Seminar

Selected Topics in Data Science

Prof. Dr. Bela Gipp
Dr. Norman Meuschke,
Dr. Terry Ruas
Members of the GippLab team

Agenda

Overview of the research group

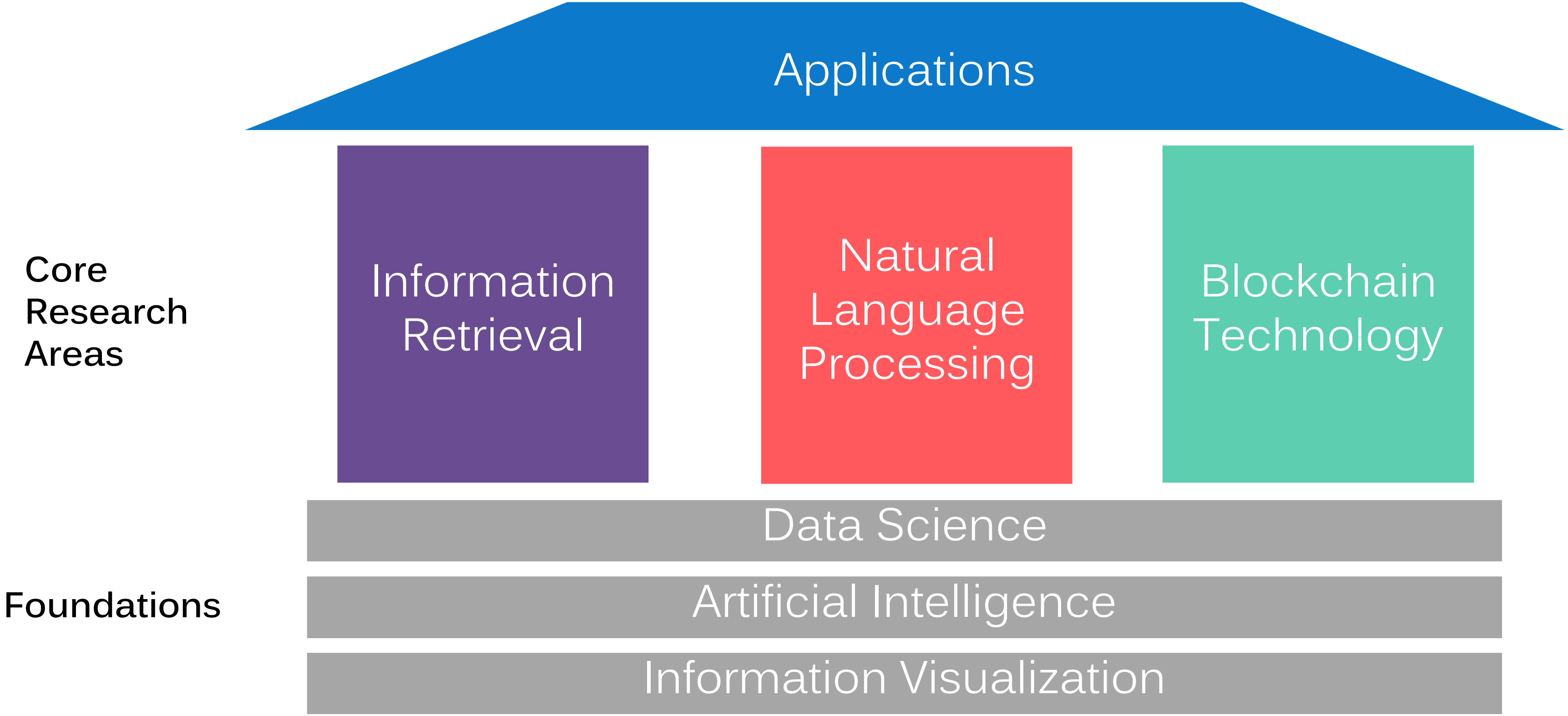
Organization of the seminar

Presentation of seminar projects

GippLab
Sci. Info. Analytics

www.gipplab.org/team

Our Research Areas



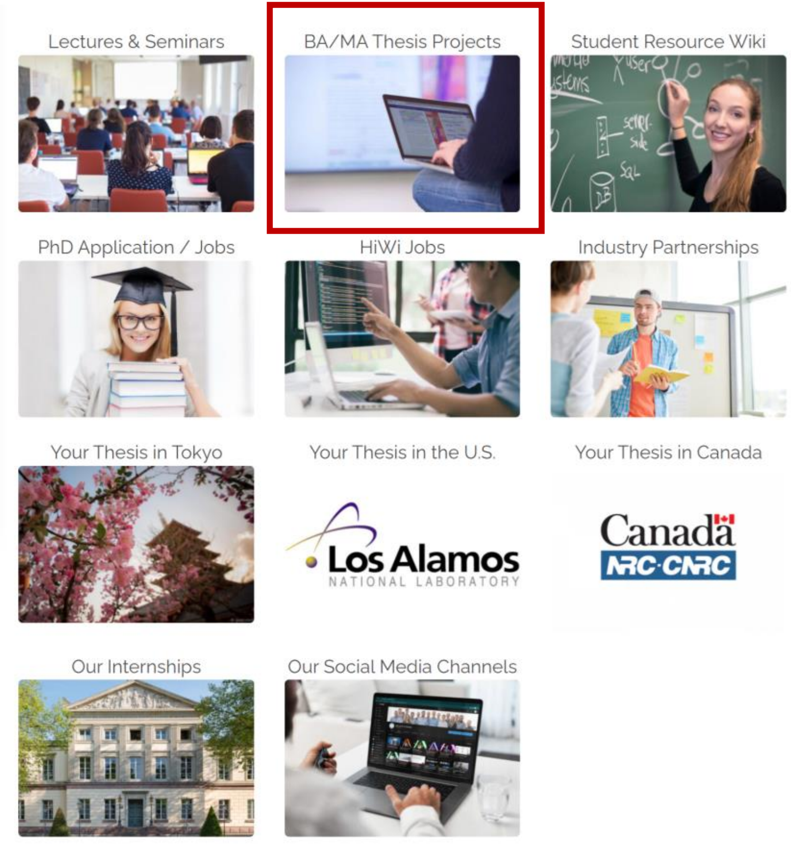
Concept of our Courses

1. Lecture
   - Spark interest in the research of our group
   - Give overview of relevant technologies
   - Teach essential skills

2. Seminar
   - Acquire in-depth knowledge of current research
   - Train methodological skills
   - Prepare potential project & thesis

3. Project & Thesis
   - Make a small, but meaningful and lasting improvement to the state of the art

Thesis at GippLab



www.gipplab.org/students-corner

Overview of the Seminar

Goals of this seminar

1. **In-depth Knowledge of Current Research**
    - Overview of state-of-the-art technologies
    - Current research trends
    - Research challenges
    - Open problems (potential BA/MA projects)

2. **Methodological Skills**
    - Find & organize literature
    - Systematically read research papers
    - Analyse, compare and contrast research
    - Determine and build upon most suitable methods and approaches
    - Structure, write, and format an academic paper
    - Presentation skills
    - Discuss work with peers

How…

- Define your research question / specific research topic

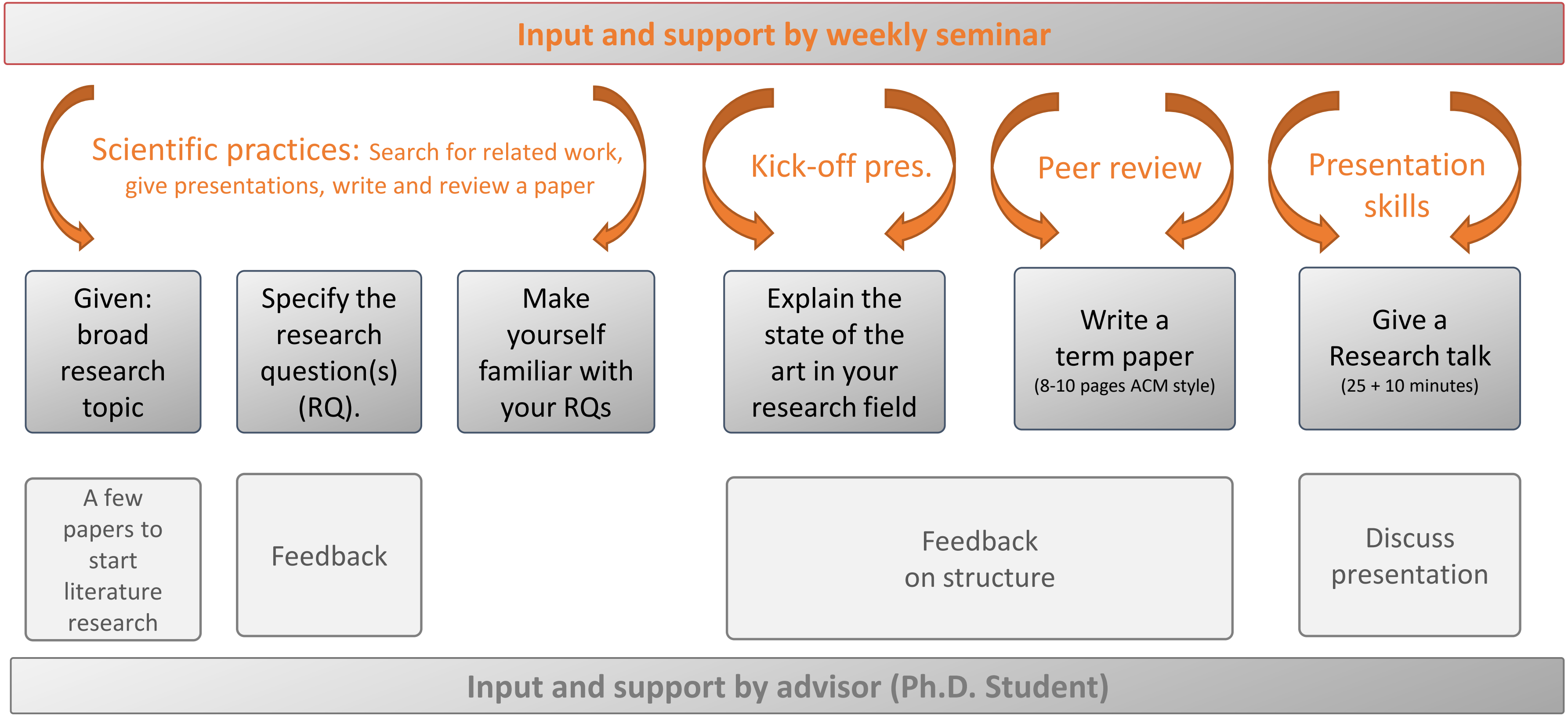
- Independent research of literature or technologies

