b

UNIVERSITÄT
BERN

Text and Data Mining (TDM): Ressourcen & APIs, Dos & Don’ts

ResearchSkills@vonRoll

TDM Week@vonRoll: Text und Data Mining für die Sozialwissenschaften

Kathi Woitas, Digital Scholarship Specialist

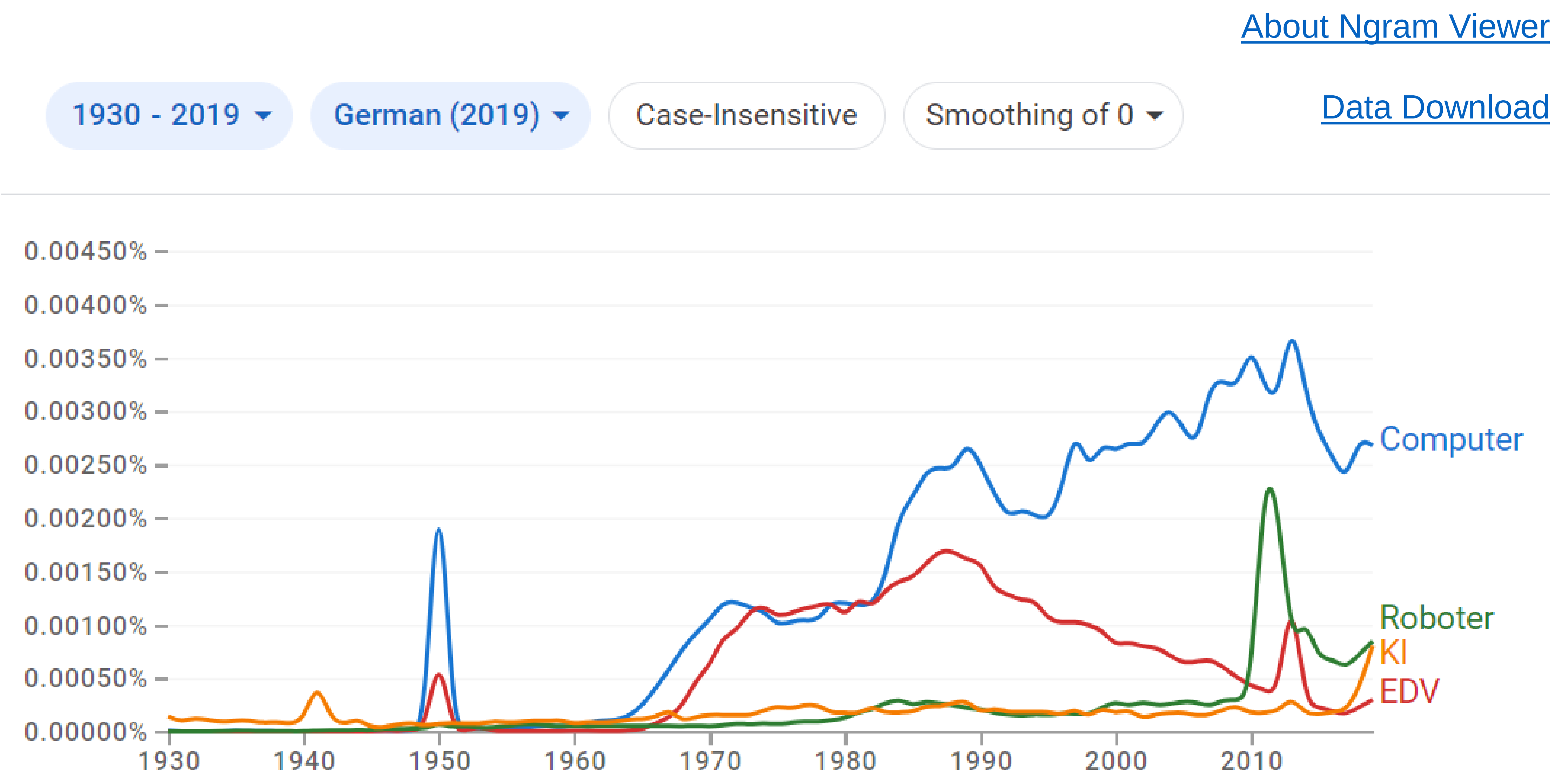
Universitätsbibliothek Bern, Digitale Dienste

