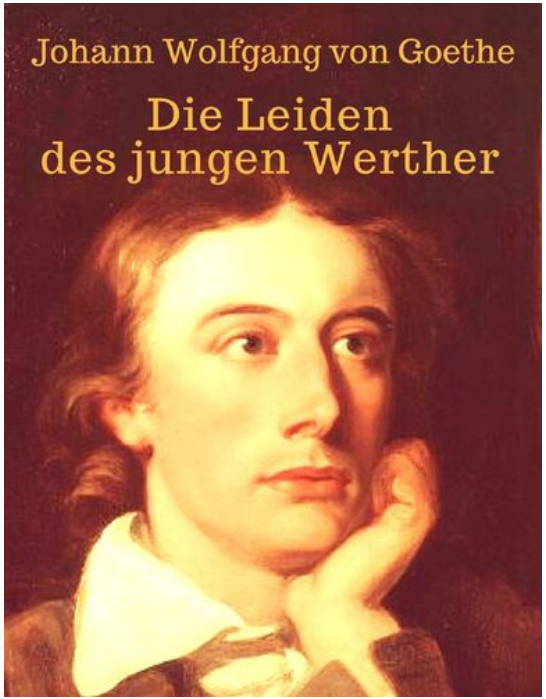
forTEXT

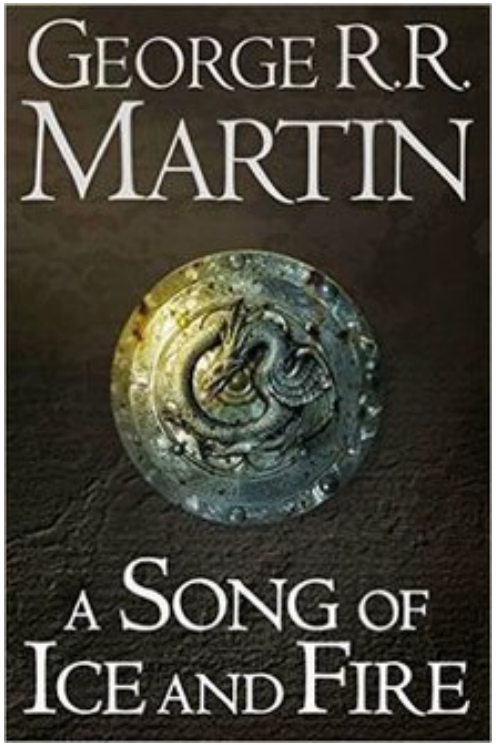
https://fortext.net

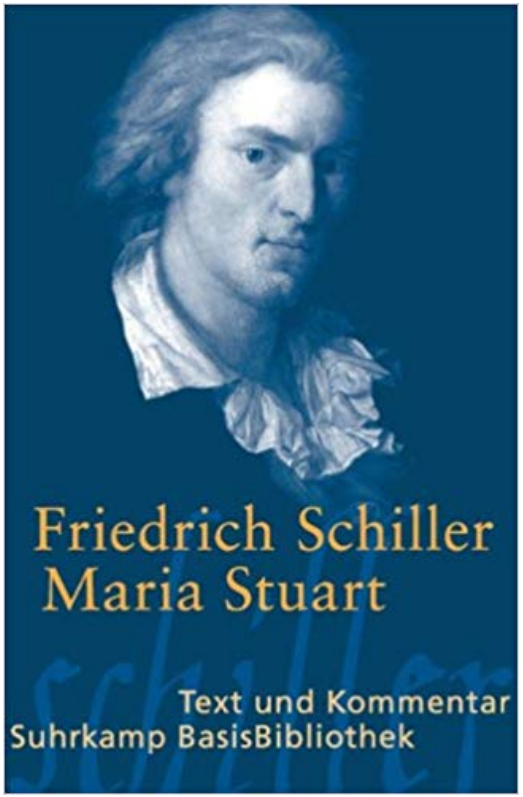
Topic Modeling mit dem DARIAH Topics Explorer

gefördert durch die
DFG Deutsche Forschungsgemeinschaft

Themen?


Johann Wolfgang von Goethe
Die Leiden des jungen Werther


GEORGE R.R. MARTIN
A SONG OF ICE AND FIRE


Friedrich Schiller
Maria Stuart
Text und Kommentar
Suhrkamp BasisBibliothek

Topics?

- Thema
- Topos
- Stoff
- Topic

Ein probabilistisches Verfahren …

$$P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi}; \alpha, \beta) = \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{t=1}^{N} P(Z_{j,t} \mid \theta_j) P(W_{j,t} \mid \varphi_{Z_{j,t}}),$$

Bitte was?

- Ein Wort kommt mit einer bestimmten Wahrscheinlichkeit in einem Topic vor
- Ein Topic kommt mit einer bestimmten Wahrscheinlichkeit in einem Dokument vor

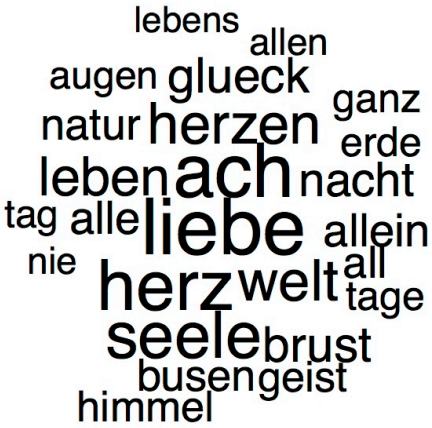
→ festlegen: Menge an Topics, Größe der Topics, Stopwords, Menge an Durchläufen, Größe der Textchunks

Was sind Textchunks?