- Consultations & peer feedback

- **Theoretical track** (term paper with a state-of-the-art literature review) vs. **Applied track** (developed project with a research report)

- In-classroom presentations

Structure of the **theoretical track** (term paper)



Input and support by weekly seminar

Scientific practices: Search for related work, give presentations, write and review a paper

Kick-off pres.

Peer review

Presentation skills

Given: broad research topic

Specify the research question(s) (RQ).

Make yourself familiar with your RQs

Explain the state of the art in your research field

Write a term paper (8-10 pages ACM style)

Give a Research talk (25 + 10 minutes)

A few papers to start literature research

Feedback

Feedback on structure

Discuss presentation

Input and support by advisor (Ph.D. Student)

Credit requirements and grading – **theoretical track**

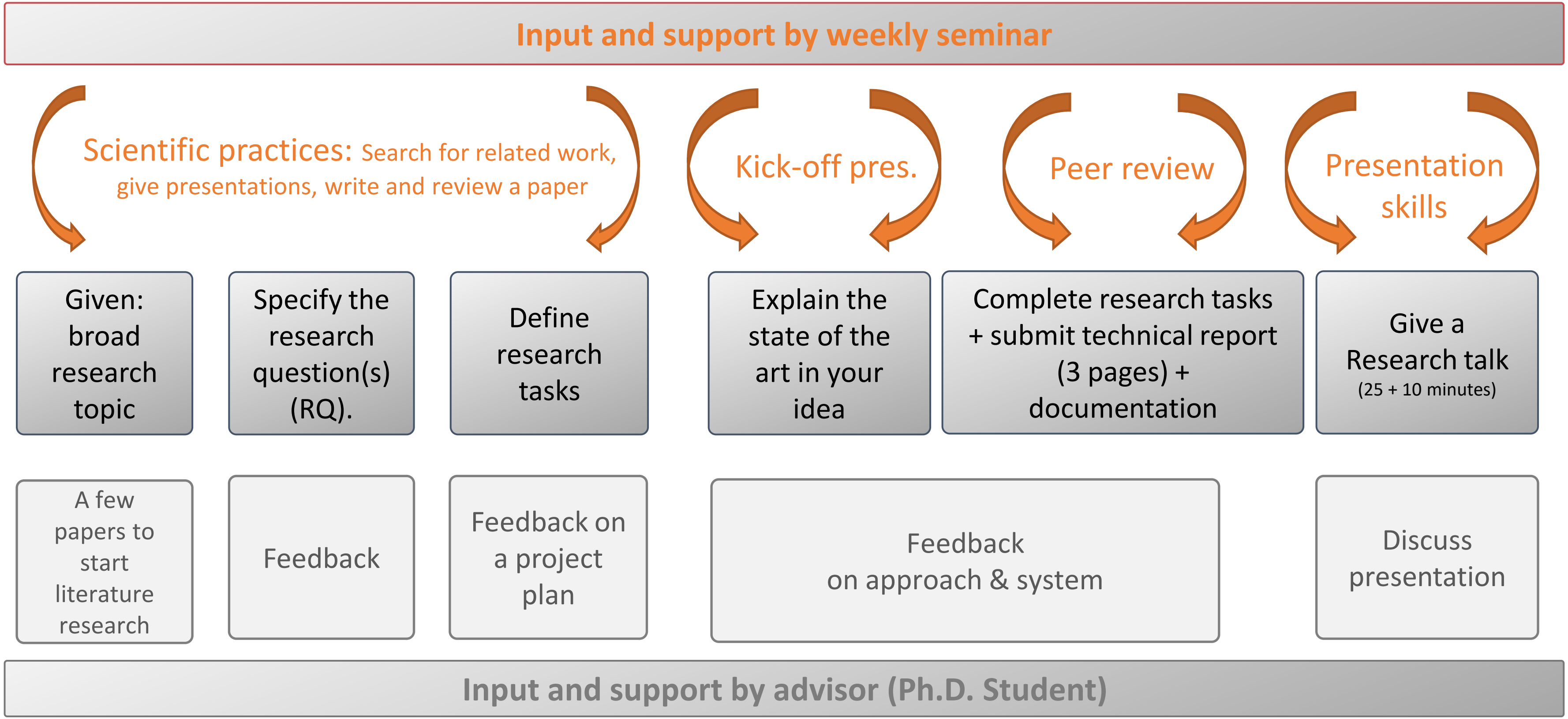
1. Term Paper (8-10 pages ACM style)
   - First draft (min 4 pages) → **10%**
   - Second draft (min 6 pages) → **20%**
   - Final paper → **35%**

2. Two Presentations
   - Teaser (5 min per topic) → **5%**
   - Full presentation (25 min + 10 min discussion) → **20%**

3. Review of other paper → **10%**

4. Attendance and Participation
   - Attendance is mandatory, missing one session is OK
   - Engagement in discussion → **bonus**

Structure of the **applied track** (project + report)



Input and support by weekly seminar

Scientific practices: Search for related work, give presentations, write and review a paper

Kick-off pres.

Peer review

Presentation skills

Given: broad research topic

Specify the research question(s) (RQ).

Define research tasks

Explain the state of the art in your idea

Complete research tasks + submit technical report (3 pages) + documentation

Give a Research talk (25 + 10 minutes)

A few papers to start literature research

Feedback

Feedback on a project plan

Feedback on approach & system

Discuss presentation

Input and support by advisor (Ph.D. Student)

Credit requirements and grading - **applied track**

1. Final Software Product (code + documentation, 4 pages) → **45%**

2. Two Presentations
   - Kick-off topic presentation (5 min) → **5%**
   - Full presentation (25 min + 10 min discussion) → **20%**

3. Project-specific milestones → **30%**

4. Attendance and Participation
   - Attendance is mandatory, missing one session is OK
   - Engagement in discussion → **bonus**

Seminar Topics

Legend for Project Descriptions

- 🧠 Project suitable for the **theory track of the course**
- 💻 Project suitable for the **applied track of the course**
- 🎓 Project can be extended to a BSc / MSc thesis
- NII Project suitable for BSc / MSc thesis at NII Tokyo

NII 🎓 💻 🧠

Natural Language Processing

Natural Language Processing

- Natural Language Processing is a **cross-disciplinary** research field that draws heavily from **artificial intelligence** (AI), **machine learning** (ML), mathematics, and linguistics.

- Personal assistants, recommender systems, fake news identification, financial stock analysis, chatbots, autocorrection, auto-completion, intelligent search engines, and automatic translation or captioning are just a few examples of how NLP and AI are helping us manage the flood of data. However, systems to process natural language are far from perfect, which leaves much space for research.

- Some of the areas we work are:
  - Natural language understanding
  - Paraphrase detection
  - Text summarization
  - Media bias/Fake news detection
  - Semantic analysis/extraction
  - Sentiment analysis

For a complete list of our research topics visit our website!

NLP02 CS-Insights – State of the art in Computer Science Publications

Background

DBLP is the largest open-access repository of scientific articles on computer science and provides metadata associated with publications, authors, and venues. We retrieved more than 6 million publications from DBLP and extracted pertinent metadata (e.g., abstracts, author affiliations, citations) from the publication texts to create the DBLP Discovery Dataset (D3). Now, on CS-Insights we devised a system (back- and front-end) to explore our dataset and uncover all the trends regarding computer science publications. As CS-Insights is an ongoing project we need to fix it's open issues and extend its functionalities.

Goal

- Solve existing issues in CS-Insights-Roadmap

Tasks

- Work on project roadmap for CS-Insights
  - Backlog and additional features
- Propose extension for CS-Insights
  - Authors features (e.g., h-index)

Jan Philip Wahle

wahle@gipplab.org

Terry L. Ruas

ruas@gipplab.org

NLP03 Identification of UN-Sustainable Development Goals (SDGs)

Background

In 2015, the United Nations Member States (UN) agreed on 17 sustainable development goals (SDG) *„for peace and prosperity for people and the planet, now and into the future“.*

We want to support this endeavor by clustering research publications based on their relevance to these SDGs. The goal is to make relevant research findable and to strengthen access and visibility to such research.