Kathi Woitas

- Digital Scholarship Services @ Universitätsbibliothek Bern
  - Beratung, Lizenzierung, Schulung von TDM-Ressourcen
  - Projektleitung Digital Collections Bern (DCB)
  - Datenbezogene Projekte

- Bibliothekswissenschaft, Europäische Ethnologie (HU Berlin)
- verschiedene WB in Data Science: Datenanalyse, Statistical Modelling, Practical Machine Learning, Big Data

Google Books Ngram Viewer



About Ngram Viewer

Data Download

1930 - 2019 ▾

German (2019) ▾

Case-Insensitive

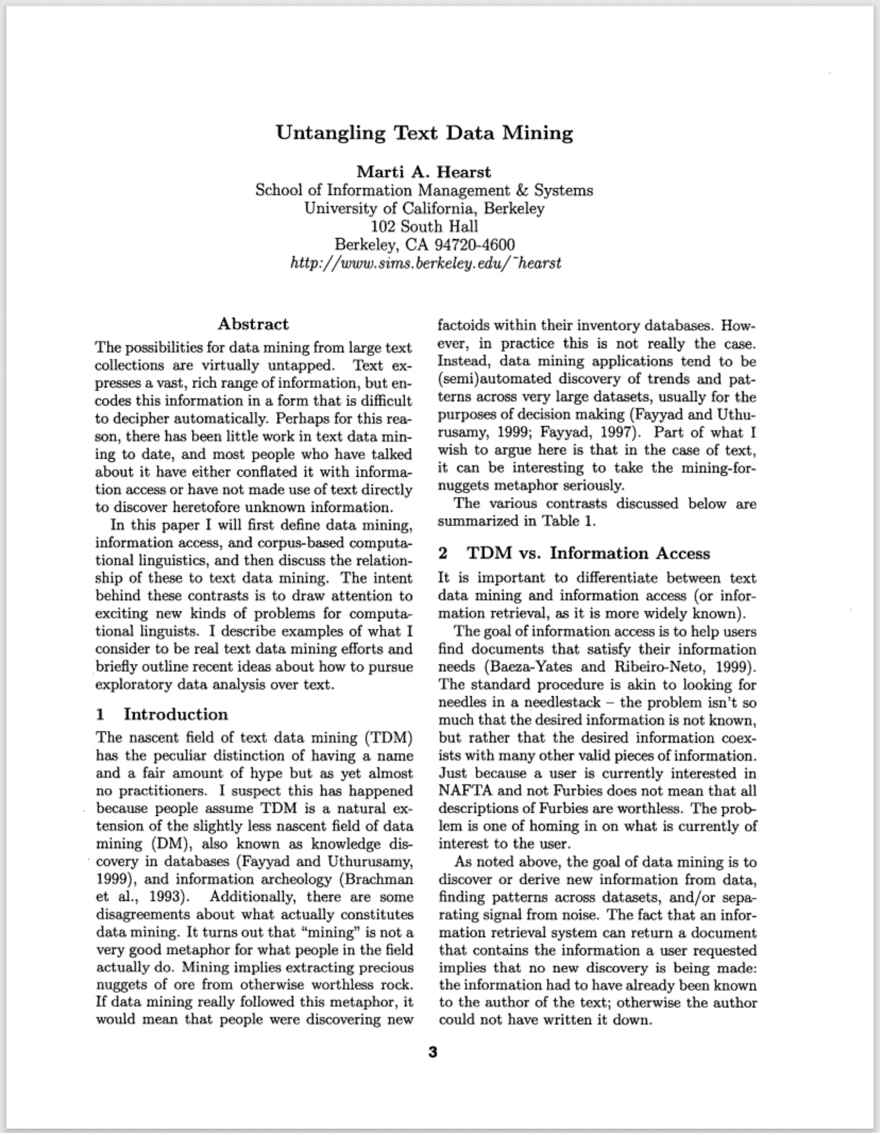
Smoothing of 0 ▾

Text and Data Mining (TDM)?

