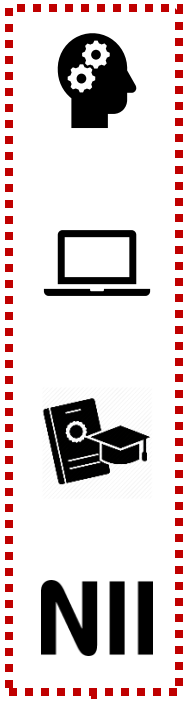
Student Projects

Summer Term 26

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
- Project suitable for BSc / MSc thesis at NII Tokyo

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

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.

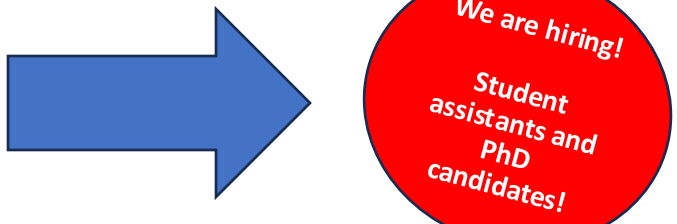
We extract specific mentions of methods from papers. This information can be part of the paper or included in tables or figures. The structure depends on the journal.

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



DIGIS

Digital Geochemistry Infrastructure for GEOROC 2.0

Daniel Kurzawe

kurzawe@sub.uni-goettignen.de

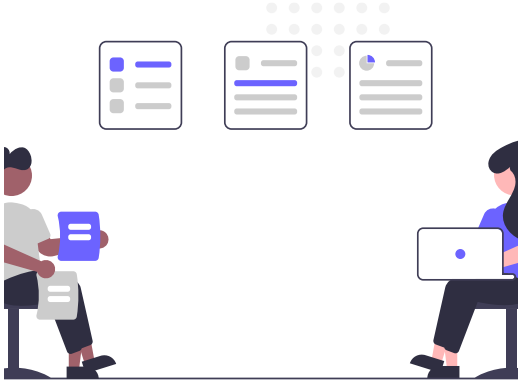
Mathias Göbel

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

- Develop a functionality to automatically add noise to the input text to assess models' robustness
- Make common automatic metrics and insightful techniques available to create a comprehensive evaluation report
- Implement a general applicable evaluation environment to test different models, datasets and metrics simultaneously

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 (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

NLP12 Quality Control of Optical Character Recognition (OCR)

Background

In the OPERANDI project, we work in a Scrum team with developers from the SUB Göttingen and the GWDG. OPERANDI strives to improve the performance and quality of OCR technology based on over 600,000 titles of German historic prints from the 16th-18th centuries. OCR results are not perfect due to poor input images or badly trained models. The quality of OCR results can be evaluated by comparing them with ground truth (GT) data (manually generated transcriptions). However, GT is rarely available as it is time consuming and cost intensive to generate. You will join the OPERANDI team to find additional methods to evaluate OCR quality, such as dictionary comparison, language models, determination of probability, or character set checking.

Goal

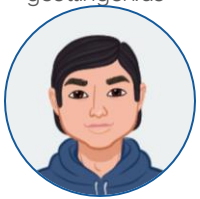
- Finding concepts and available open-source solutions for OCR quality evaluation that do not need ground truth to determine quality metrics like character error rate (CER) or word error rate (WER)

Tasks

- Review concepts, algorithms, and tools for OCR quality control
- Decide which of these would work best for the OPERANDI project
- Optional: implement the best quality control solution for a provided data sample
- Bonus: an opportunity to join our Scrum team and learn about agile ways of working

