Exploring Historical Sources with Language Technology: Results and Perspective

Carsten Schnober

TECHNISCHE UNIVERSITÄT DARMSTADT

Welt der Kinder
Children and their World

http://welt-der-kinder.gei.de/

https://www.ukp.tu-darmstadt.de/research/current-projects/welt-der-kinder/

UKP

Welt der Kinder: People and Institutions

- Funded by Leibniz Association
- May 2014 – May 2017
- Georg Eckert Institute for International Textbook Research
  - Maik Fiedler, Andreas Weiß, Dr. Robert Strötgen, Prof. Dr. Simone Lässig
- Ubiquitous Knowledge Processing Lab (UKP) TU Darmstadt, and German Institute for International Educational Research (DIPF)
  - Carsten Schnober, Dr. Richard Eckart de Castilho, Prof. Dr. Iryna Gurevych
- Institute for Information Science and Language Technology in Hildesheim (IWIST)
  - Ben Heuwing, Prof. Dr. Thomas Mandl, Prof. Dr. Christa Womser-Hacker

Welt der Kinder – Children and their World

TECHNISCHE UNIVERSITÄT DARMSTADT

- Background
- Project Goals
- Research Questions
- Methods
  - Natural Language Processing
  - User-centered development
  - Visualization

Welt der Kinder: Background

TECHNISCHE UNIVERSITÄT DARMSTADT

- Digital humanities: cooperation between historians, information scientists, and computer scientists
- Research subject: German textbooks from 1850 until 1918
  - 6,000 digitized textbooks and juvenile books (900,000 pages)
  - available at GEI Digital: http://gei-digital.gei.de
- Period from 1850 till 1918 of particular historical interest
  - Accelerated production of knowledge
  - Dominated by globalization and nationalization simultaneously
- Sources reflect contemporary world interpretation patterns and elements of cultural memory at the time

Welt der Kinder: Goals

TECHNISCHE UNIVERSITÄT DARMSTADT

1) Historical research about representations and interpretations of the world
    - Textbooks as main information source for young adults
2) Exploration of specific media types: textbooks and juvenile literature
    - Consider media type, coverage, and transformation of collective knowledge
3) Combine established hermeneutic methodology with innovative methods
4) Fundamental research in computational linguistics
    - Develop and adapt methods of topic modeling, semantic analysis, and opinion mining

Welt der Kinder: Research Questions

TECHNISCHE UNIVERSITÄT DARMSTADT

1) “Foreign” continents, countries, peoples, colonies, discoveries, expeditions, missions
2) War and peace, political alliances, regents and statesmen
3) Science and culture, urbanisation and industrialisation, pauperism and social policies
4) Transportation, communication, trade, industry, technology, environment

OCR and Post-Processing

TECHNISCHE UNIVERSITÄT DARMSTADT

- OCR done mostly with Abbyy Finereader in a previous project
  - Estimated accuracy rate: 90-95% (characters)
- OCR error analysis
  - Segmentation: “DieNordarmeestand unter”, “Hilfs-truppen”
  - Character recognition: “alle feine Lander”, “den Â©rund zu”, “Dtto”
  - Garbage: "ijiilaAll"
- Spelling variations in the original document (unrelated to OCR)
  - Differences to today's language and within the corpus (canonical German orthography was introduced in 1905)
  - Spelling mistakes by authors

Topic Modelling: Sample Topics

TECHNISCHE UNIVERSITÄT DARMSTADT

- Words beginning with capital letters (200 topics, top 10 words):
  - Sieg, Eroberung, Zug, Tod, Herrschaft, Krieg, Iii, Kampf, Unterwerfung, Aufstand
  - Alexander, Perser, Babylon, Cyrus, Syrien, Israel, Euphrat, Aegypten, Asien, Darius
  - Reich, deutsch, Reiche, deutsche, Grenze, groß, Mark, Nation, Herrscher, Nachfolger
- Automatically detected named entities (200 topics, top 10 words):
  - Athen, Athener, Sparta, Athens, Griechenland, griechischen, Salamis, athenischen, Xerxes, Aristides
  - Joseph, Isaak, Abraham, Juden, Jakob, Gott, Kanaan, Israel, Abrahams, Gottes
  - Westfalen, Bielefeld, Minden, Paderborn, Dortmund, Herford, Arnsberg, Soest, Hagen, Hamm
- Sentiment Words (50 topics, top 10 words):
  - Recht, Schuld, hart, streng, Klage, verbieten, bestrafen, schwer, Weise, schuldig
  - Tod, Erbe, Besitz, Anspruch, Streit, Macht, Recht, anerkennen, Vertrag, Konstanz
  - führen, Macht, erklären, mächtig, unterstützen, unabhängig, vollständig, glänzend, Unterstützung, kräftig