*As noted above, the goal of **data mining** is to **discover** or **derive new information** from data, **finding patterns** across datasets, and/or separating signal from noise.*

*If we **extrapolate** from data mining (as practiced) on numerical data **to data mining from text collections**, we discover that there already exists a field engaged in text data mining: corpus-based computational linguistics!*

→ Text as Data, Textdatenanalyse



Untangling Text Data Mining

Marti A. Hearst
School of Information Management & Systems
University of California, Berkeley
102 South Hall
Berkeley, CA 94720-4600
http://www.sims.berkeley.edu/˜hearst

Abstract

The possibilities for data mining from large text collections are virtually untapped. Text expresses a vast, rich range of information, but encodes this information in a form that is difficult to decipher automatically. Perhaps for this reason, there has been little work in text data mining to date, and most people who have talked about it have either conflated it with information access or have not made use of text directly to discover heretofore unknown information.

In this paper I will first define data mining, information access, and corpus-based computational linguistics, and then discuss the relationship of these to text data mining. The intent behind these contrasts is to draw attention to exciting new kinds of problems for computational linguists. I describe examples of what I consider to be real text data mining efforts and briefly outline recent ideas about how to pursue exploratory data analysis over text.

1 Introduction

The nascent field of text data mining (TDM) has the peculiar distinction of having a name and a fair amount of hype but as yet almost no practitioners. I suspect this has happened because people assume TDM is a natural extension of the slightly less nascent field of data mining (DM), also known as knowledge discovery in databases (Fayyad and Uthurusamy, 1999), and information archeology (Brachman et al., 1993). Additionally, there are some disagreements about what actually constitutes data mining. It turns out that "mining" is not a very good metaphor for what people in the field actually do. Mining implies extracting precious nuggets of ore from otherwise worthless rock. If data mining really followed this metaphor, it would mean that people were discovering new factoids within their inventory databases. However, in practice this is not really the case. Instead, data mining applications tend to be (semi)automated discovery of trends and patterns across very large datasets, usually for the purposes of decision making (Fayyad and Uthurusamy, 1999; Fayyad, 1997). Part of what I wish to argue here is that in the case of text, it can be interesting to take the mining-for-nuggets metaphor seriously.

The various contrasts discussed below are summarized in Table 1.

2 TDM vs. Information Access

It is important to differentiate between text data mining and information access (or information retrieval, as it is more widely known).

The goal of information access is to help users find documents that satisfy their information needs (Baeza-Yates and Ribeiro-Neto, 1999). The standard procedure is akin to looking for needles in a needlestack – the problem isn't so much that the desired information is not known, but rather that the desired information coexists with many other valid pieces of information. Just because a user is currently interested in NAFTA and not Furbies does not mean that all descriptions of Furbies are worthless. The problem is one of homing in on what is currently of interest to the user.

As noted above, the goal of data mining is to discover or derive new information from data, finding patterns across datasets, and/or separating signal from noise. The fact that an information retrieval system can return a document that contains the information a user requested implies that no new discovery is being made: the information had to have already been known to the author of the text; otherwise the author could not have written it down.

Hearst, M. A. Untangling text data mining. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. S. 3–10 (Association for Computational Linguistics, 1999). DOI: 10.3115/1034678.1034679