Lilja Sautter

sautter@sub.uni-goettingen.de

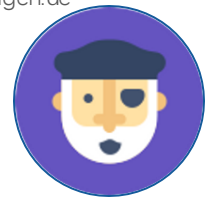


Kay Liewald

liewald@sub.uni-goettingen.de

Jörg-Holger Panzer

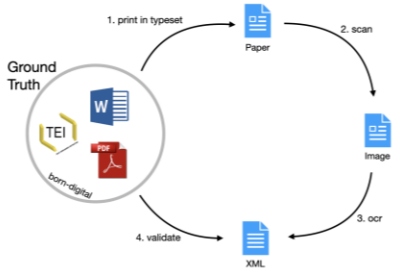
panzer@sub.uni-goettingen.de



NLP20 Automated Ground Truth (GT) Creation for OCR

Background

In the OPERANDI project, we work in a Scrum team with developers from the SUB Göttingen and the GWDG. OPERANDI strives to improve the performance and quality of OCR technology based on over 600,000 titles of German historic prints from the 16th-18th centuries. The quality of OCR results can be evaluated by comparing them with ground truth data (manually generated transcriptions). However, GT is rarely available as it is time consuming and cost intensive to generate. You will join the OPERANDI team to automate GT creation by creating GT from born-digital materials, providing a new way to evaluate OCR quality.



Goal

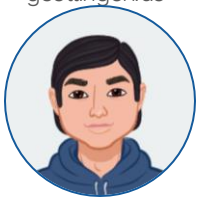
- Establishing a workflow and evaluating the usability of born-digital materials as a basis for assessing OCR quality

Tasks

- Print born-digital materials, scan and OCR these materials
- Compare the results with the ground truth from the born-digital materials to evaluate the quality of OCR results
- Repeat this in different script types (specific historic fonts)
- Bonus: an opportunity to join our Scrum team and learn about agile ways of working

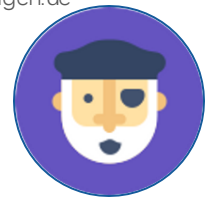
Lilja Sautter

sautter@sub.uni-goettingen.de



Kay Liewald

liewald@sub.uni-goettingen.de



Jörg-Holger Panzer

panzer@sub.uni-goettingen.de

NLP13 Do Machines Have No Heart?

Background

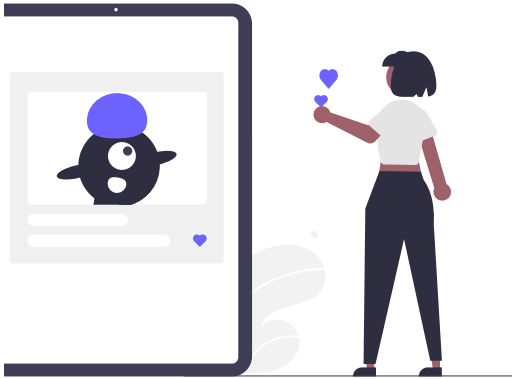
This project proposes an analysis of the sentiment embodied in text generated by large language models (LLMs), such as GPT-4. Using sentiment analysis methodologies, we aim to assess the sentiment polarity (positive, negative, neutral) and emotion classification (joy, anger, surprise) inherent in machine generated text across a diverse prompts and contexts. The proposed study will focus on understanding how LLMs, despite their lack of emotional states or personal perspectives, can potentially generate text embedding a wide spectrum of sentiment expressions. A significant aspect of our research will be identifying any sentiment inconsistencies in the model outputs, particularly in the face of ambiguous or complex prompts and comparing with existing human experiments

Goal

- Explore the sentiment embedded in LLM using machine-generated text and comparing it with human behavior

Tasks

- Literature review on sentiment/emotion analysis on language models
- Probe selected LLM to generate text following prompts/instructions
- Sentiment analysis and exploration of LLM's output
- Correlation between human and machine text



Jan Philip Wahle

wahle@gipplab.org

Terry L. Ruas

ruas@gipplab.org

NLP14 Paraphrase Types: Data and Task Generation

Background

Current paraphrase generation and detection systems are yet unaware of the lexical variables they manipulate. Generative models cannot be asked to perform certain types of perturbations, and detection models are unable to understand which paraphrase types they detect or learn limited language aspects (e.g., primarily syntax). The shallow notion of what composes paraphrases used by these systems limit their understanding of the task and makes it challenging to interpret detection decisions in practice. Thus, we need to leverage existing datasets and tasks used in Paraphrasing with more granular information so we can assess the problem better and develop more robust techniques.