Topic Modelling in *Welt der Kinder*

TECHNISCHE UNIVERSITÄT DARMSTADT

- Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003)
- Model variants
  - Number of topics: 20, 50, 100, 150, 200, 1000
  - Lemmatization, stopword filtering, ...
  - Word filtering
    - capitalized words (~nouns in German orthography)
    - named entities
    - sentiment words

Identifying Requirements: User-Centered Development

TECHNISCHE UNIVERSITÄT DARMSTADT

- Interviews and contextual inquiry with historians about previous projects
- Scenarios and wireframes as “boundary objects”
  - Virtual users from observations of real-world historic research processes
- Models and tools continuously adapted to emerging requirements
- Specificity vs. Sustainability
  - Approaches and technologies developed for project requirements, but as generical as possible

Topic Modelling: Identified Requirements

TECHNISCHE UNIVERSITÄT DARMSTADT

- Topics have to be homogeneous and easy to label
- Comparisons between topics, instead of analysis of terms within topics
- Topic correlations are relevant
  - Correlations between topics, co-occurrences of distinct topics
  - Examples: `France ~ Trade`? `France ~ War`? `France ~ Peace`?
  - Differences in correlations of topics within sub-collections
- Only a small part of the topics is helpful for research
- Macro-analysis with exemplary documents

Topic Modeling and Visualization

TECHNISCHE UNIVERSITÄT DARMSTADT

- Identify *relevant* topics for document / collection / task
- Give (visual) impression about the distribution of topics
  - Compare sub-collections by time, school type (confession, gender, level), region, …
- Approaches:
  - Display document-topic distribution over time in the current facet
  - DiTop (Oelke et al., 2014)
  - Serendip (Alexander et al., 2014)

