Task Area Collections

Text-Mining-Pipelines für unstrukturierten Text

Florian Barth (SUB Göttingen / Göttingen Centre for Digital Humanities), José Calvo Tello (SUB Göttingen), Stefan Funk (SUB Göttingen), Mathias Göbel (SUB Göttingen), Daniel Kurzawe (SUB Göttingen), Nanette Rißler-Pipka (Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen), Ubbo Veentjer (SUB Göttingen)

Sammlungen an der SUB Göttingen

- das **TextGrid Repository** beinhaltet u. a. literarische Korpora (Digitale Bibliothek), digitale Editionen (Architrave, Bibliothek der Neologie) oder Digitalisate (Virtuelles Skriptorium St. Matthias)
- **DARIAH-DE Repository** und Portal **DiscussData** als Plattformen für geisteswissenschaftliche Forschungsdaten
- **DigiZeitschriften**, das deutsche digitale Zeitschriftenarchiv, umfasst wissenschaftliche Zeitschriftenbestände aus dem geisteswissenschaftlichen Bereich
- im **Verzeichnis deutschsprachiger Drucke des 17. sowie des 18. Jahrhunderts (VD17, VD18)** werden vom Göttinger Digitalisierungszentrum und nationalen Partnern alle Druckwerke der entsprechenden Jahrhunderte digitalisiert

TextGrid Virtuelle Forschungsumgebung für die Geisteswissenschaften · DARIAH-DE Digitale Forschungsinfrastruktur für die Geistes- und Kulturwissenschaften · DISCUSS DATA · DigiZeitschriften DAS DEUTSCHE DIGITALE ZEITSCHRIFTENARCHIV · VD17 · VD18

Unstrukturierter Text

Unterschiedlicher Stand der Digitalisierung und textinternen Anreicherung in SUB-Ressourcen:

- Texte mit umfassenden Strukturinformationen
- Plain-Texte mit schwankender OCR-Qualität
- als Bilddateien digitalisierte Faksimiles (VD17, VD18)

Ziele

- Erzeugung qualitativ hochwertiger Plain-Texte für Ressourcen mit Bilddateien in Zusammenarbeit mit dem OCR-D-Projekt
- Text-Mining-Verfahren zur Anreicherung mit textinterner Struktur
- Bereitstellung abgeleiteter Textformate (Schöch u. a. 2020)

OCR-D DFG-funded Initiative for Optical Character Recognition Development

Ressourcen

Import

- Ingest kompletter Sammlungen (Text und Metadaten)
- Support für spezifische Daten- und Metadatenformate, z. B. TEI-XML-Varianten von Community-relevanten Korpora

TextGrid Virtuelle Forschungsumgebung für die Geisteswissenschaften · **KOLIMO** · European Literary Text Collection (ELTeC)

Anreicherung

- Data Reconciliation: Überprüfung, Vervollständigung und Erweiterung der Metadaten
- Enrichment: Verknüpfung der Metadaten mit Normdaten und Knowledge Bases

VIAF · GND · WIKIDATA · DBpedia

Export

- Serialisierung angereicherter Metadaten in gewünschtes Format (Eingangsformat oder Neuformat)