Goal

- Extend current datasets used in paraphrase related tasks to include paraphrase types

Tasks

- Literature review on paraphrase types (atomic paraphrase types)
- Probe existing LLM to generate/classify pair sentences including selected paraphrase types (e.g., prompting, few-, or zero-shot) using the ETPC dataset as a reference
- Correlate (e.g., BLEU, similarity, ROUGE, BERTScore) generated paraphrase with existing data and select the best paraphrase types
- Extend the best paraphrase types to generate/classify new data from other paraphrase datasets
- Propose new tasks for the BIG-bench and/or GEM benchmarks based on Paraphrase Types

Jan Philip Wahle

wahle@gipplab.org

Terry L. Ruas

ruas@gipplab.org

NLP15 The Paraphrase Type Taxonomy

Background

This project proposes an extensive literature review to identify and critically evaluate various paraphrase types that have been proposed in linguistic, computational, and educational domains. By synthesizing these diverse perspectives, the project aims to develop a cohesive framework that categorizes paraphrase types based on linguistic features, context, and communicative intent. Through rigorous analysis and categorization, the project aims to establish a comprehensive taxonomy of paraphrase types. Furthermore, the research team plans to develop an open-access online repository, where the findings and the framework will be made available to the public, promoting collaboration and further research in this domain.

Goal

- Investigate and (re)organize available taxonomies and language models used in paraphrase types

Tasks

- Investigate available taxonomies used in paraphrase (types)
- Critical evaluation of existing ones (agreement and disagreements between them)
- Investigate available models used in paraphrase generation and detection
- Propose a new taxonomy (with definitions, examples, and instructions) for paraphrase types (generation and detection)

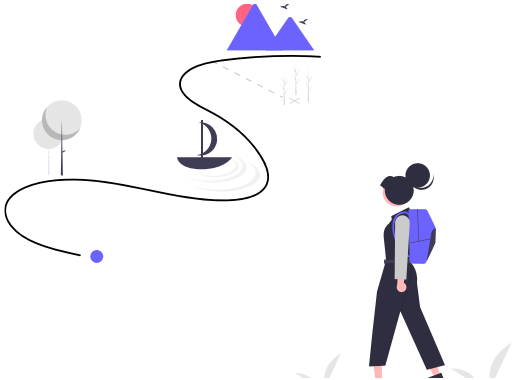
Jan Philip Wahle
wahle@gipplab.org

Terry L. Ruas
ruas@gipplab.org

NLP17 LLMs and the Search for the Holy Prompt

Background

The advancements in the capabilities of large language models (LLMs) have ushered in a new era in artificial intelligence, with applications spanning diverse sectors (e.g., healthcare, education, entertainment). However, extracting precise and desired information from these models is not trivial. An emerging understanding of "prompt engineering" plays a crucial role in determining the efficacy, precision, and utility of the response from LLMs. Investigating the importance of prompt engineering is hence crucial, not only to improve the practical deployment of LLMs but also to delve deeper into understanding the intricacies of their internal representation and response mechanisms.



Goal

- Systematically explore and quantify the impact of prompt engineering on the performance of LLMs in paraphrase-related/text generation tasks and develop best practices for finding the best prompts

Tasks

- Literature Review: Examine existing literature on prompt engineering
- Empirical Study: Design experiments using various prompts across multiple tasks to measure the variability in LLMs' performance based on prompt differences.
- Framework Development: Construct a framework or guideline, based on empirical results, for crafting effective prompts that maximize desired outcomes when interacting with LLMs.
- Evaluate models, tasks, and prompts in selected tasks

Jan Philip Wahle

wahle@gipplab.org

Terry L. Ruas

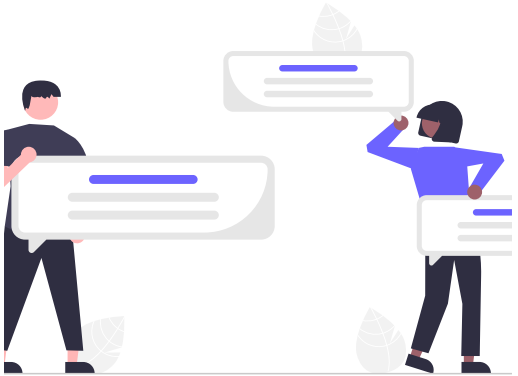
ruas@gipplab.org

NII

NLP18 It's not What you say it, but How you say it

Background

Large language models (LLM) have revolutionized Natural Language Processing (NLP) due to their ability to understand and generate human-like text. Their efficacy in producing meaningful outputs relies significantly on the way they are prompted. The same way as people, by slight alterations in prompts can lead to considerable differences in the generated content, which may affect both the quality and the nature of the response. This also raises the questions if specific models have a certain bias towards prompts. This phenomenon underscores the need for an in-depth analysis of the relationship between prompts, output and LLMS.



