

forTEXT

Literatur digital erforschen

forTEXT

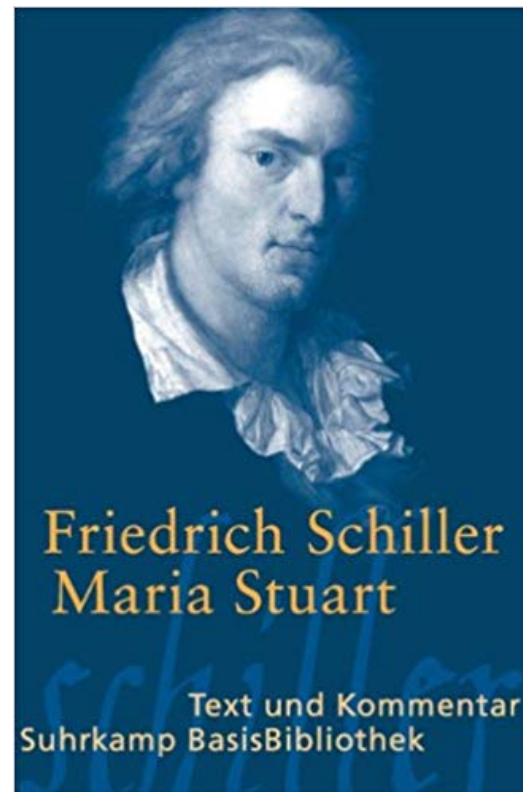
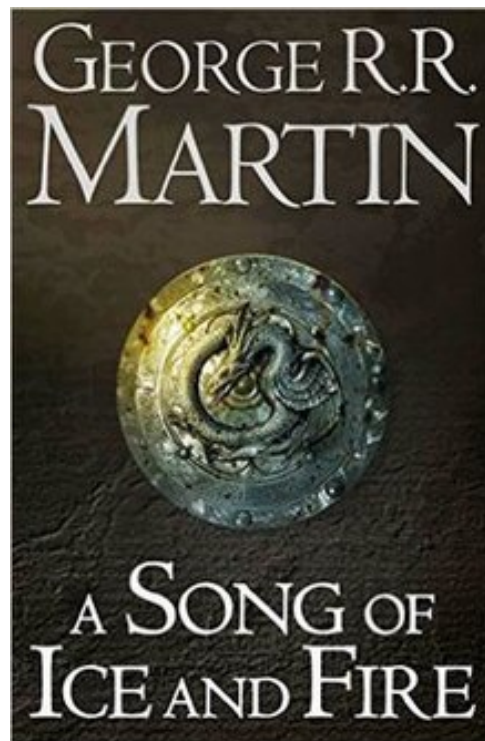
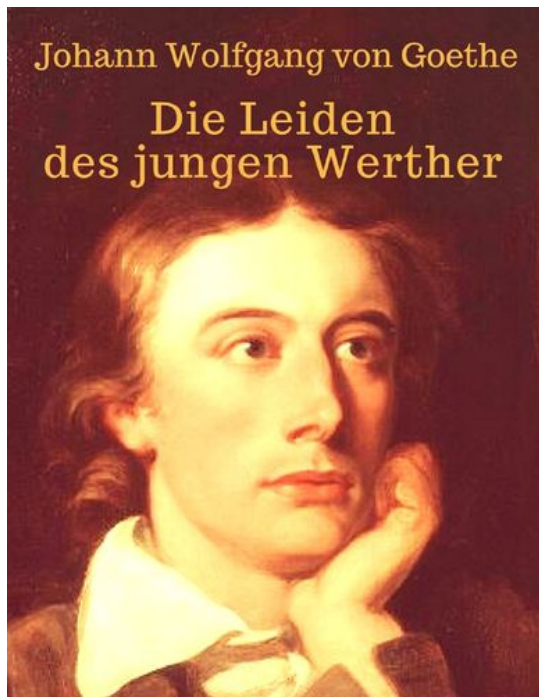
<https://fortext.net>

Topic Modeling mit dem DARIAH Topics Explorer

gefördert durch die

DFG Deutsche
Forschungsgemeinschaft

Themen?



Topics?

- Thema
- Topos
- Stoff
- Topic



Ein probabilistisches Verfahren ...

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

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

schwanken
 donner schoenen
 sirenen
 tales land schiffer
 ans schiff ruder
 inseln see ufer sturm
 wind wasser segel
 stelle kahn bald woge
 wellen regen
 haenden

begierde genuss
 freund leicht groesse
 dachte dichter erfunden
 innersten seele wornach
 lebendig jagd
 gewerbe ideen zugleich
 waeren welt irrst gefuehle
 jeher besitzthuemer
 papieren hinueber
 empfindungen

instrumente
 saenger
 melodie sang
 alte lied stimme
 harfe ohr gesang ton
 stimmen begann musik toene
 singen lieder
 kennst hoeren dahin
 toenen tanz
 instrument

werner
 geschrieben
 schrieb briefe kapitel
 schriff blatt lesen capitel
 worte buch buecher
 erstes inhalt brief las schriften
 begriff schreiben lehrjahre
 gelesen papier
 meisters