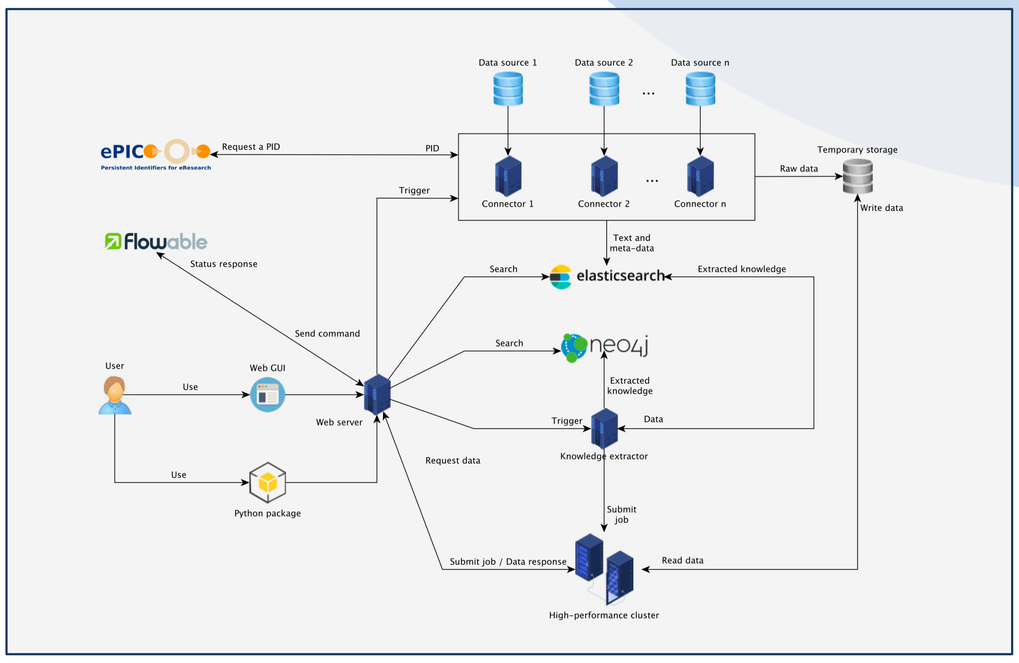
TEI · METS · MODS · RDF

Corpus Reader

MINE

- **Data Warehouse**
  normalisierte Einbindung distinkter Ressourcen
- **Elastic Search**
  Repository-übergreifende Suche für Text und Metadaten
- **Knowledge Graph (neo4j)**
  Ressourcen-übergreifende Verknüpfung von Informationen auf Dokumentebene und textinternen Anreicherungen

MINE



Natural Language Processing Pipeline

NLP-Pipeline *MONAPipe* (Dönicke u. a. 2022):

- entwickelt im Projekt *Modes of Narration and Attribution* (MONA) innerhalb des Göttingen Centre for Digital Humanities (GCDH)
- basierend auf spaCy 2 (Honnibal und Montani, 2017)
- Standardkomponenten und Custom-Komponenten (Re-Implementierungen / Wrapper bestehender Systeme sowie Eigenentwicklungen)

Anreicherung textinterner Strukturinformationen, u. a.:

- linguistische Strukturen (Zeitformen, Zeitausdrücke, Sprachnormalisierung, Satz-/Teilsatzsplitting)
- Named Entities (inkl. Koreferenzen sowie Entity Linking)
- narrative Phänomene (Redewiedergabe, Sprechererkennung und -attribution, Ereignisse, Reflexionen)

Preprocessing

Tokenizer (spaCy)	Tagger (spaCy)	Lemmatizer (spaCy)	NER (spaCy)	Normalizer (Dönicke, u. a. 2022)

Morphosyntaktische Analyse

Sentencizer (spaCy / NLTK)	Dependency Parser (spaCy)	Clausizer (Dönicke, 2020)	Analyzer (Altinok, 2018)	TenseTagger (Dönicke, 2020)

Semantische Analyse

Temponym Tagger (Strötgen und Gertz, 2010 / 2015)	Emotions Tagger (Mohammad und Turney, 2010; Dönicke, u. a. 2022)

Sprecher/Rede, Koreferenzresolution

Speech Tagger (Brunner u. a. 2022)	Speaker Extractor (Dönicke, u. a. 2022)	Coref (Krug, u. a.; Dönicke, u. a. 2022)

Modes of Narration and Attribution

Event Tagger (Vauth, u. a. 2021)	Comment Tagger (Weimer u. a., im Erscheinen)	Generalization Tagger (Gödeke, im Erscheinen)	Entity Linker (Barth u. a., 2022)	Attribution Tagger (Dönicke u. a., 2022)

Referenzen

Duygu Altinok. 2018. DEMorphy, German language morphological analyzer. arXiv:1803.00902.