Goal

- Analyze how varying prompts influence the outputs of LLMs across selected NLP tasks and derive insights that can guide effective prompting strategies. Understand the differences between prompt alternation and selected LLM

Tasks

- Literature Review: Examine existing research and documented observations on how prompts influence large language models (select models and tasks)
- Experimental Design: Create a diverse set of prompts for selected NLP tasks. This set should include varied lengths, tones, styles, and implicit biases. (manual or auto)
- Data Collection: Use the selected prompts on consistent LLMs and collect/evaluate the outputs for each prompt (against gold standard)
- Analysis and Interpretation: Evaluate the data to discern patterns and relationships between prompt variations and model outputs.

Jan Philip Wahle

wahle@gipplab.org

Terry L. Ruas

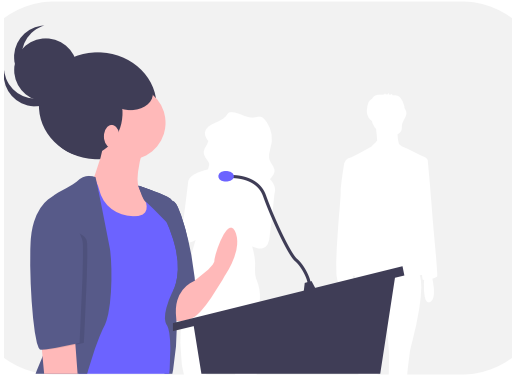
ruas@gipplab.org

NII

NLP19 What Are We Talking (and Doing) About in the EU?

Background

In an era with so much data available, knowing what to extract and how to structure it is essential for solving any problem. The structured compilation of extensively discussed topics at the European Union Parliament not only empowers policymakers, researchers, and analysts with a comprehensive overview of the legislative landscape but also grants citizens a clear overview of the issues that shape their continent. This project is not just a technical undertaking, but a venture that lays the foundation for transparency, accountability, and progress. This project focuses on the organization and exploration of the European Parliament's Open Data into meaningful structures so further investigations can be carried out.



Goal

- Analyze and organize the (selected) data of the European Parliament's into a more accessible way so specific investigations can be carried out.

Tasks

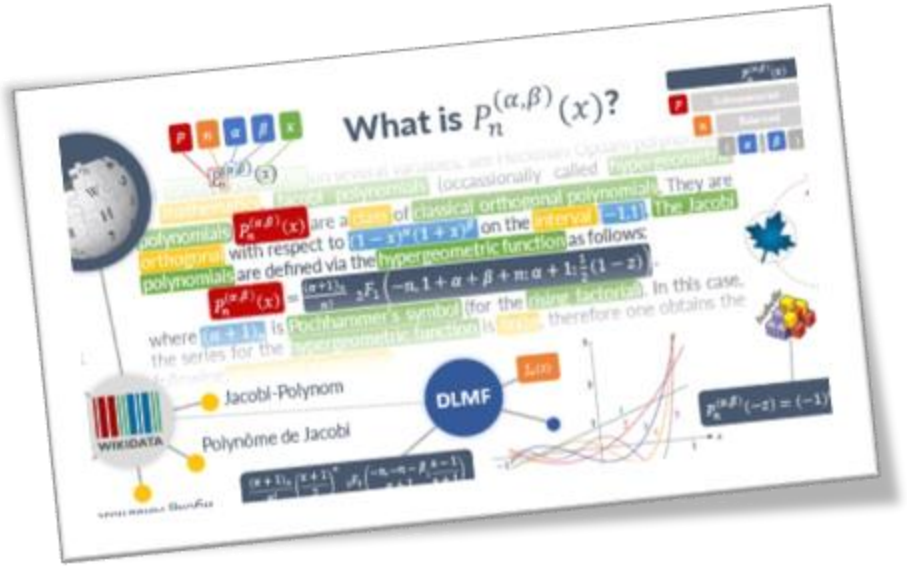
- Understand the structure of the European Parliament's Open Data Portal
- Identify major categories and topics we would like to gather and organize data (specific lexicons might be used to curate such data)
- Implement a solution to extract, categorize, and store data on selected topics from minutes, plenary sessions, speakers, etc;
- Propose sub-topic organization for the data
- Provide an initial (data science) analysis on selected topics