Solr and Solritas

TECHNISCHE UNIVERSITÄT DARMSTADT

UKP

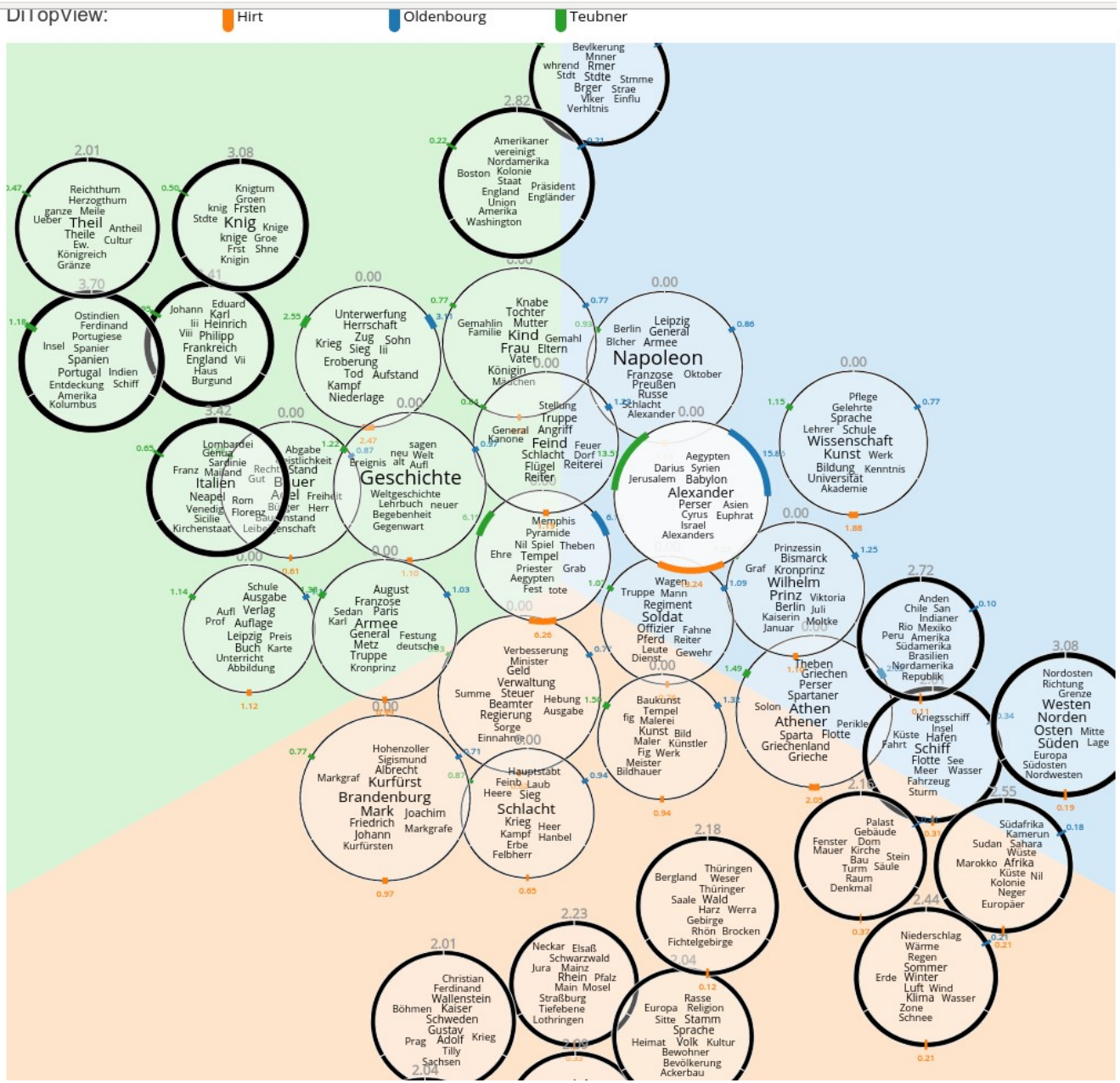
Solr and Solritas

TECHNISCHE UNIVERSITÄT DARMSTADT

- Scenario: users interested in certain topics, exploring and searching
- Add documents and topic proportions to Solr index
- Solritas web interface
  - Text searc
  - Topic proportion range queries:
    `topic1:[0.5 TO *] topic2:[0.0 TO 0.1]`
  - Most relevant topics shown for each document
  - Aggregation / intersection of document selection through facets
    - Documents associated with topic *X* from model *A*
    - Documents from publisher *P*, period *Y*, …

