TopicExplorer - Qualitative Text Mining for Humanities and Social Sciences

Alexander Hinneburg\textsuperscript{1} Christian Oberländer\textsuperscript{2} Christian Papilloud\textsuperscript{3} Anne Purschwitz\textsuperscript{4}

\textsuperscript{1}Computer Science, Martin-Luther-University Halle-Wittenberg

\textsuperscript{2}Japanese Studies, Martin-Luther-University Halle-Wittenberg

\textsuperscript{3}Social Science, Martin-Luther-University Halle-Wittenberg

\textsuperscript{4}History, IZEA, Martin-Luther-University Halle-Wittenberg

May 2017

https://blogs.urz.uni-halle.de/topicexplorer/
Outline

Introduction

TopicExplorer
   Visualization techniques and user interface
   TopicFrames
   Topics and Meta Data

OCR and Topic Models

Conclusion
Introduction

- Qualitative text analysis is a methodological foundation for many disciplines in the humanities and social sciences
- Main challenges for software support
  - corroboration of results
  - semi-automatic, content-based preparation, indexing and summarization of large text collections
- Goal of the DFG grant application
  - build an infrastructure for the TopicExplorer system to offer qualitative text mining as a self service
  - Use case: media impact analysis
  - Use case: deep interview analysis & history of sociology
  - Use case: Halle journals of the 18th century includes application of OCR software for historic printings
- TopicExplorer
  - Visualization techniques and user interface
  - Topics and meta data for context information
Outline

Introduction

**TopicExplorer**
- Visualization techniques and user interface
- TopicFrames
- Topics and Meta Data

**OCR and Topic Models**

Conclusion
Short Overview of Topic Models [BNJ03, Ble12]

- Latent Dirichlet Allocation (LDA) assumes that a document is generated as follows
  - draw topic mixture proportions $\theta$ of document from Dirichlet distribution
  - Generate words of the document as follows
    - draw topic $z$ of the word to be generated from multinomial with parameters $\theta$
    - draw word $w$ from topic specific multinomial over vocabulary that is indicated by $z$ and parameterized by $\varphi$

- LDA-Model

- Inference estimates distributions over hidden variables $z$, $\theta$, $\varphi$ ⇒ each word is assigned to a topic
Outline

Introduction

**TopicExplorer**
Visualization techniques and user interface
  TopicFrames
  Topics and Meta Data

OCR and Topic Models

Conclusion
Overview of TopicExplorer User Interface [HPS12, HRPO14]

https://blogs.urz.uni-halle.de/topicexplorer/
Similarity-based Layout of Topics

- **Algorithm**
  - Compute all pairwise topic similarities
  - Compute hierarchical clustering
  - Layout dendrogram tree, each leaf is a topic
- **Rainbow color mapping reflects computed topic order**
  - Color acts like a visual hash function: similar color indicates similar topic with high likelihood.
- **Color provides orientation for the user at three levels**
  - Topics
  - Document browsing and ranking
  - Word assignments of single Documents
Document Browsing and Ranking

- Documents are ranked by decreasing topic affinity
  - Colored circles indicate the four most important topics
  - Color acts as visual hash function
Document Inspection

- Document is represented with topic assignments of words
  - Color acts as visual hash function
Hierarchical Topics

- Initial situation

- Preview of topic merging according to topic hierarchy

- Merged topic, (allows interactive splitting to reverse merging)
Outline

Introduction

**TopicExplorer**
- Visualization techniques and user interface
  - **TopicFrames**
  - Topics and Meta Data

**OCR and Topic Models**

Conclusion
Motivation of TopicFrames [HRPO14]

- **TopicExplorer**

- **Problem**
  - Topic models give no guarantees on interpretability
  - Not all topics are well interpretable
  - Topics represented by word lists may leave room for several different interpretations
Towards interpretable Topic Representations

- Most probable words is not always the best option
- More context of top words is helpful
- Results on topic coherence
  - Humans find a topic interpretable if pairs of top words often appear close together in documents [Lau et al. EACL 2014]
- Observation
  - Top word lists of topics are often dominated by nouns, while verbs are less prominent
- Frames
  - Basic communication units for semantic contents
  - Noun-verb co-locations
- Topic Frames
  - Topic specific noun-verb co-location
- How to define and compute topic frames?
Topic Frames

Definition

- **Topic frame** occurs when a noun and a verb appear close together in a document and the topic model assigned both tokens to the same topic.