Jan Philip Wahle

wahle@gipplab.org

Terry L. Ruas

ruas@gipplab.org



Mathematical Information Retrieval (Wikidata & Wikipedia & Translations)

Mathematical Information Retrieval

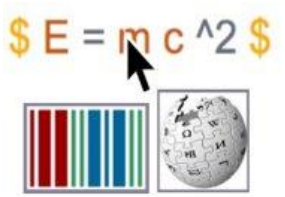
- Mathematical Information Retrieval focuses on extracting of mathematical knowledge from digital libraries for search-, recommendation- and assistance-systems.
- The project investigates fundamental methods and tools for making mathematical knowledge accessible to information retrieval tools.
- A wide 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,

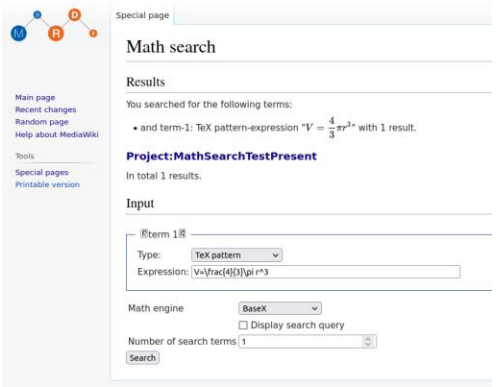
For a complete list of our research topics visit our website!

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.





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

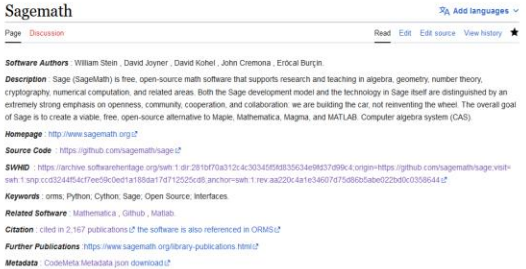
Moritz Schubotz

Moritz.Schubotz@fiz-karlsruhe.de

MR04 Identify citations for research software in scientific articles

Background

For years, Graph Neural Networks have been growingly adopted for cases where data are not independent and identically distributed. Drug-protein prediction, social network clustering and scientific article/software recommendations are examples of applications where GNN has succeeded. However, it is not straightforward to model research articles and research software in the same graph, primarily since these last ones deal with complex heterogeneous metadata contents, formats, and source code containing both programming and natural language.



Goal

We aim to use Heterogeneous Graph Neural Networks to identify software in mathematical research articles, emphasizing software metadata. The approach must be then implemented in Julia using standard libraries.

Tasks

- Set up a Julia environment on your computer
- Build a pipeline to prepare swMATH and zbMATH (restricted to ArXiv sources) data and metadata
- Use standard libraries to inject the relevant data in a GNN model and train this model to identify software in scientific articles

Maxence Azzouz

Maxence.Azzouz-Thuderoz@fiz-karlsruhe.de

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.

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 (ʃ,⨍)

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.