Workflow Text and Data Mining / NLP

1 Datenzugang

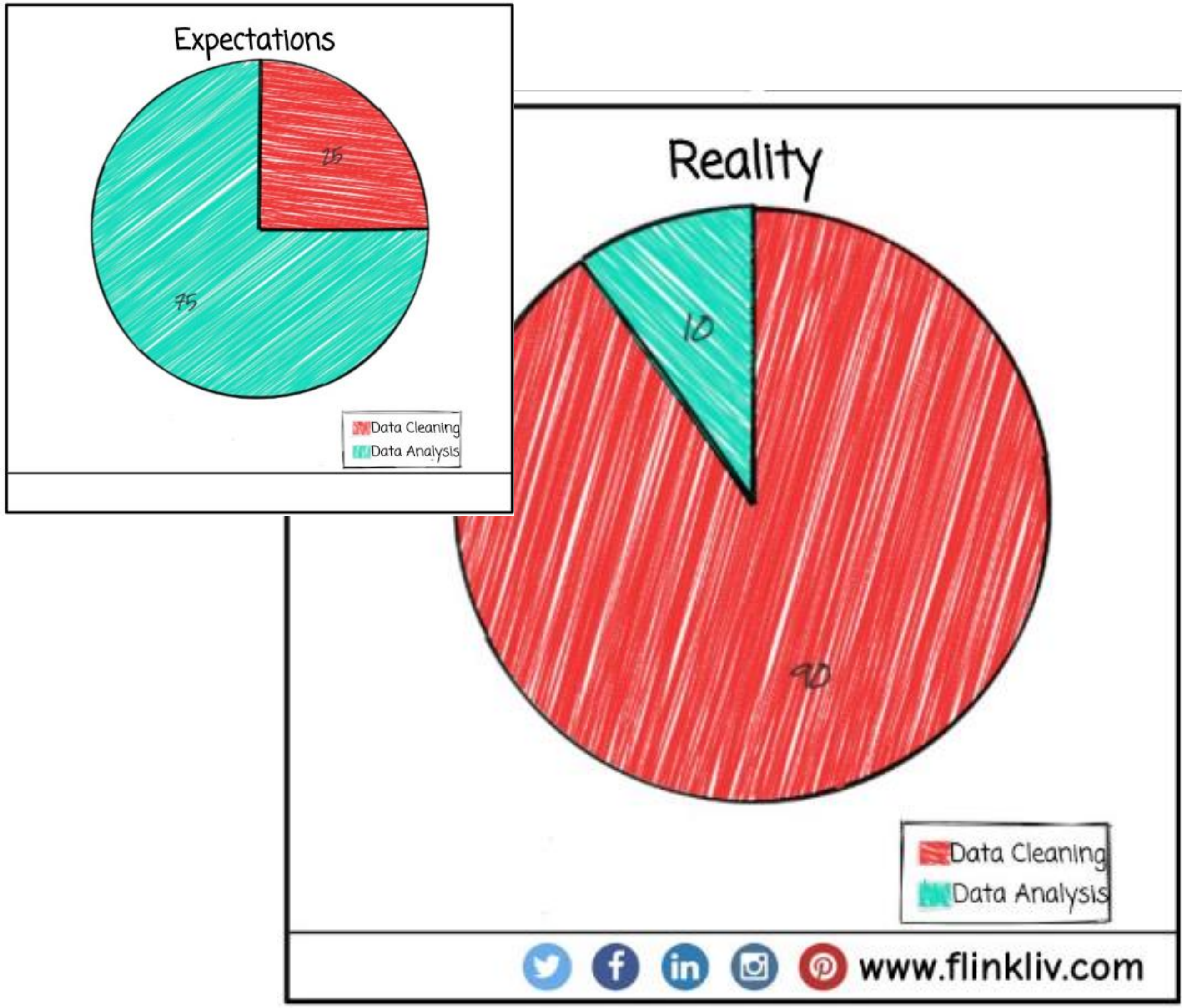
2 Pre-Processing

3 Statistische Analyse

4 Vektorisierung

5 Training/Modellentwicklung

6 Auswertung/Interpretation



Details + Bsp. siehe [DOI: 10.5281/zenodo.6417539](https://doi.org/10.5281/zenodo.6417539)