Florian Barth, Hanna Varachkina, Tillmann Dönicke und Luisa Gödeke. 2022. Levels of non-fictionality in fictional texts. In Proceedings of the 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation within LREC2022, S. 27–32, Marseille, France. European Language Resources Association.

Annelen Brunner, Ngoc Duyen Tanja Tu, Lukas Weimer und Fotis Jannidis. 2020. To BERT or not to BERT – comparing contextual embeddings in a deep learning architecture for the automatic recognition of four types of speech, thought and writing representation. In SwissText/KONVENS.

Tillmann Dönicke. 2020. Clause-level tense, mood, voice and modality tagging for German. In Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories, S. 1–17, Düsseldorf, Germany. Association for Computational Linguistics.

Tillmann Dönicke, Florian Barth, Hanna Varachkina und Caroline Sporleder. 2022. MONAPipe: Modes of Narration and Attribution Pipeline for German Computational Literary Studies and Language Analysis in spaCy. In Proceedings of KONVENS (Konferenz zur Verarbeitung natürlicher Sprache/Conference on Natural Language Processing).

Tillmann Dönicke, Hanna Varachkina, Anna Mareike Weimer, Luisa Gödeke, Florian Barth, Benjamin Gittel, Anke Holler und Caroline Sporleder. 2022. Modelling speaker attribution in narrative texts with biased and bias-adjustable neural networks. Frontiers in Artificial Intelligence, 4.

Luisa Gödeke, Florian Barth, Tillmann Dönicke, Anna Mareike Weimer, Hanna Varachkina, Benjamin Gittel, Anke Holler und Caroline Sporleder. Im Erscheinen. Generalisierungen als literarisches Phänomen. Charakterisierung, Annotation und automatische Erkennung. Zeitschrift für digitale Geisteswissenschaften.

Matthew Honnibal und Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Markus Krug, Frank Puppe, Fotis Jannidis, Luisa Macharowsky, Isabella Reger und Lukas Weimar. 2015. Rule-based coreference resolution in German historic novels. In Proceedings of the Fourth Workshop on Computational Linguistics for Literature, S. 98–104, Denver, Colorado, USA. Association for Computational Linguistics.

Saif Mohammad und Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, S. 26–34, Los Angeles, CA. Association for Computational Linguistics.

Christof Schöch, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann und Jörg Röpke. 2020. Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. In: Zeitschrift für digitale Geisteswissenschaften. Wolfenbüttel. text/html Format. DOI: 10.17175/2020_006

Jannik Strötgen und Michael Gertz. 2010. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In Proceedings of the 5th International Workshop on Semantic Evaluation, S. 321–324, Uppsala, Sweden. Association for Computational Linguistics.

Jannik Strötgen und Michael Gertz. 2015. A baseline temporal tagger for all languages. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, S. 541–547, Lisbon, Portugal. Association for Computational Linguistics.

Michael Vauth, Hans Ole Hatzel, Evelyn Gius und Chris Biemann. 2021. Automated event annotation in literary texts. In Proceedings of the Conference on Computational Humanities Research 2021 (CHR 2021), S. 333–345, Amsterdam, the Netherlands.

Anna Mareike Weimer, Florian Barth, Tillmann Dönicke, Luisa Gödeke, Hanna Varachkina, Anke Holler, Caroline Sporleder und Benjamin Gittel. Im Erscheinen. The (in-)consistency of literary concepts – formalising, annotating and detecting literary comment. Journal of Computational Literary Studies.

DFG Deutsche Forschungsgemeinschaft

Das NFDI-Konsortium Text+ wird gefördert durch die Deutsche Forschungsgemeinschaft (DFG) unter der Projektnummer 460033370.

Text+ ist Teil der nfdi Nationale Forschungsdaten Infrastruktur

Poster: Florian Barth, José Calvo Tello, Stefan Funk, Mathias Göbel, Daniel Kurzawe, Nanette Rißler-Pipka, Ubbo Veentjer

Text+ Plenary 2022,
12./13. September 2022,
Mannheim