SUSTAINABLE DEVELOPMENT GOALS
1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS

Goal

- Evaluation of the existing approach and enhancement of implementation

Tasks

1. Implementation
   - Refactoring of existing R code
   - Build a Python implementation
2. Data science
   - Evaluation of used methods
   - Enlarge data set and compare results

Daniel Kurzawe

kurzawe@sub.uni-goettignen.de

Mathias Göbel

goebel@sub.uni-goettingen.de

NLP04 Information Extraction from Research Papers for DIGIS

Background

The objective is to devise approaches for the automated extraction of geochemical data and metadata from research papers and implement them prototypically pipeline for the geochemical data infrastructure DIGIS.

We extract specific mentions of methods from papers. This information can be part of the paper or included in tables or figures. The structure depends on the journal.

DIGIS

Digital Geochemistry Infrastructure for GEOROC 2.0

Goal

- A protoype for extraction specific information from research papers

Tasks

- Compare existing approaches for information extraction for a given set of papers
- Implement a prototype
- Draft a data pipeline

Daniel Kurzawe

kurzawe@sub.uni-goettignen.de

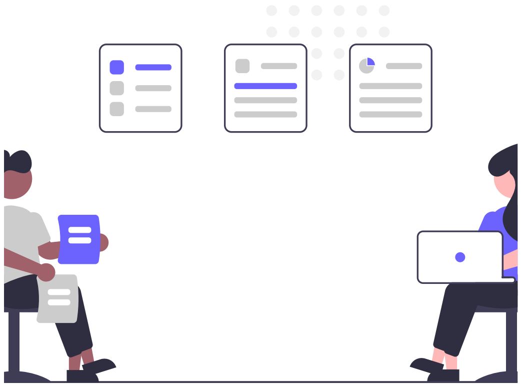
Mathias Göbel

goebel@sub.uni-goettingen.de

NLP07 Meeting Summarization System Testbench

Background

The field of natural language processing has seen a significant amount of research in recent years on the task of meeting summarization. With the increasing availability of meeting transcripts, there is a growing need for efficient methods to automatically summarize the content of these meetings. As of now, due to the different formats of meetings and the dynamic, idiosyncratic nature, many domain- and problem-specific techniques have been introduced. However, the area lacks a standardized benchmark for evaluating these methods. Thus, it is difficult to compare and identify the strengths and weaknesses of the individual techniques.



Goal

- Design and develop a unified framework to test meeting summarization techniques (evaluation harness).

Tasks

- Design a solution to transform any kind of input format for models and datasets into one common form
- Implement a general applicable evaluation environment to test different models, datasets and metrics simultaneously
- Evaluate the most prominent techniques

Frederic Kirstein

kirstein@gipplab.org

Terry L. Ruas

ruas@gipplab.org

NLP08 Multi-Source Meeting Summarization

Background

An increase in the number of online meetings made clear that typically meetings only have few key topics and a limited amount of relevant information for all participants. Therefore, the extraction of their key topics and their summarization became more sought after. Meetings differ from traditional text. The multi-party setting, deviant formats, idiosyncratic nature, and different semantic styles promote a complex scenario. Short meetings can easily reach thousands of tokens in just a few minutes. Thus, techniques that produce high quality summaries from multiple sources (e.g., transcripts, email, chat), including the most important ideas discussed, are still necessary. For now, we seek which techniques related to the meeting summarization domain, e.g., text summarization and generation, can be adapted to meetings.

Goal

- Explore the automatic text summarization task (Extractive/Abstractive) applied to meetings [low resource languages]

Tasks

- Study which models, datasets and metrics can be used in this task (from meeting summarization directly and related domains)
- Define describing criteria for models, datasets and metrics and organize these according to the criteria (e.g., relation graph, clustering)
- Evaluate current state of the art models in a scalable process and incorporate the results into the individual descriptions / organizations

Frederic Kirstein

kirstein@gipplab.org

Terry L. Ruas

ruas@gipplab.org

NLP09 AI Solving 😃 + ❓

Background

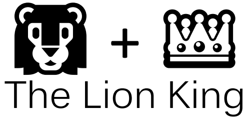
With the digital era and the need to communicate complex information quickly and concise, emojis are used to abbreviate otherwise lengthy textual descriptions. Due to higher resolutions, emojis evolved into rich and detailed graphics for a large variety of use cases. The sheer number of existing graphics motivated emoji riddles, puzzles that try to reverse engineer the intended textual meaning of a set of graphics. Due to the ambiguities of simplified graphics, these emoji riddles often require a certain degree of creativity to solve. Meanwhile, all the information necessary to solve such puzzles are generally accessible to modern algorithms. Hence the pressing question, can an AI be creative and solve emoji riddles?

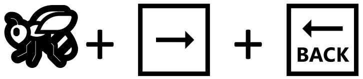
Goal

- Develop a model that converts a set of emoji graphics into textual descriptions.

Tasks

- Study which models, datasets, metrics can be used in this task.
- Develop a concept model that trains connections between graphics and textual descriptions.
- Train and evaluate the model.





Terry L. Ruas

ruas@gipplab.org

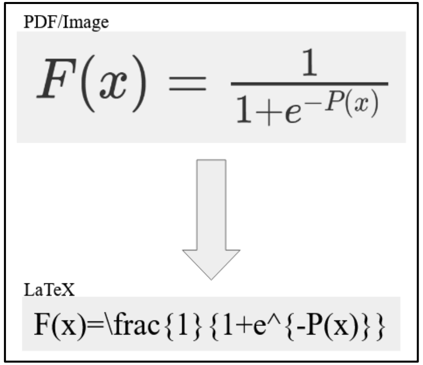
André Greiner-Petter

greinerpetter@gipplab.org

NLP11 Optical Character Recognition for math formulae (MathOCR)

Background

The extent of mathematics literature is estimated to be more than 120 million pages. However, the majority of them are not available in a machine-processable format. Hence, valuable mathematical content cannot be used by today's technologies for developing models capable of doing math-related tasks. There are MathOCR tools such as MathPix, InftyReader, etc. However, they are commercial and not evaluated for their performance. In this project, you will work on an OCR tool capable of converting math equations from PDF or image to machine-processable forms such as LaTeX or MathML. You will work on neural networks capable of detecting math symbols and the structure of math formulae. You can choose either theory or applied track for this project.



$$F(x) = \frac{1}{1+e^{-P(x)}}$$

```
F(x)=\frac{1}{1+e^{-P(x)}}
```

Goal

- Develop a MathOCR.

Tasks

- Textual and non-textual elements detection from a PDF or an Image input.
- Math symbol identification.
- Formula structure recognition using computationally inexpensive methods.
- Evaluate MathOCR methods on mathematical formulae of varied complexity.

Ankit Satpute

Ankit.Satpute@fiz-karlsruhe.de

André Greiner-Petter

greinerpetter@gipplab.org

Moritz Schubotz

Moritz.Schubotz@fiz-karlsruhe.de

NLP12 Quality Control of Optical Character Recognition (OCR)

Background

OCR results are not perfect due to poor input images or badly trained models. With the help of Ground Truth (GT) data, which are manually generated transcriptions, the quality of the OCR can be determined by comparing them with the GT and thereby the quality of the OCR processors and workflows. However, it is time-consuming and cost-intensive to generate GT data, so GT is rarely available. For the OPERANDI project, which deals with the performance and quality improvement of OCR technology based on German prints from the 16th to the 18th century (about 600,000 titles), other methods are needed in this respect, such as dictionary comparison, language models, determination of probability or character set checking.

Goal

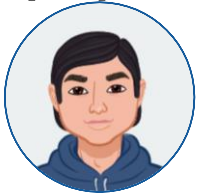
- Finding concepts and available open-source solutions for OCR quality evaluation, which do not need Ground Truth

Tasks

- Review different concepts, algorithms, and tools for quality control of OCR results
- Decide which concept, algorithm or tool would work best for OPERANDI
- Optional: implementing the best quality control solution for a provided data sample

Lilja Sautter

sautter@sub.uni-goettingen.de



Kay Liewald

liewald@sub.uni-goettingen.de

Jörg-Holger Panzer

panzer@sub.uni-goettingen.de

NLP13 Retro-Creation of Archaeological Corpora

Background

Extracting data from archaeological texts represents one of the archaeology’s most leading challenges. In recent years, Natural Language Processing has been also adopted in the archaeological domain, but we are still far away from achieving robust results.

Large corpora like the Corpus Vasorum Antiquorum (www.cvaonline.org) offer a big corpus of archaeological descriptions. These descriptions vary in length and also differ slightly depending on the author. Publication languages are mainly German and English (but also Italian and French).

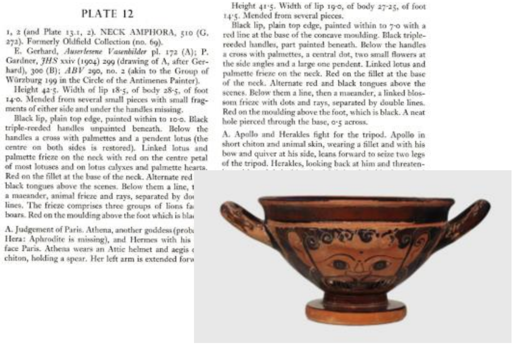


PLATE 12

1, 2 (and Plate 13.1, 2). NECK AMPHORA, 510 (G. 272). Formerly Oldfield Collection (no. 69).