Datenzugänge

APIs vs. Dumps

Dumps

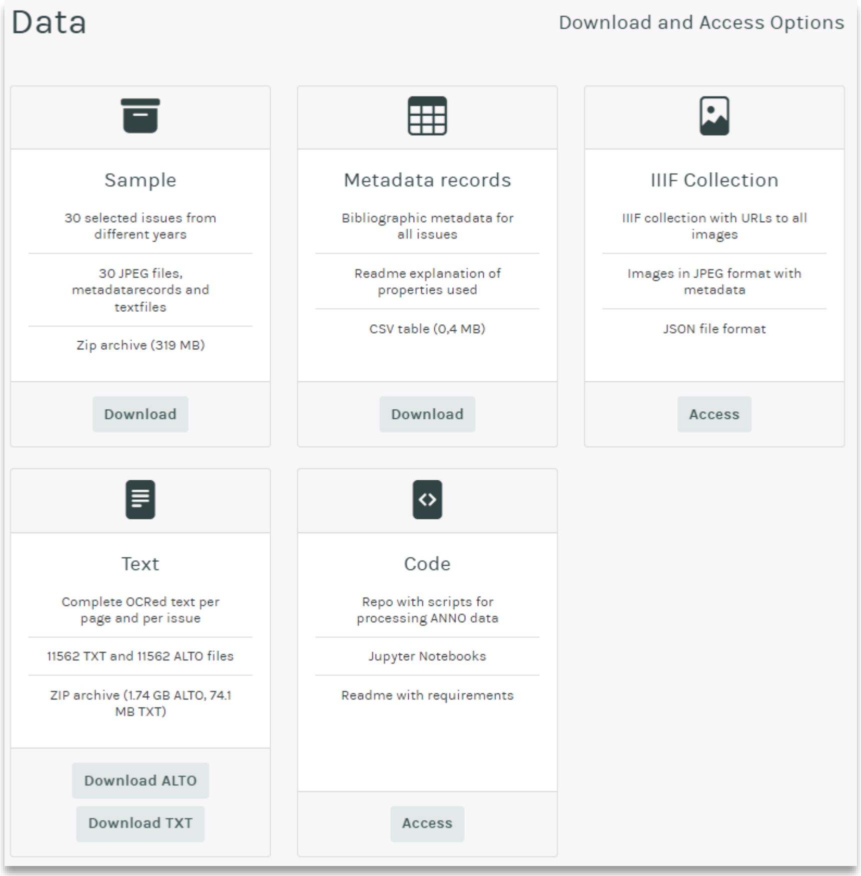
- «Gesamtdownload» zu einem bestimmten Zeitpunkt (Snapshot, Bulk)
- je nach Datenquelle/-inhalt sehr gross und Handling schwierig
- zumeist geringere Aktualität

APIs

- zielgerichtete Datenabfrage
- zumeist grössere Aktualität
- Kennenlern-Aufwand
  - Endpoints, Inhalte
  - Abfragesyntax, Suche, Filter etc.
- oft Registrierung + Authentifizierung per User Key/API Token

Datenzugang

Bsp. frei zugängliche Quellen



Data

Download and Access Options

Sample

30 selected issues from different years

30 JPEG files, metadatarecords and textfiles

Zip archive (319 MB)

Metadata records

Bibliographic metadata for all issues

Readme explanation of properties used

CSV table (0,4 MB)

IIIF Collection

IIIF collection with URLs to all images

Images in JPEG format with metadata

JSON file format

Text

Complete OCRed text per page and per issue

11562 TXT and 11562 ALTO files

ZIP archive (1.74 GB ALTO, 74.1 MB TXT)

Code

Repo with scripts for processing ANNO data

Jupyter Notebooks

Readme with requirements

Bibliografische Daten & drumherum

- OpenAlex (CC0, umfangreiche Daten-Aggregation)
- Bulk Bibliographic Metadata (Internet Archive)

Sammlungen aus Archiven, Bibliotheken etc.

- z.B. Europeana, diverse Nationalbib., Bundesarchiv

Online-Quellen

- CLARIN Corpora, awesome-public-datasets, …

→ Übersicht TDM-Ressourcen + APIs auf DS-Website

Datenzugang

Bsp. lizenzierte Quellen

Bibliografische Daten & drumherum

- Dimensions (auf Anfrage); ~~WoS, Scopus (minimal)~~

Volltexte von wiss. Publikationen (non-OA)