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

Johannes Stegmüller

Johannes.Stegmueller@fiz-karlsruhe.de

André Greiner-Petter

greinerpetter@gipplab.org

Moritz Schubotz

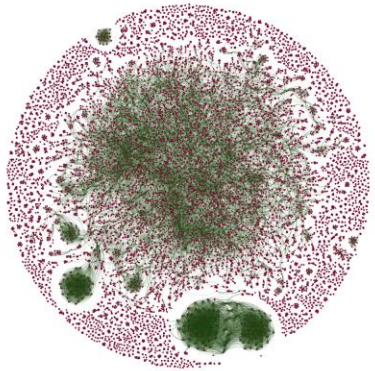
Moritz.Schubotz@fiz-karlsruhe.de

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

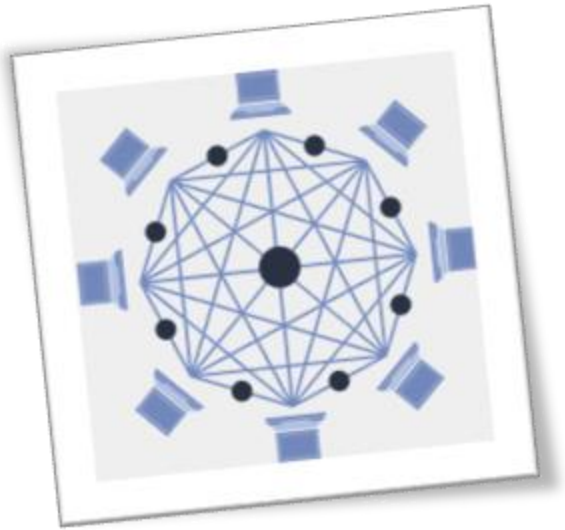
- Extraction theory components theory and transformation into a structural theory description

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



Decentralized Open Science

Decentralized Open Science

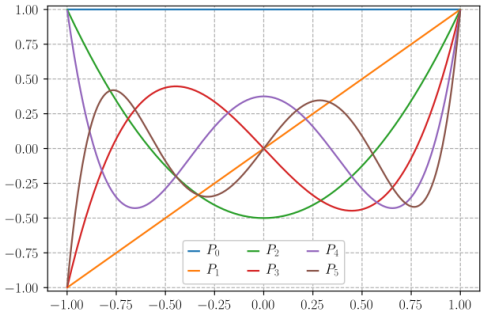
- The Decentralized Open Science aims to employ decentralized information technology to foster the open science movement.
- As described in the twelve Vienna principles, Open Science aims to make scientific processes more transparent and results more accessible. However, there are many incentives to 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

- Synthesize the existing reserach for privacy-enhancing tools for IPFS to devise recommendations for further IPFS development

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

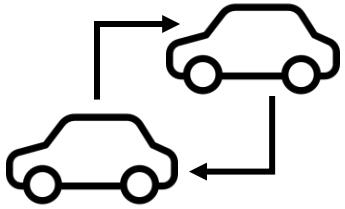
Moritz Schubotz

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.

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

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.

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

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.

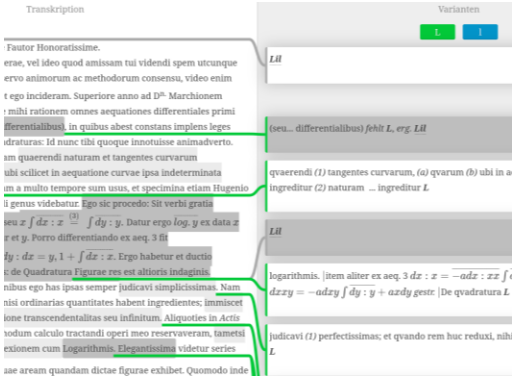
What are the shared approaches of plagiarism detection and textual criticism? Can they benefit from each other?

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)



Daniel Kurzawe

kurzawe@sub.uni-goettingen.de

Mathias Göbel

goebel@sub.uni-goettingen.de

PD08: Authorship verification using LLMs

Background

Cloze test proved to be a useful tool for testing text comprehension. Some universities use it during a disciplinary procedure when a student is suspect from submitting a work authored by someone or something else (plagiarism, contract cheating, unallowed use of generative AI). Authors of the text are more likely to fill in correct words.

The project aims to find a method that identifies words to be masked such that the cloze test can reliably discriminate between authors and non-authors. LLMs are trained to predict the word in given context. Previous experiments showed that nouns that the model would not guess correctly are good candidates.

Goal

- To extend the existing project by conducting more experiments with LLMs and users
- To improve existing method (better discriminate between authors and non-authors)

Tasks

- Employ more language models to identify masked word (so far only MT-5 was used)
- Experiment with probability of the word in given context (so far only rank was used)
- Investigate the influence of language (English, German, etc.; native / non-native)
- Investigate the influence of time (authors forget their text and achieve lower scores)

The project aims to find a _____ that identifies words to be masked such that the cloze test can reliably _____ between authors and non-authors. LLMs are trained to predict the word in given context. Previous _____ showed that nouns that the model would not guess correctly are good candidates.