E. Gerhard, *Auserlesene Vasenbilder* pl. 172 (A); P. Gardner, *JHS* xxiv (1904) 299 (drawing of A, after Gerhard), 300 (B); *ABV* 290, no. 2 (akin to the Group of Würzburg 199 in the Circle of the Antimenes Painter).

Height 42·5. Width of lip 18·5, of body 28·5, of foot 14·0. Mended from several small pieces with small fragments of either side and under the handles missing.

Black lip, plain top edge, painted within to 10·0. Black triple-reeded handles unpainted beneath. Below the handles a cross with palmettes and a pendent lotus (the centre on both sides is restored). Linked lotus and palmette frieze on the neck with red on the centre petal of most lotuses and on lotus calyxes and palmette hearts. Red on the fillet at the base of the neck. Alternate red black tongues above the scenes. Below them a line, t a maeander, animal frieze and rays, separated by do lines. The frieze comprises three groups of lions fa boars. Red on the moulding above the foot which is bla

A. Judgement of Paris. Athena, another goddess (prob Hera: Aphrodite is missing), and Hermes with his face Paris. Athena wears an Attic helmet and aegis chiton, holding a spear. Her left arm is extended forw

Height 41·5. Width of lip 19·0, of body 27·25, of foot 14·5. Mended from several pieces.

Black lip, plain top edge, painted within to 7·0 with a red line at the base of the concave moulding. Black triple-reeded handles, part painted beneath. Below the handles a cross with palmettes, a central dot, two small flowers at the side angles and a large one pendent. Linked lotus and palmette frieze on the neck. Red on the fillet at the base of the neck. Alternate red and black tongues above the scenes. Below them a line, then a maeander, a linked blossom frieze with dots and rays, separated by double lines. Red on the moulding above the foot, which is black. A neat hole pierced through the base, 0·5 across.

A. Apollo and Herakles fight for the tripod. Apollo in short chiton and animal skin, wearing a fillet and with his bow and quiver at his side, leans forward to seize two legs of the tripod. Herakles, looking back at him and threaten-

Goal

- Make archaeological data accessible in a structured way in a database. Another goal is to train the type and scope of an archaeological object description in such a way that new descriptions can be generated (f.e. with the help of computer vision)..

Tasks

- Develop a formalised vocabulary of archaeological terms
- Extract data about shape, depicted figures and objects, measurements, style and scientific discussion
- (try to) build a model that is able to describe an artefact in archaeological terms

Martin Langner

martin.langner@uni-goettingen.de

Norman Meuschke

meuschke@gipplab.org

NLP14 Analyzing Ancient Text on 3D Models

Background

3D modeling for archaeological applications is a growing field, as researchers work to better understand and analyze ancient artifacts and texts. However, there is a lack of standardized models and clear guidelines for the level of detail that should be included in these models. Additionally, there is a growing need for large collections of 3D shapes that have been annotated and organized in a meaningful way. There are already resources available for this type of work, but they need processing and organization into common formats. This project gives you the opportunity to help shape the future of this field by working with cutting-edge technology and using recent advancements in NLP to analyze and annotate these 3D models.

Goal

- Develop and promote standards in 3D modeling for analyzing natural language content on archeological findings.

Tasks

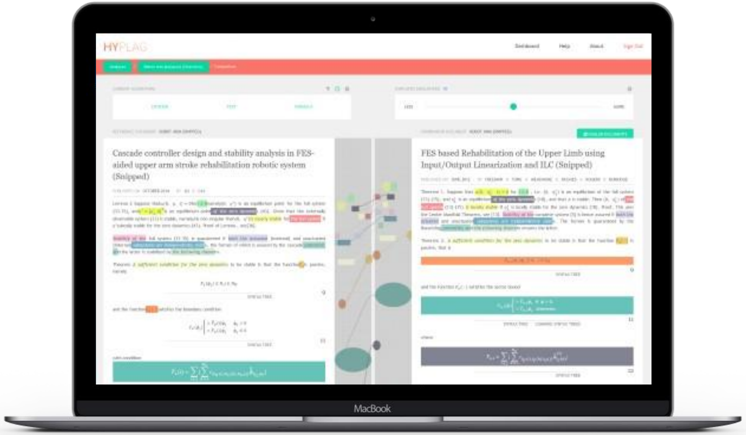
- Develop model categories and detail requirements.
- Create an annotated repository of 3D shapes and extend ShapeNet with additional synsets of WordNet.
- Provide web-based prototype for visualization and analysis.

Martin Langner

martin.langner@uni-goettingen.de

Norman Meuschke

meuschke@gipplab.org



Cascade controller design and stability analysis in FES-aided upper arm stroke rehabilitation robotic system (Snipped)

FES based Rehabilitation of the Upper Limb using Input/Output Linearization and ILC (Snipped)

Plagiarism Detection

Plagiarism Detection

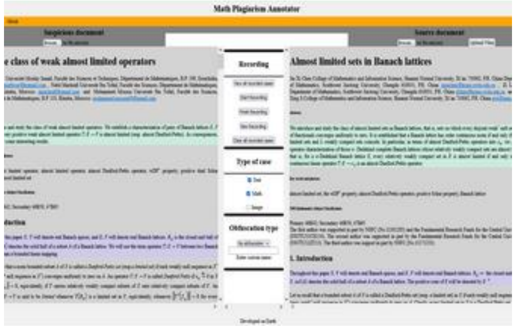
- The problem of **academic plagiarism** has been present for **centuries**.
    - The rapid and continuous advancement of **information technology** has **made plagiarizing easier** than ever.
- **Academic plagiarism** is one of the severest forms of research **misconduct** and has strong negative impacts on academia and the public.
- Plagiarized research papers impede the scientific process, e.g., by distorting the mechanisms for tracing and correcting results.
- As plagiarism detection is a multi-variable complex problem, our solutions must also be. Therefore, we tackle a myriad of sub-areas in our research projects:
    - Citation extraction
    - Image similarity
    - Mathematical-based fingerprint
    - Text analysis via semantic and syntactic similarity

A complete list of Plagiarism Detection topics visit our pages for PD and NLP!

PD01 Developing a Plagiarism Detection System

Background

Recent developments in language models and services such as chatGPT have allowed people to make legitimate-looking copies of texts without knowing the sources. Using ideas, and concepts without citing the sources could lead to plagiarism. A Plagiarism Detection System (PDS) helps in finding instances of copied elements in a document from potential source documents. In this project, you will work on developing a PDS. Specifically, you will learn about how documents are handled in a PDS and similarity in document pairs is calculated. Along with textual reuse detection, you will also get to work with the detection of non-textual reuse such as mathematical formulae, images, etc.


Math Plagiarism Annotator
Suspicious document
Source document
e class of weak almost limited operators
Recording
Almost limited sets in Banach lattices
Type of case
Text
Math
Image
Obfuscation type
1. Introduction

Goal

- Developing a Plagiarism Detection System.

Tasks

- Building a system interface to select a document under inspection and potential source documents.
- Integrate big data analytics platforms to handle a large number of documents.
- Implementing a document retrieval algorithm.
- Highlight detected reuse(potentially plagiarized) instances.

Ankit Satpute

Ankit.Satpute@
fiz-karlsruhe.de

André Greiner-Petter

greinerpetter@gipplab.org

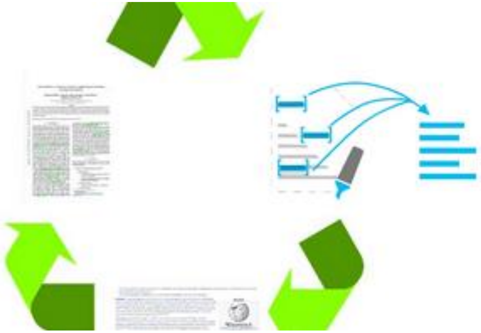
Moritz Schubotz

Moritz.Schubotz@
fiz-karlsruhe.de

PD02 Do Plagiarism Detection Systems Really Detect Plagiarism?

Background

Existing plagiarism detection systems (PDS) detects slightly altered copies of the text. However, plagiarism occurring in scientific documents is highly disguised. In this project, you will evaluate if the existing PDS works on naturally occurring cases of plagiarism or not. There exist artificial corpora of plagiarism such as PAN but they don’t represent naturally occurring cases of plagiarism because the plagiarism is artificially created. Hence, you will utilize corpora with naturally occurring cases of plagiarism such as Vroniplag, Dissernet, etc. Eventually, you would build a PDS of tomorrow that detects highly disguised cases.



Goal

- Evaluating plagiarism detection systems (open source) on naturally occurring cases of plagiarism.

Tasks

- Studying plagiarism detection approaches.
- Using corpora representing naturally occurring cases of plagiarism as test cases.
- Recording character positions of detected reuse.
- Working on non-textual reuse detection approaches.

Ankit Satpute

Ankit.Satpute@
fiz-karlsruhe.de

André Greiner-Petter

greinerpetter@gipplab.org

Moritz Schubotz

Moritz.Schubotz@
fiz-karlsruhe.de

PD05 Identifying Plagiarism of ChatGPT

Background

With the advent of advanced AI-powered language models like ChatGPT, the threat of machine-paraphrased plagiarism has become a serious concern. These models can generate text that is virtually indistinguishable from human writing, making it easy for individuals to commit plagiarism, but difficult for existing systems to detect. As these models become more accessible and widely adopted, the problem of plagiarism is expected to escalate, making it imperative for research institutions, publishers, and schools to have robust automated solutions in place.