Topics with top word list that are not clearly interpretable.

Topics frames make topics more clearly interpretable.
Workflow and Implementation

- Linguistic data preparation
  - Tokenization, important for Japanese documents
  - Part-of-speech tagging
  - Lemmatization
- Topic modeling
  - work with word lemmata
  - translate topic assignments back to original documents
- Topic frame detection
  - find noun-verb co-locations with tokens assigned to same topic
  - filter frames occurrences with sentence delimiters
  - count different topic frames and number of frame occurrences

Topic Frame Occurrence

Keywords with Topics
Outline

Introduction

**TopicExplorer**
Visualization techniques and user interface
**TopicFrames**
**Topics and Meta Data**

**OCR and Topic Models**

Conclusion
Topics and Meta Data: Time

- Document Collection consists of Blogs with Time Stamps
- Temporal Information is not used for Topic Modeling
- Instead: token assignments to topics are analyzed over time
Topics and Meta Data: General Case

- Documents come with all kinds of meta data
- Texts, topic model and meta data are stored in a relational database [RH16]
  - Derived linguistic meta data: POS, NER, Sentiment, ...
  - Geographic entities mentioned in documents
  - Images, Videos, Links, ...
- Token assignments to topics can be analyzed in context of meta-data entities using
  - Aggregation and/or Intra-Document Proximity
- Topic-Explorer uses automatically configured workflows to allow future extension to all kind of meta data:
Topics and Meta Data: General Case

- Documents come with all kinds of meta data
- Texts, topic model and meta data are stored in a relational database [RH16]
  - Derived linguistic meta data: POS, NER, Sentiment, ...
  - Geographic entities mentioned in documents
  - Images, Videos, Links, ...
- Token assignments to topics can be analyzed in context of meta-data entities using
  - Aggregation and/or Intra-Document Proximity
- Topic-Explorer uses automatically configured workflows to allow future extension to all kind of meta data:
Outline

Introduction

TopicExplorer
  Visualization techniques and user interface
  TopicFrames
  Topics and Meta Data

OCR and Topic Models

Conclusion
OCR for historic printings

- Abbyy (commercial) and Tesseract (open source) need extensive training to handle historic printings
- New neural networks techniques for OCR are available: OCRopus
- Neural networks may require less and simpler training [Spr16]
- DFG funds OCR-D project\(^1\) for enhancing available OCR software

\(^1\)http://ocr-d.de
Combining OCR and Topic Models

▶ Evaluating models of latent document semantics in the presence of OCR errors [WLR10]

Though model quality declines as errors increase, simple feature selection techniques enable the learning of relatively high quality models even as word error rates approach 50%.

▶ Towards Noise-resilient Document Modeling [YL11]

▶ On handling textual errors in latent document modeling [YL13]

Using both real and synthetic data sets with varying degrees of errors, our TDE-LDA model outperforms: (1) the traditional LDA model by 16%-39% (real) and 20%-63% (synthetic); and (2) the state-of-the-art N-Grams model by 11%-27% (real) and 16%-54% (synthetic).
Outline

Introduction

TopicExplorer
  Visualization techniques and user interface
  TopicFrames
  Topics and Meta Data

OCR and Topic Models

Conclusion
Conclusion

- Topic models are helpful for qualitative text analysis
- TopicExplorer User-Interface helps to go from statistical models towards semantic interpretations
- Future Work
  - evaluate new OCR methods for historic printings
  - combine OCR and topic modeling
David M. Blei.

Probabilistic topic models.

David M Blei, Andrew Y Ng, and Michael I Jordan.

Latent dirichlet allocation.

Alexander Hinneburg, Rico Preiss, and René Schröder.

Topicexplorer: Exploring document collections with topic models.

Alexander Hinneburg, Frank Rosner, Stefan Pessler, and Christian Oberländer.

Exploring document collections with topic frames.
Frank Rosner and Alexander Hinneburg.
Translating bayesian networks into entity relationship models.

Uwe Springmann.
OCR für alte Drucke.

Evaluating models of latent document semantics in the presence of ocr errors.

Tao Yang and Dongwon Lee.
Towards noise-resilient document modeling.

Tao Yang and Dongwon Lee.

On handling textual errors in latent document modeling.