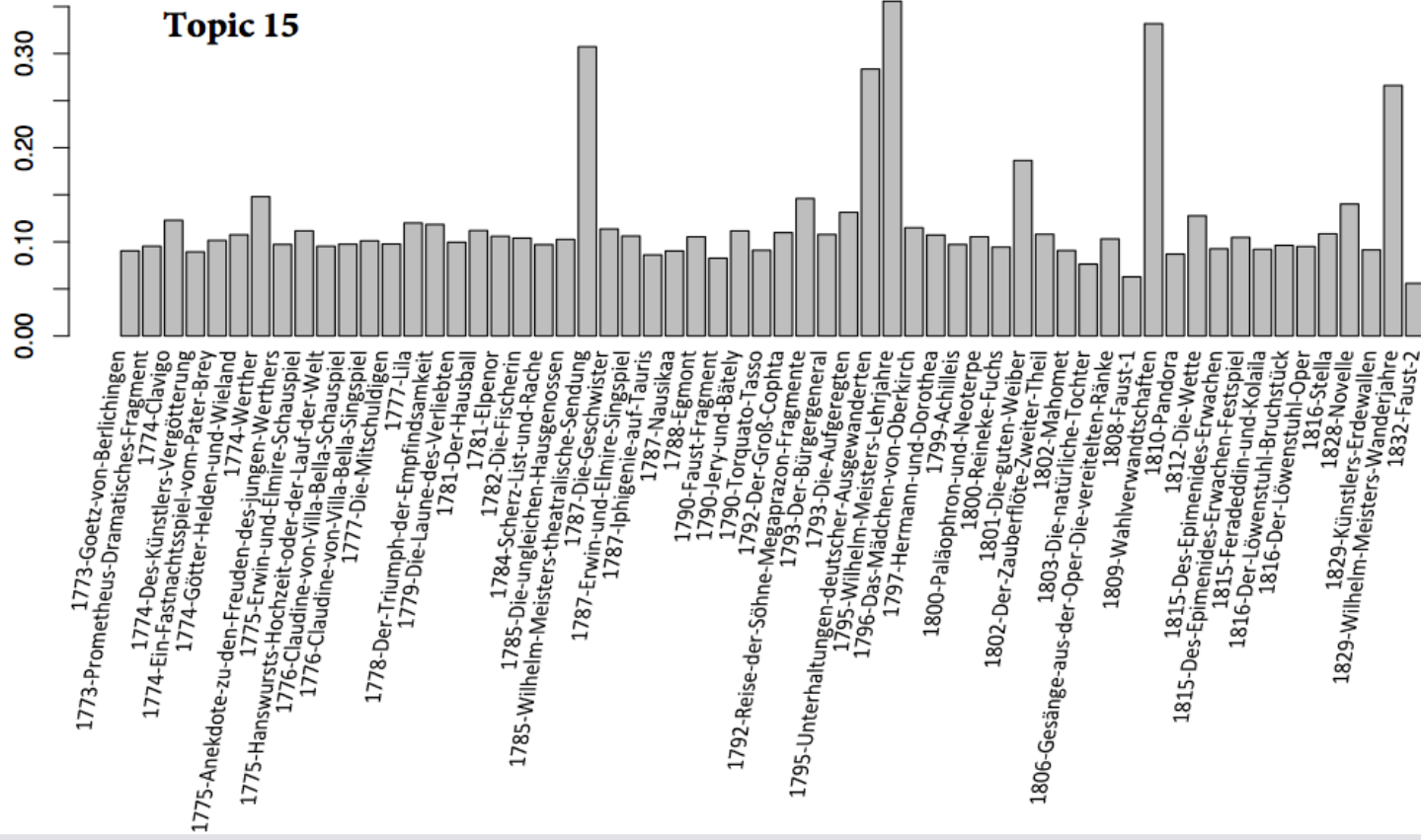
lebens allen
 augen glueck ganz
 natur herzen erde
 leben ach nacht
 tag alle liebe allein
 nie herz welt all
 seele brust
 busengeist
 himmel

vergleichen
 stuecken dichtkunst
 verwunderung gegenden
 aussen welt
 mahren zugleich
 meere dichter innen
 held menschen
 grund frische leicht goetter
 gegoennt huedsch dichtung
 gedichte entwickelt
 verliebten
 poetischen

zuschauer
 stellen spielen
 oft sehr ganz
 rollen stuecke stueck recht
 groessten theater spielte
 schauspieler spiel
 einige rolle uebrigen
 mensch beifall gespielt
 personen besonders
 gesellschaft

schmuck
 laube goldenen
 baeume schoenen
 schoen blaetter goldene
 zweige garten frucht
 kleider blumen rosen
 krone fruechte tragen
 strauss haupt kranz baum
 schoene wiese
 schoensten voegel

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