Goal

- The goal of this project is to identify plagiarism of ChatGPT and other AI models in written works automatically.

Tasks

- Research and understand the inner workings of ChatGPT and other AI language models.
- Develop a method/training architecture for detecting generated and plagiarized text
- Evaluate the performance of the tool using quantitative and qualitative assessments.

Jan Philip Wahle

wahle@gipplab.org

Terry L. Ruas

ruas@gipplab.org

PD06 Stylometric Features of AI-generated Text

Background

Whereas extrinsic plagiarism detection aims at the identification of potential sources of plagiarism (similar documents within a database), intrinsic plagiarism detection looks for dissimilarities within one document. The goal of the intrinsic approach is to identify parts with different writing styles that may indicate different authors or different sources that should be investigated. One of these sources may be ChatGPT or other AI-based text generators. A typical intrinsic approach constitutes of identification of stylometric features, which indicate writing style.

Goal

- Investigate stylometric features of AI-generated text, and to what extent these features depend on prompt.

Tasks

- Create a corpus of documents (human written, AI-generated, machine-translated, machine-paraphrased)
- Investigate stylometric features (web and command line app available)
- Explore possibilities of Chat GPT to vary stylometric features of generated text
- Discuss identifiability of AI-generated text via stylometric features

Tomáš Foltýnek

foltynek@fi.muni.cz

Norman Meuschke

meuschke@gipplab.org

PD07 Textual Criticism and Plagiarism

Background

“The identification of textual variants, or different versions, of either manuscripts or of printed books” (Wikipedia) is a major task in philology entitled “textual criticism”. The analysis of a single text in different variants starting from the very first sketch up to the latest authorized version is provided with a historical-critical edition. Before the digital age, these editions used an obnoxious amount of signs marking and categorizing these differences, like the Leiden convention has standardized. However, newer visualization technologies provide more and more interactive views to these editions.

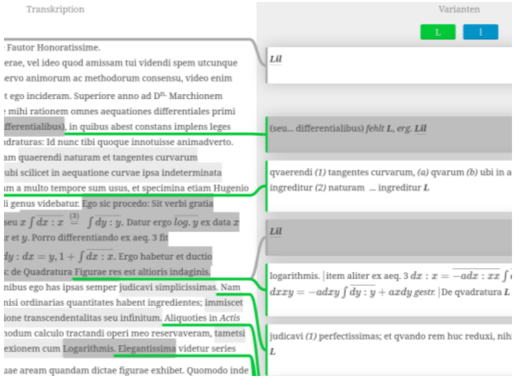
What are the shared approaches of plagiarism detection and textual criticism? Can they benefit from each other?

Goal

- Investigate if/how a software for plagiarism detection can be utilized to deal with a set of documents that represent the same text. Comparison of both methods.

Tasks

- Input material selection (assisted)
- Data conversion
- Usage of PD software/visualization, publish result using web technology
- Workflow to automatize main steps and to scale up (deal with as many editions as possible)



Fautor Honoratissime.
erae, vel ideo quod amissam tui videndi spem utcunque
ervo animorum ac methodorum consensu, video enim
t ego incideram. Superiore anno ad D$^{n.}$ Marchionem
mihi rationem omnes aequationes differentiales primi
fferentialibus), in quibus abest constans implens leges
idraturas: Id nunc tibi quoque innotuisse animadverto.
am quaerendi naturam et tangentes curvarum
ubi scilicet in aequatione curvae ipsa indeterminata
im a multo tempore sum usus, et specimina etiam Hugenio
li genus videbatur. Ego sic procedo: Sit verbi gratia
seu $x\int\overline{dx:x} \overset{(3)}{=} \int\overline{dy:y}$. Datur ergo $\overline{log.\,y}$ ex data x
r et y. Porro differentiando ex aeq. 3 fit
$dy:dx = y, 1 + \int\overline{dx:x}$. Ergo habetur et ductio
s: de Quadratura Figurae res est altioris indaginis.
nibus ego has ipsas semper judicavi simplicissimas. Nam
nisi ordinarias quantitates habent ingredientes; immiscet
ione transcendentalitas seu infinitum. Aliquoties in *Actis*
nodum calculo tractandi operi meo reservaveram, tametsi
exionem cum Logarithmis. Elegantissima videtur series
uae aream quandam dictae figurae exhibet. Quomodo inde

Lil

(seu... differentialibus) *fehlt* **L**, *erg.* ***Lil***

qvaerendi *(1)* tangentes curvarum, *(a)* qvarum *(b)* ubi in a
ingreditur *(2)* naturam ... ingreditur **L**

Lil

logarithmis. |item aliter ex aeq. 3 $dx:x = \overline{-adx:xx}\int$
$dxxy = -adxy\int\overline{dy:y} + axdy$ *gestr.* |De qvadratura **L**

judicavi *(1)* perfectissimas; et qvando rem huc reduxi, nih
L

Daniel Kurzawe

kurzawe@sub.uni-goettingen.de

Mathias Göbel

goebel@sub.uni-goettingen.de

Recommender Systems

Recommender Systems

- **Recommender Systems** help users discover content that is relevant to their current interests and which they might have missed otherwise.

- Today, these systems drive our consumption of media and information and thus directly influence our views on certain topics. This results in many new research questions, such as how can we reduce bias or make recommender systems more transparent?

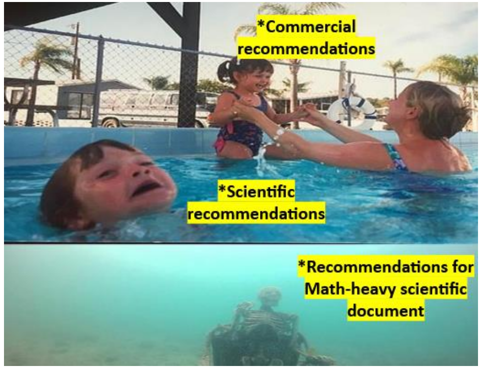
- In our group, we especially focus on improving the recommendation of **literature**, and on supporting **researchers**:
  - ➢ How can we recommend to researchers the **research papers** that are most relevant to their current interests?
  - ➢ How can we recommend suitable **scientific collaborators**?

- Areas of Research:
  - Feature extraction/ analysis
  - Semantic analysis
  - Similarity measures
  - Novel UIs & Information Visualization
  - Evaluation of recommendation interfaces

RS01 Recommender system for math-heavy scientific documents

Background

Do you also experience that the recommendations you see are not fulfilling your information needs? Especially when you are looking for information on a scientific topic and would like to understand the topic more. If yes, then we are in the same boat. These days it is easy to get “Helmet” as a recommendation when buying a “bike” online, but it is hard to get relevant scientific recommendations, especially in math-heavy STEM (Science, Technology, Engineering, Mathematics). In this project, you work on the problem of generating recommendations for scientific documents with mathematical contents. You develop methods and perform experiments to generate relevant recommendations. You will utilize a dataset that represents ideal recommendation.


*Commercial recommendations
*Scientific recommendations
*Recommendations for Math-heavy scientific document

Goal

- Generating recommendations for math-heavy scientific documents using non-textual features.

Tasks

- Analyzing citation patterns to generate recommendations.
- Generating recommendations by finding similar math formulae.
- Identifying and formulating non-textual features from scientific documents.

Ankit Satpute

Ankit.Satpute@
fiz-karlsruhe.de

André Greiner-Petter

greinerpetter@gipplab.org

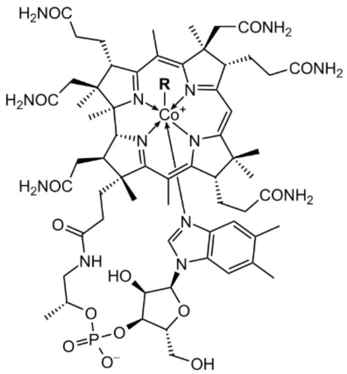
Moritz Schubotz

Moritz.Schubotz@
fiz-karlsruhe.de

RS02 Designing Chemical Feature Analysis for Chemists

Background

Scientists use recommender systems to quickly find the most relevant scientific literature in their field. However, in the chemistry domain, chemists are not only interested in textual relevance, but also want to find and make sense of chemical formulas contained in the literature. For example, within a publication: $C_9H_8O_4$, C9H8O4, acetylsalicylic acid, and ASA all refer to the same compound: Aspirin! How can we design an interface that highlights such information and lets chemists quickly get an overview of all chemical compound names, chemical formulas, and drug names discussed in a publication?