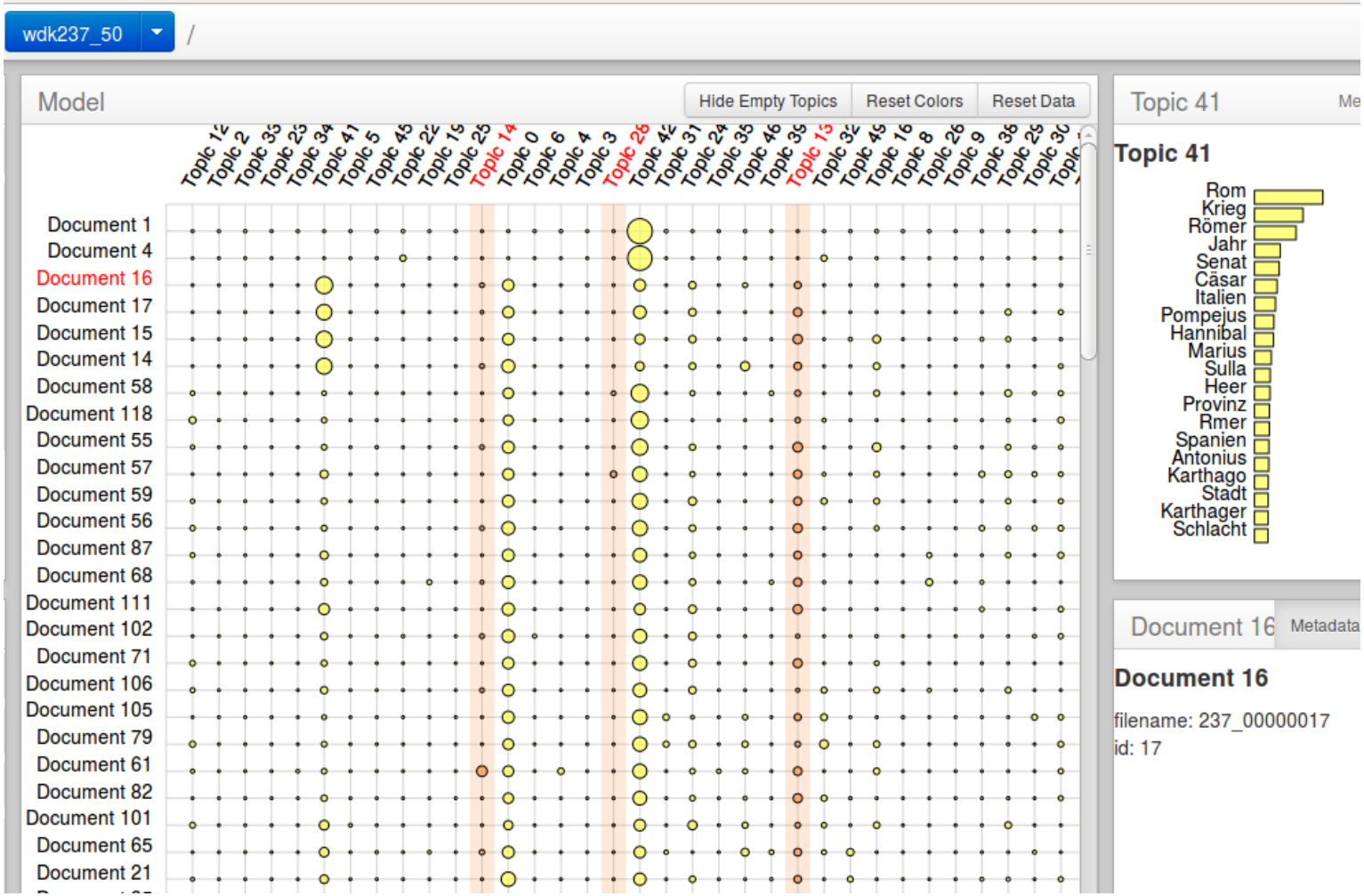
DiTop

- Oelke, Daniela, Hendrik Strobelt, Christian Rohrdantz, Iryna Gurevych, and Oliver Deussen. “Comparative Exploration of Document Collections: A Visual Analytics Approach.” In Computer Graphics Forum, 33:201–10. Wiley Online Library, 2014. http://onlinelibrary.wiley.com/doi/10.1111/cgf.12376/full.



Serendip

- Eric Alexander, Joe Kohlmann, Robin Valenza, Michael Witmore, and Michael Gleicher. “Serendip: Topic Model-Driven Visual Exploration of Text Corpora”. In *Proceedings of the 2014 IEEE Conference on Visual Analytics Science and Technology (VAST)* — November 2014



Opinion Mining: *Welt der Kinder* Examples

TECHNISCHE UNIVERSITÄT DARMSTADT

- Obvious opinions
  - `Luther war der größte Mann seiner Zeit.`
    "Luther was the greatest man of his time."
  - `der schreckliche Bauernkrieg`
    "the terrible Peasant War"
  - `hochmütige Franzosen`
    „arrogant Frenchmen"
- Subtle / complex opinion (requiring world knowledge)
  - `Und so förderte die Bibelübersetzung das Reformationswerk unermeßlich.`
    "So the translation of the bible immensely promoted the reformation work."
  - `Japan, das asiatische England`
    "Japan, the Asian England"

Opinion Mining

TECHNISCHE UNIVERSITÄT DARMSTADT

- Little annotated data available in German in general
- Even less (no) data for historical data
- Weakly supervised approaches
  - “Joint Sentiment/Topic Model for Sentiment Analysis” (Lin and He 2009)
  - “Thumbs up or Thumbs Down?”, (Turney, 2002)
  - “Sentiment Analysis on the People’s Daily.” (Li and Hovy, 2014)
    - In a single document (news article), sentiment towards an entity is consistent
    - Over a certain period of time, sentiments towards an entity are inter-related

Summary

TECHNISCHE UNIVERSITÄT DARMSTADT

- Welt der Kinder
  - Research on historic German textbooks
    - Digitized corpus of 6,000 textbooks and juvenile books
  - Combine traditional methods and natural language processing technologies
- Natural Language Processing
  - OCR (post-processing)
  - Topic modelling
  - Opinion mining
  - Named Entity Recognition
  - ...

Thanks

- Georg Eckert Institute for International Textbook Research
- Ubiquitous Knowledge Processing Lab (UKP) TU Darmstadt, and German Institute for International Educational Research (DIPF)
- Institute for Information Science and Language Technology in Hildesheim (IWIST)
- Institute of Popular Culture Studies, University of Zurich
- Bavarian State Library
- Göttingen Centre for Digital Humanities
- Universitätsbibliothek Technische Universität Braunschweig