alexander Brinkmann


3. How to Structure Your Paper / Presentation
Goal of Seminar Paper

A seminar paper differs significantly from a master thesis

- The topic is already defined
- No need to implement or develop algorithms
- No need to perform experiments
- Primarily: reproduction and re-organization of content that is already available

Goals of seminar paper

1. Describe the problem / task
2. describe several existing methods/systems for handling the task,
3. compare the methods/systems and their evaluation using a systematic set of comparison criteria
How to Structure Your Paper?

1. Introduction and Problem Statement
   - Which problem/task is addressed? Why is the problem important?
   - Structure of your paper

2. Description of Existing Approaches
   - Overview of existing methods and features used by the methods
   - Detailed description of selected methods
   - Comparison of the selected methods using a set of comparison criteria

3. Evaluation
   - Comparison and discussion of the evaluation tasks, metrics
   - Comparison of the evaluation results

4. Conclusion
   - What did the comparison of the methods and evaluation results show?
   - Can something be concluded for future work?

5. Bibliography
Learn from Examples

- Read survey articles and identify the structure from the previous slide
  - Why can this paragraph be found at that position?
  - What is the purpose of some section / subsection?

- Important
  - Read survey articles!
  - Read conference or journal papers.

- Textbook on how to write a thesis

- University Library: Academic Writing Consultancy
  - https://www.bib.uni-mannheim.de/en/writing-consultancy/
Citing different Types of Publications

- **Journal article**
  - Good to cite, current research results
  - Survey articles (very good for an overview)

- **Conference and workshop paper**
  - Good to cite, current research results

- **Books (sometimes cited)**
  - Textbooks
  - Collections of articles/papers => Cite specific paper in book

- **Websites**
  - better not cited, exceptions are, e.g., W3C Specifications
  - Wikipedia is not an exception!!! *Do not cite Wikipedia, ever!!*

- **Slide sets (especially from our lectures)**
  - Never cite!
How to Find Relevant Publications?

- Use Standard Search Engines
- **Use Google Scholar**
  - we use it a lot ourselves
- Search Engines of the University’s library
  - see slides from the library course
- **Exploit references**: Given a relevant document $x$
  - Follow references in the past: papers $y$ that $x$ has cited
  - Follow references in the future: papers $y$ that cited $x$ ("cited by" functionality in Google scholar)
4. Questions?