R = 5'-deoxyadenosyl, CH_3, OH, CN

Goal

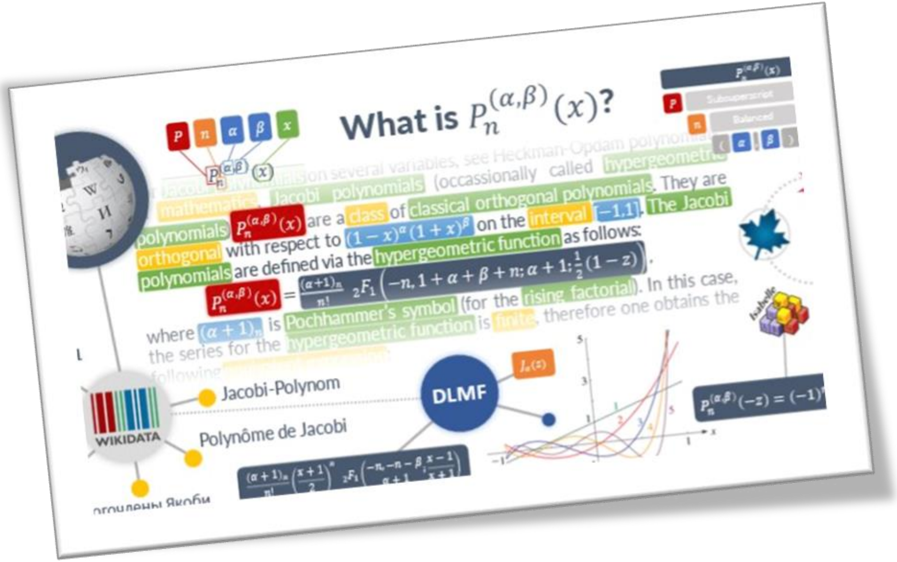
- Improve the ease of finding, comparing and understanding chemical named entities (e.g. methanol, acetone, hexane, H20, MG(OH)3, CH2Cl2) in scientific literature.

Tasks

- Research and compare existing toolkits for the automated extraction of chemical information from scientific literature, e.g., ChemDataExtractor (python)
- Build a prototype that visualizes the chemical information present in scientific papers to help chemist more quickly find, compare, and make sense of the chemical information present.

Corinna Breitinger
breitinger@uni-wuppertal.de

André Greiner-Petter
greinerpetter@gipplab.org



What is $P_n^{(\alpha,\beta)}(x)$?

Jacobi polynomials on several variables, see Heckman-Opdam polynomials

mathematics, Jacobi polynomials (occasionally called hypergeometric polynomials) $P_n^{(\alpha,\beta)}(x)$ are a class of classical orthogonal polynomials. They are orthogonal with respect to $(1-x)^\alpha(1+x)^\beta$ on the interval $[-1,1]$. The Jacobi polynomials are defined via the hypergeometric function as follows:

$$P_n^{(\alpha,\beta)}(x) = \frac{(\alpha+1)_n}{n!}\,{}_2F_1\left(-n, 1+\alpha+\beta+n; \alpha+1; \tfrac{1}{2}(1-z)\right),$$

where $(\alpha+1)_n$ is Pochhammer's symbol (for the rising factorial). In this case, the series for the hypergeometric function is finite, therefore one obtains the following

P
n
α
β
x
$P_n^{(\alpha,\beta)}(x)$
Subsuperscript
Balanced
WIKIDATA
Jacobi-Polynom
Polynôme de Jacobi
ргочлены Якоби
DLMF
$J_\alpha(z)$
Isabelle
$P_n^{(\alpha,\beta)}(-z) = (-1)^n$
$\frac{(\alpha+1)_n}{n!}\left(\frac{x+1}{2}\right)^n {}_2F_1\left(\frac{-n,-n-\beta}{\alpha+1};\frac{x-1}{x+1}\right)$

Mathematical Information Retrieval (Wikidata & Wikipedia & Translations)

Mathematical Information Retrieval

- Mathematical Information Retrieval focuses on extracting of mathematical knowledge from digital libraries for search-, recommendation- and assistance-systems.

- The project investigates fundamental methods and tools for making mathematical knowledge accessible to information retrieval tools.

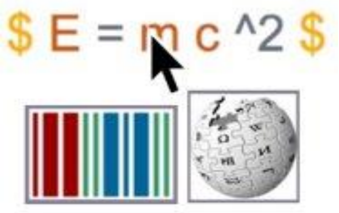
- A wide variety of applications would benefit from advancements to mathematical information retrieval:
  - academic literature search
  - literature recommendation
  - plagiarism prevention
  - tutoring assistance tools
  - patent search
  - enterprise search,

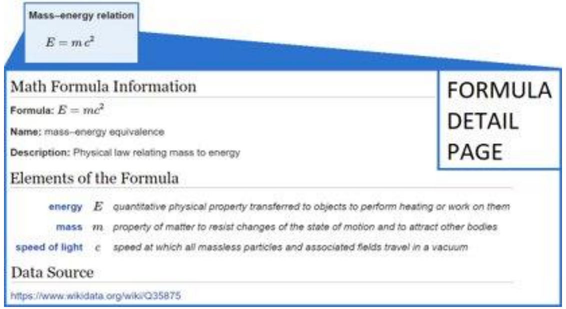
For a complete list of our research topics visit our website!

MR01 Wikipedia Formula Annotation

Background

Mathematical Question Answering systems, such as https://mathqa.wmflabs.org require labeled formulae. However, very little data is available so far.





Mass–energy relation

$E = m c^2$

FORMULA DETAIL PAGE

Math Formula Information

Formula: $E = mc^2$

Name: mass–energy equivalence

Description: Physical law relating mass to energy

Elements of the Formula

energy	E	*quantitative physical property transferred to objects to perform heating or work on them*
mass	m	*property of matter to resist changes of the state of motion and to attract other bodies*
speed of light	c	*speed at which all massless particles and associated fields travel in a vacuum*

Data Source

https://www.wikidata.org/wiki/Q35875

Goal

- Develop a formula annotation recommender system for Wikipedia.

Philipp Scharpf

scharpf@gipplab.org

Tasks

- Build an API for the AnnoMathTeX (https://annomathtex.wmflabs.org) annotation recommendations.
- Build a seeding service to transfer the annotated formula concept knowledge to Wikidata.
- Build an integrated system and deploy it to Wikipedia. If successful, you can proudly claim to be a Wikipedia developer.

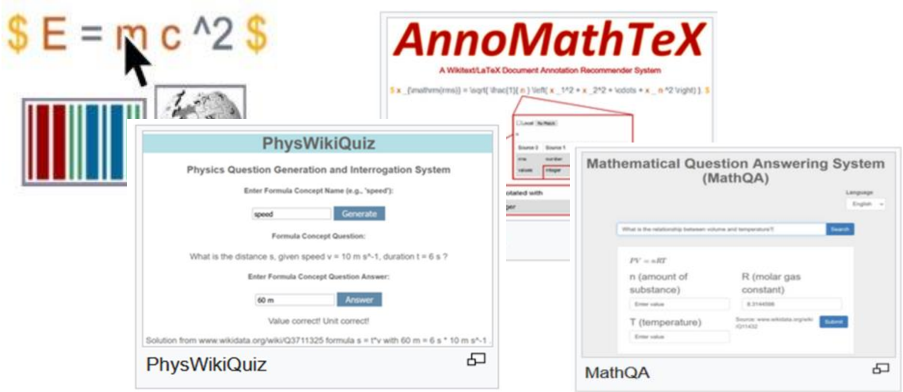
Moritz Schubotz

Moritz.Schubotz@
fiz-karlsruhe.de

MR02 A Large Dataset for Mathematical Entity Linking

Background

Mathematical Information Retrieval systems (MathIR)
and Machine Learning applications
require labeled formulae.
However, very little data is available so far.



$ E = m c ^2 $

AnnoMathTeX

A Wikitext/LaTeX Document Annotation Recommender System

$ x _{\mathrm{rms}} = \sqrt{ \frac{1}{ n } \left(x _1^2 + x _2^2 + \cdots + x _ n ^2 \right) } $

PhysWikiQuiz

Physics Question Generation and Interrogation System

Enter Formula Concept Name (e.g., 'speed'):

speed Generate

Formula Concept Question:

What is the distance s, given speed v = 10 m s^-1, duration t = 6 s ?

Enter Formula Concept Question Answer:

60 m Answer

Value correct! Unit correct!

Solution from www.wikidata.org/wiki/Q3711325 formula s = t*v with 60 m = 6 s * 10 m s^-1 .

PhysWikiQuiz

Mathematical Question Answering System (MathQA)

Language: English

What is the relationship between volume and temperature? Search

$PV = nRT$

n (amount of substance): Enter value

R (molar gas constant): 8.3144598

T (temperature): Enter value

Source: www.wikidata.org/wiki/Q11432 Submit

MathQA

Goal

- Develop a large annotated dataset of annotated (linked) mathematical formulae.

Philipp Scharpf

scharpf@gipplab.org

Tasks

- Research design principles for large annotated datasets (annotation guidelines).
- Annotate formulae in a corpus (e.g., Wikipedia articles or arXiv documents) using the AnnoMathTeX formula and identifier annotation recommender system.
- Test selected MathIR systems (e.g., MathQA or PhysWikiQuiz) on your dataset.
- Document insights from the annotation and test process to improve the systems.