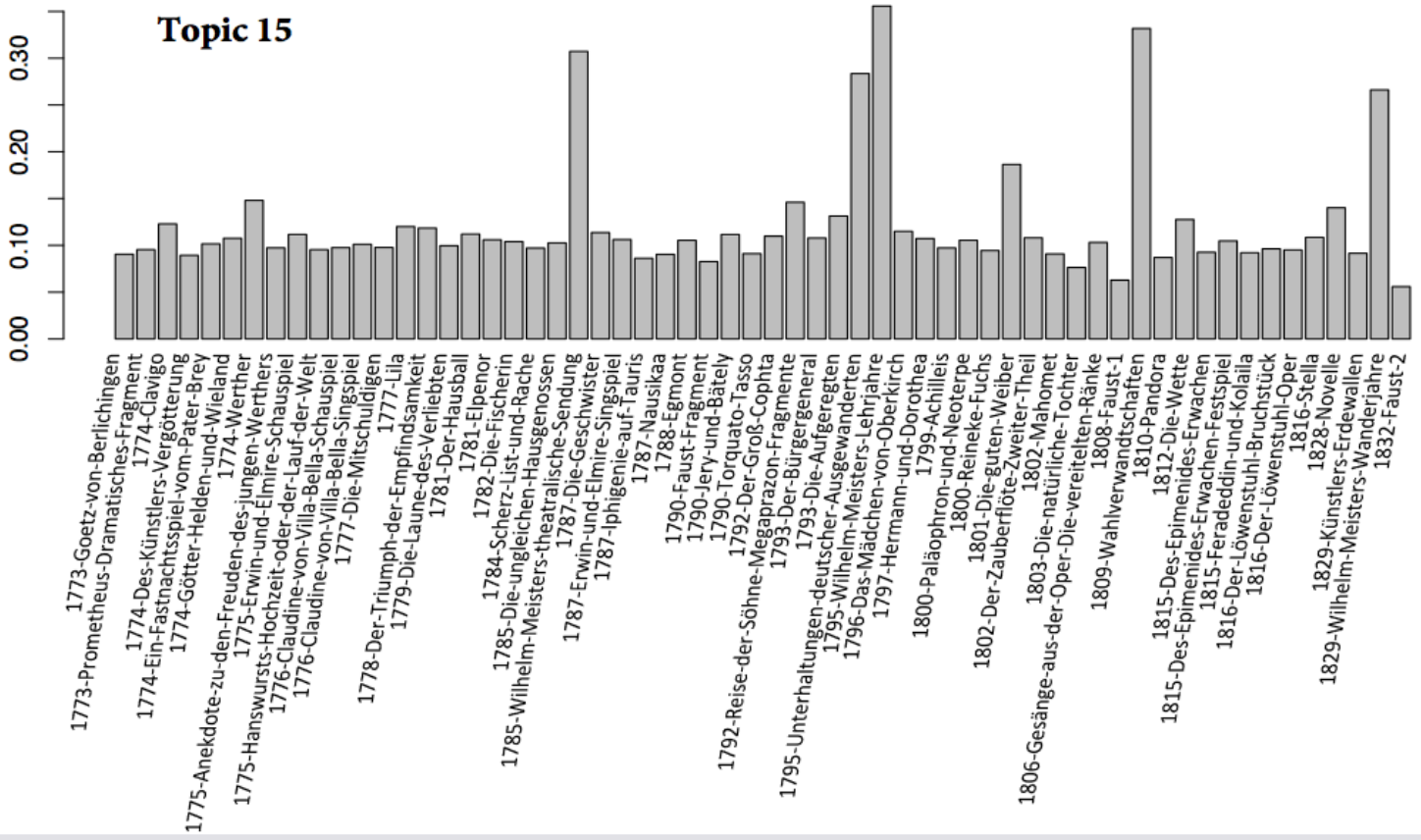
Preprocessing

- Die Texte des Korpus werden in gleich große Segmente (= *chunks*) aufgeteilt (z.B. 1000 Wörter)
- Ein *Part-of-Speech-Tagging* (POS-Tagging) ermöglicht, nur bestimmte Wortarten zu modellieren
- Eine Lemmatisierung vereinheitlicht Wortformen
- Eine NER (*Named Entity Recognition*) ermöglicht es, Eigennamen gebündelt auszuschließen
- Die Stoppwortliste enthält die MFW (*Most Frequent Words*) und kann beliebig erweitert werden

Beispieltopics Goethe



Verteilung eines Topics auf die Texte des Goethekorpus



Danke!

forTEXT: https://fortext.net
CATMA: https://catma.de

Follow us on …

Twitter: @fortext_catma
YouTube: forTEXT & CATMA
Pinterest: forTEXT – Digital Humanities
Facebook: @forTEXTundCATMA

forTEXT
Literatur digital erforschen

CATMA
FOR UNDOGMATIC TEXTUAL MARKUP AND ANALYSIS