Tomáš Foltýnek
foltynek@fi.muni.cz

Terry L. Ruas
ruas@gipplab.org

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.

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

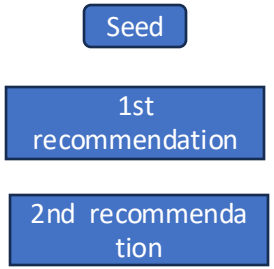
Moritz Schubotz

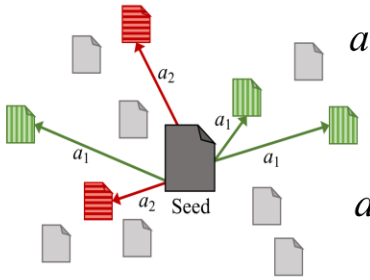
Moritz.Schubotz@fiz-karlsruhe.de

RS03 Building a Scientific Document Recommender System

Background

Recommender System (RS) suggests relevant scientific articles from vast amount of scientific literature. RS helps students, researchers and professionals to keep up with new developments in their area of interest. RS typically do not consider in what aspect two documents are similar. In this project, you implement a RS that produces recommendations based on specific aspects (math formulae, keywords, etc).





a_1 : Recommendations generated with similar math formulae.

a_2: Recommendations generated with similar keywords.

Goal

- Building an aspect based scientific document recommender recommender system.

Tasks

- Setting up an Elasticsearch cluster with scientific documents from arXiv, zbMATH Open.
- Build a prototype that visualizes recommendations for a seed document retrieved from Elasticsearch based on aspect such as math formulae, text, Keywords, etc.

Ankit Satpute

Ankit.Satpute@fiz-karlsruhe.de

André Greiner-Petter

greinerpetter@gipplab.org

Moritz Schubotz

Moritz.Schubotz@fiz-karlsruhe.de

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".

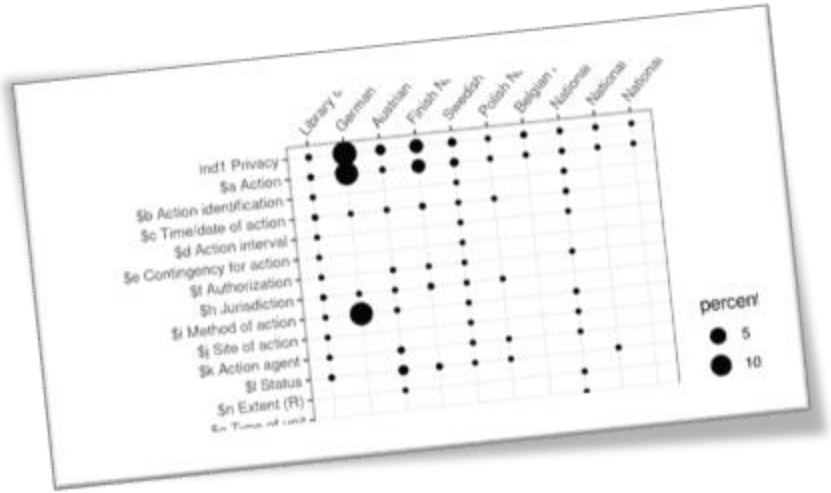
Goal

- 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

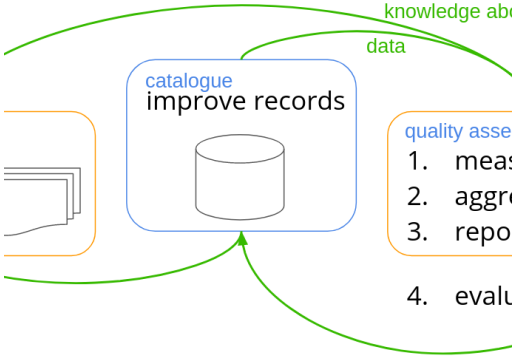


Cultural Analytics

CA01 (Meta)data quality assessment

Background

Everybody recognizes bad data, but it is not easy to define what makes data good or bad. Quality assessment is a special data analysis process aiming to highlight some features of a dataset called quality dimensions. This analysis can be used in a later step of data analysis, such as data cleaning or exploratory data analysis. In this course we use cultural heritage metadata (library, archival and museum catalogues) as our research data. We will learn about theories such as data quality dimensions. We will use and contribute to the development of assessment tools to detect quality related problems. Finally we will discuss the results with metadata experts of the data provider institutions.