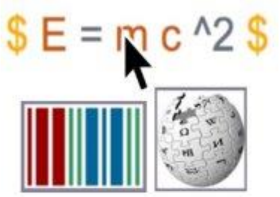
Moritz Schubotz

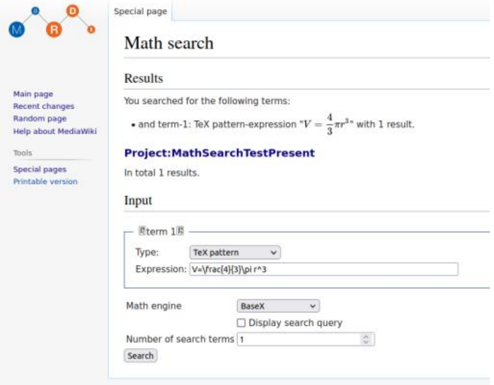
Moritz.Schubotz@
fiz-karlsruhe.de

MR03 Develop a Formula Validator for Math Search

Background

The *MathSearch* MediaWiki extension can search for wikipages based on the formulas they contain.
MediaWiki is the technological backbone for Wikipedia.





Math search

Results

You searched for the following terms:

- and term-1: TeX pattern-expression "$V = \frac{4}{3}\pi r^3$" with 1 result.

Project:MathSearchTestPresent

In total 1 results.

Input

term 1

Type: TeX pattern

Expression: V=\frac{4}{3}\pi r^3

Math engine BaseX

□ Display search query

Number of search terms 1

Goal

- Integrate a LaTeX validator to the MediaWiki extension Math Search (PHP), adapt GUI elements and user experience.

Tasks

- Set up a local development environment for MediaWiki extensions
- Write code which integrates the input validator and tests the functionality
- Create front-end elements which display the validation output for Math Search.

Johannes Stegmüller

Johannes.Stegmueller@
fiz-karlsruhe.de

Moritz Schubotz

Moritz.Schubotz@
fiz-karlsruhe.de

MR04 Research Software Infrastructures

Background

Since the European Open Science Cloud program has been launched by the European Commission in 2015, building efficient research software infrastructures has raised as a first order problematics to make European researchers properly equipped to face scientific challenges in the future. FIZ Karlsruhe supplies the main mathematical software database through its portal, swMATH, and is an active member of the EOSC Program.

But what is the situation now?

Goal

- Build a state-of-the-art literature review on research software infrastructures

Maxence Azzouz

Maxence.Azzouz-Thuderoz@
fiz-karlsruhe.de

Tasks

- Compile a list of articles in the mathematical research software area
- Identify the next challenges in the area

Moritz Schubotz

Moritz.Schubotz@
fiz-karlsruhe.de

MR05 Improve the Software of Wikipedia

Background

MediaWiki is the software running Wikipedia. While editing Wikipedia is straight forward, editing the MediaWiki source code requires a bit more effort. In this project, you will be guided to your first contribution to the open-source project MediaWiki.

Unuaiga, Wikimedia Hackathon Barcelona 2018 (04), CC BY-SA 4.0

Goal

Improve the MediaWiki software in production by fixing a bug or implementing a feature requested by the community. For example
[Add an integral symbol with a short horizontal bar in the middle (⨍,⨎)]

Tasks

1. Understand the problem and develop an implementation plan
2. Get Community Consensus
3. Set up a local development environment
4. Develop unit and integration tests
5. Interactively improve your code according to the suggestions
6. Get your code deployed and test it in production
7. Update the documentation and issue tracking software.

Johannes Stegmüller

Johannes.Stegmueller@
fiz-karlsruhe.de

André Greiner-Petter

greinerpetter@gipplab.org

Moritz Schubotz

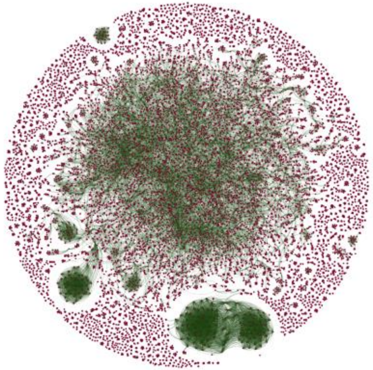
Moritz.Schubotz@
fiz-karlsruhe.de

MR06 Non-Statement View: A Set-theoretic Description of Theories

Background

The non-statement view (or structuralistic theory concept) uses set theory to describe a scientific theory through its internal structure and in conjunction with larger theory networks. This philosophical framework allows a generic theory description.

There are several publications about structural reconstructions of scientific theories, e.g. Newton particle mechanics. Due to its set-theoretical nature, a (semi-)automatic approach for such a reconstruction might be possible. This project explores this approach.



Goal

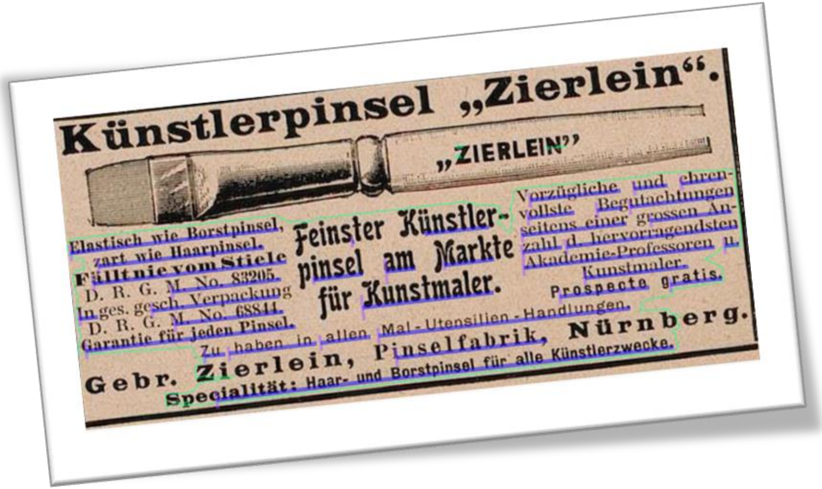
- Extraction theory components theory and transformation into a structural theory description

Tasks

- Explore concept for a semi-automatic reconstruction process
- Mapping semantic and concepts
- Build a theory network
- Implement a parser and transformer for a specific domain

Daniel Kurzawe

kurzawe@sub.uni-goettignen.de



Künstlerpinsel „Zierlein".

Elastisch wie Borstpinsel,
zart wie Haarpinsel.
Fällt nie vom Stiele
D. R. G. M. No. 83205.
In ges. gesch. Verpackung
D. R. G. M. No. 68844.
Garantie für jeden Pinsel.

Feinster Künstlerpinsel am Markte für Kunstmaler.

Vorzügliche und ehrenvollste Begutachtungen seitens einer grossen Anzahl d. hervorragendsten Akademie-Professoren u. Kunstmaler.

Prospecte gratis.

Zu haben in allen Mal-Utensilien-Handlungen.

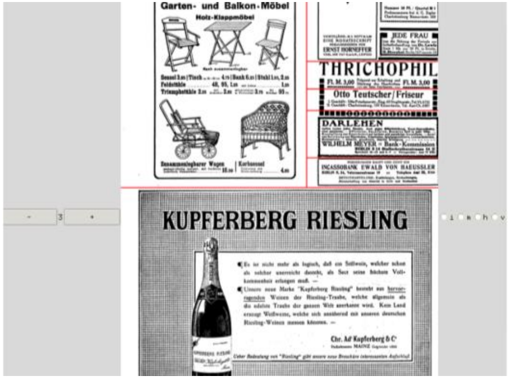
Gebr. Zierlein, Pinselfabrik, Nürnberg.
Specialität: Haar- und Borstpinsel für alle Künstlerzwecke.

Multimodal Digital Editions

MMD01 Interactive Page Segmentation

Background

While the problem of Page Segmentation can be considered 'solved' in many applications, reliably detecting and extracting visual material from historical newspapers remained a problem. Recently the "Newspaper Navigator" project gained some remarkable success by training a model based on detectron 2 with millions of crowd-sourced annotations on american historical newspapers. It would be desirable to make this model usable for page segmentation in various applications in the field of digital editing.



Goal

- Incorporate the predictions from Newspaper Navigator into a graphical user interface that allows to view, but also easily correct the predicted bounding boxes of visual material.

Tasks

- Get familiar with Newspaper Navigator and Detectron 2 and set up a pipeline to perform visual element detection
- Show the results of the process in a graphical user interface. Depending on the scope of the project and interest of participants, an existing python-based interface can be used or a new one can be created
- Export the segmented images and a json file containing the coordinates of the bounding boxes.

Johanna Sophia Störiko

johanna.stoeriko@uni-goettingen.de

MMD02 Evaluating the Newspaper Navigator Model

Background

While the problem of Page Segmentation can be considered 'solved' in many applications, reliably detecting and extracting visual material from historical newspapers remained a problem. Recently the "Newspaper Navigator" project gained some remarkable success by training a model based on detectron 2 with millions of crowd-sourced annotations on American historical newspapers. In this project you will examine the accuracy of the model on advertisement pages from the German cultural magazine "Die Jugend".

Seidenstoffe
Adolf Grieder & Cie Zürich
Humor des Auslandes
Patente
Brockhaus Lexikon
Revidierte
Jubiläums-Ausgabe
Erscheint
Locomobilen
R. Wolf
Magdeburg-Buckau
Odonta
Oscar Consée

Goal

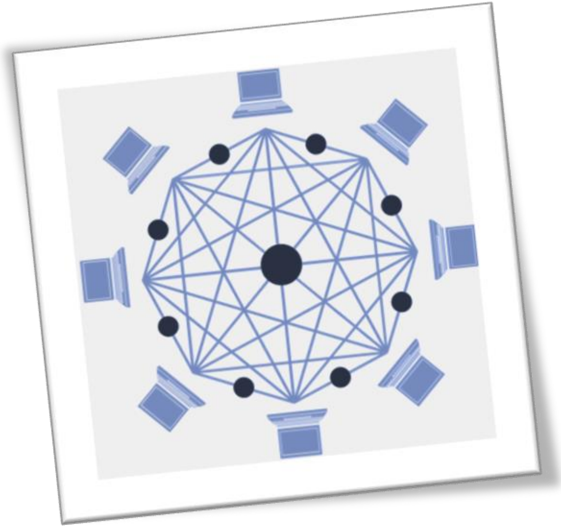
- Process the provided data with the Newspaper Navigator Model and compare the results to the given ground truth.

Tasks

- Find a common data representation for both, the outputs from newspaper navigator and the provided ground truth
- Decide on a metric to use to compute accuracy
- Compute the accuracy of the visual element detection on the data given

Johanna Sophia Störiko

johanna.stoeriko@uni-goettingen.de



Decentralized Open Science

Decentralized Open Science

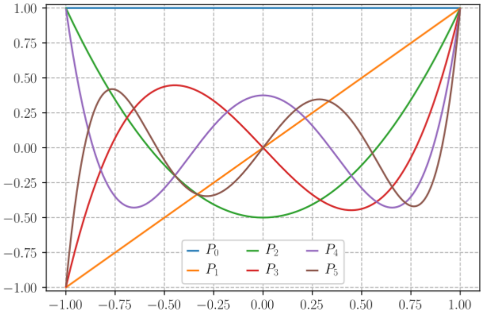
- The Decentralized Open Science aims to employ decentralized information technology to foster the open science movement.
- As described in the twelve Vienna principles, Open Science aims to make scientific processes more transparent and results more accessible. However, there are many incentives to abstain from doing Open Science, e.g., confidentiality, to keep a competitive advantage.
- Decentralization is the final iteration towards transparency and openness. We want to eliminate data silos and the dependency of Open Science tools on non-transparent central service providers.
- The project focuses on the following fields:
  - Content Protection
  - Intellectual Property Protection
  - Similarity Detection
  - Reliable Data Stores
  - Shared Computational Infrastructure

A complete list of Decentralized Open Science topics visit our website!

DOS04 PolySim – Similarity Detection Based on Polynomials

Background

Similarity detection plays an important role for information retrieval to detect similar documents. In open science, we not always own the right to share and process documents. Hence, we aim to mask the contents of input documents in the similarity detection process to keep the document plaintext hidden and protected.



Goal

- Develop a Python program to calculate the similarity between input documents based on polynomials

Tasks

- Transform document features and their positions into coordinates
- Approximate polynomials which are unique to each document
- Calculate the similarity between these polynomials

Cornelius Ihle

ihle@gipplab.org

Moritz Schubotz

Moritz.Schubotz@
fiz-karlsruhe.de

DOS05 Literature Review on Privacy-enhancing Tools for IPFS

Background

IPFS is an open network for sharing data. However, privacy is not a built-in feature of the network. Therefore, users typically rely on third-party tools to encrypt their data and anonymization tools like VPNs and TOR to protect their privacy. It is your task to provide a scoping overview on the currently existing tools to improve user privacy in IPFS.

Goal

- Develop a Python program to calculate the similarity between input documents based on polynomials

Tasks

- Systematically search public source code repositories for privacy enhancing projects that are suitable for IPFS
- Review articles for methods to improve privacy in public distributed hash tables.

Cornelius Ihle

ihle@gipplab.org

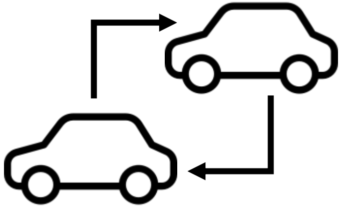
Moritz Schubotz

Moritz.Schubotz@
fiz-karlsruhe.de

DOS06 Literature Review on Peer-to-Peer (P2P) in Automotive

Background

While P2P networks were introduced over twenty years ago, it can be observed that automotive companies still heavily rely on centralized architectures. This centralized approach introduces a dependency on mobile network coverage, points of failures and other effects. In this seminar we aim to generate a comprehensive overview of P2P utilization within the automotive field including the related benefits and drawbacks.



Goal

- Carry out a systematic literature review on peer-to-peer networks in the automotive space

Tasks

- Compile a list of existing approaches
- Come up with a list of core properties that distinguish those
- Compare and contrast each proposal according to those properties

Vadim Weis

weis@gipplab.org

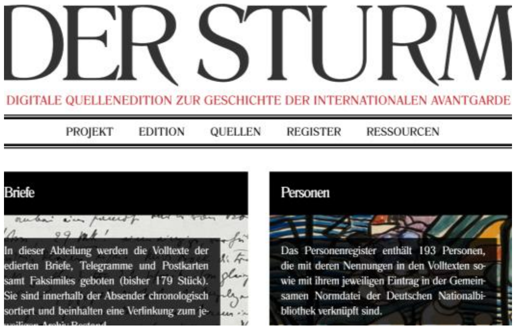
Moritz Schubotz

Moritz.Schubotz@
fiz-karlsruhe.de

DOS07 Exploring the Web3 Stack for Digital Editions

Background

Research in the humanities is undergoing a shift toward the use of digital data, methods, and tools. The sustainable, durable, and secure storage of such research data is a critical issue from an infrastructural perspective. For example, digital editions are typically stored on centralized library servers. This practice is contrary to our aspirations for decentralized open science, where we strive for availability, accessibility, and durability. One approach to bring digital editions to the decentralized Web3 is the use of web applications stored on FileCoin with the complementary sharing solution IPFS.



DER STURM

DIGITALE QUELLENEDITION ZUR GESCHICHTE DER INTERNATIONALEN AVANTGARDE

PROJEKT EDITION QUELLEN REGISTER RESSOURCEN

Briefe

In dieser Abteilung werden die Volltexte der edierten Briefe, Telegramme und Postkarten samt Faksimiles geboten (bisher 179 Stück). Sie sind innerhalb der Absender chronologisch sortiert und beinhalten eine Verlinkung zum je-

Personen

Das Personenregister enthält 193 Personen, die mit deren Nennungen in den Volltexten sowie mit ihrem jeweiligen Eintrag in der Gemeinsamen Normdatei der Deutschen Nationalbibliothek verknüpft sind.

Goal

- Explore and reverse engineer an existing digital edition (e.g. DER STURM. Digitale Quellenedition), using a Web3 software stack (React, Angular, etc.)

Tasks

- Transform TEI (XML files) into HTML-Tags and render them in a simple web app using CSS to mimic the original look and feel of the selected digital edition.
- Reverse engineer interactive elements of the digital edition like navigation bars etc. to mimic the original look and feel.
- Extend the web app so that further integrated data (e.g. images) will be loaded from IPFS instead of URLs.

Marco Beck

beck@gipplab.org

Moritz Schubotz

Moritz.Schubotz@
fiz-karlsruhe.de

Cornelius Ihle

ihle@gipplab.org

Seminar Organization

Course Schedule

Week	Date	Description	Deadlines
1	2023-04-11	Course Introduction	**17.04. (10:00)**: Submit topic interests
2	2023-04-18	**How to Find and Manage Academic Literature**	**Indiv.**: Discussion of project with supervisor, send kick-off presentation draft to supervisor
3	2023-04-25	**Kick-off Presentations** (1:1 with supervisor)	
4	2023-05-02	**How to Write Academic Papers**	
5	2023-05-09	Optional: 1:1 with supervisor	
6	2023-05-16	Optional: 1:1 with supervisor	
7	2023-05-23	**Intermediate milestone presentation** (1:1 with supervisor)	**22.5. (10:00)**: Submit intermediate milestone for applied projects OR: term paper draft (min. 4 pages)
8	2023-05-30	Optional: 1:1 with supervisor	
9	2023-06-06	**How to Give Academic Presentations**	
10	2023-06-13	optional: 1:1 with supervisor	**12.06. (10:00)**: Submit paper draft (min. 6 pages)
11	2023-06-20	optional: 1:1 with supervisor	**19.06. (10:00)**: Submit review for other paper
12	2023-06-27	optional: 1:1 with supervisor	**26.06. (10:00)**: Send draft of final presentation to supervisor
13	2023-07-04	**Maybe: Final Presentations**	**03.07. (10:00)**: Submit final presentation in StudIP
14	2023-07-11	**Final Presentations**	
	2023-08-31	**Submit Final Deliverables (Paper / Code)**	

Next Steps

- Vote on project topics: https://ogy.de/sds23-voting by **Monday, April 17, 10am**
  - Select up to 3 topics
  - Indicate order of preference 1st, 2nd , 3rd choice

Project Voting

Word of advice

- Our seminar requires work…
  - 5 credit points = 150h
  - 7 in-classroom sessions / presentations = 11h
  - 139h self-guided work = 10h / week

- …but is worth it if you commit:
  - good trial run for research and thesis projects anywhere
  - perfect for framing topic of your thesis at GippLab
  - close interaction with highly motivated supervisor

Resources

Questions?

Dr. Norman Meuschke
meuschke@uni-goettingen.de

Dr. Terry Lima Ruas
ruas@uni-goettingen.de

Prof. Dr. Bela Gipp
gipp@uni-goettingen.de

You can find guidelines, tutorials, and templates to support you throughout your studies in our wiki:

https://gipplab.org/students-corner/wiki/