Goal

- Understanding the full life cycle of data quality assessment (study data quality dimensions, tools, and a metadata standard, assess quality with a relevant tool and communicate the result with data curators).

Tasks

- Review literature about (meta)data quality
- Understand the metadata schema of a selected cultural heritage data source
- Adapt a quality assessment tool (e.g. SHACL, JSON Schema, QA catalogue) to measure quality dimensions
- Visualize the results and communicate with metadata experts of the data provider

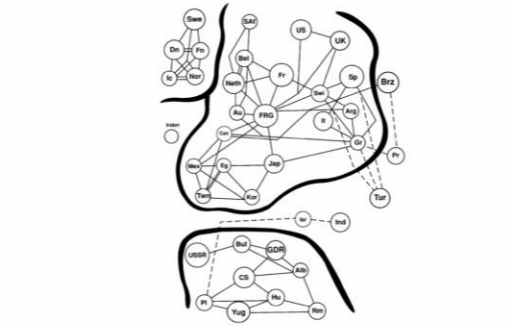
Péter Király

peter.kiraly@gwdg.de

CA02 Bibliographic data science

Background

Bibliographic data contains factual historical dimensions, such as personal names of authors and contributors (occasionally with additional properties), place and date of publication, name of publishers/printers/scriptors, genre, subject description of the content (keywords, classification terms), materiality, provenance (current and previous holding institutions, owners). After data cleaning and normalization all these information shed light to historical patterns, such as how the roles of different languages changed regionaly, how the literary canon evolved, who were the important authors and books in a particular periods, enduring and ephemeral best-sellers, how the media changed, and how all these correlated with each other?



Political and S-C divides in Europe in the 1970s & 1980s (Šajkevič 1992), Index Translationum data

Goal

- Run historical data analysis on library catalogues (understand, extract, normalize, analyze and visualize bibliographic data, compare result with qualitative sources).

Tasks

- Review literature about bibliographic data science
- Formulate research questions
- Understand the metadata schema of a selected cultural heritage data source
- Use R/Python/Java to clean, analyze and visualize data
- Check literature if your result is a novelty and compare with state-of-the-art research

Péter Király

peter.kiraly@gwdg.de

Computer Vision

CV01 Individual tree segmentation

Background

As part of the Biodiversity Exploratories, a drone equipped with a LiDAR sensor was employed to survey 23 forest sites in Germany. Each site generated approximately 150 million 3D points. The primary objective of the survey is to extract single tree variables such as crown shape, horizontal biomass, tree height, and other 3D characteristics. The assignment of LiDAR points to individual trees is necessary to achieve this objective. Forest managers and modellers are particularly interested in tree crown variables as they can identify productive and healthy trees. In addition, forest modellers can establish relationships between crown characteristics and other tree variables, which is an important subject in remote sensing applications.

Goal

- Individual tree segmentation from UAS derived LiDAR point clouds

Tasks

- Review literature about semantic segmenation in point clouds e.g. PointNet
- Generating a training and independent test database
- (try to) train a model able to segment individual trees and (try to) classifiy the corresponding tree species
- Compute the accuracy against the background of spatial autocorrelation

Dr. Nils Nölke

nnoelke@gwdg.de

Open Scholarly Data

OS01 Open Scholarly Data Warehouse

Background

At the SUB Göttingen, we run an **open scholarly data warehouse** using the **GWDG Scientific Cluster** and **Google Cloud BigQuery**. We need **data engineering support** to maintain and enhance our data warehouse and the underlying data pre-processing and import workflows.

Goal

- To maintain and enhance the data warehouse and underlying workflows. We want to automate as much of the process as possible.

Tasks

- Monthly data updates
- Maintenance of existing Python scripts
- Monitoring of (schema) changes in the data sources
- Design of an automated workflow (eg Apache Airflow)

Najko Jahn
najko.jahn@sub.uni-goettingen.de

Nick Haupka
nick.haupka@sub.uni-goettingen.de