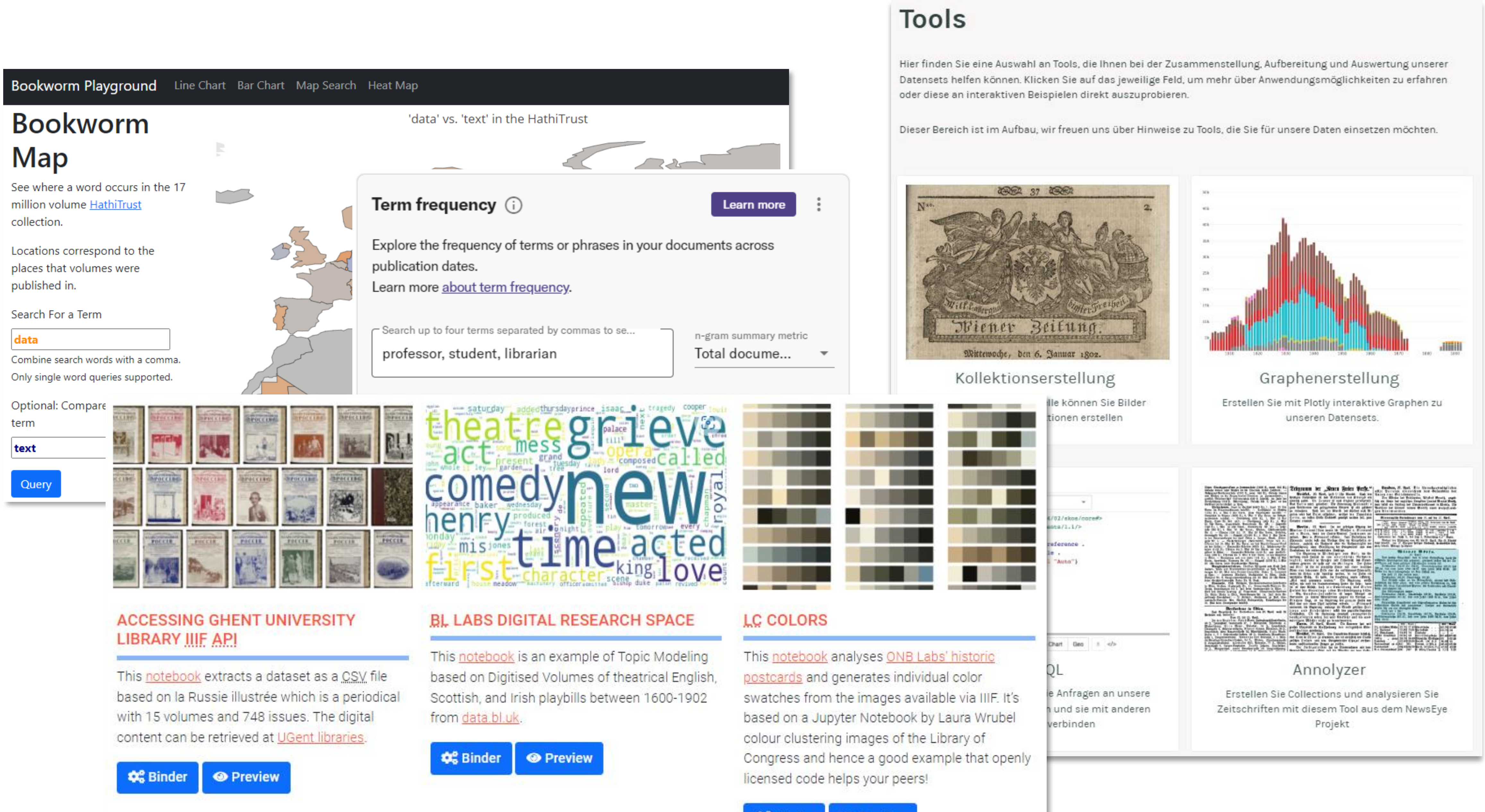
- TDM-Lizenz: Elsevier, Springer Nature, Wiley, IEEE

weiterer “Bibliotheks-Content”

- z.B. Bücher: HathiTrust Research Center
- CH-Zeitungen: SwissDox@LiRI (→ Donnerstag!)

→ Übersicht TDM-Ressourcen + APIs auf DS-Website

Tools für Visualisierung + Analyse



Bookworm Playground Line Chart Bar Chart Map Search Heat Map

Bookworm Map

See where a word occurs in the 17 million volume HathiTrust collection.

Locations correspond to the places that volumes were published in.

Search For a Term

data

Combine search words with a comma.
Only single word queries supported.

Optional: Compare term

text

Query

'data' vs. 'text' in the HathiTrust

Term frequency ⓘ

Learn more

Explore the frequency of terms or phrases in your documents across publication dates.
Learn more about term frequency.

Search up to four terms separated by commas to se...

professor, student, librarian

n-gram summary metric

Total docume...

Tools

Hier finden Sie eine Auswahl an Tools, die Ihnen bei der Zusammenstellung, Aufbereitung und Auswertung unserer Datensets helfen können. Klicken Sie auf das jeweilige Feld, um mehr über Anwendungsmöglichkeiten zu erfahren oder diese an interaktiven Beispielen direkt auszuprobieren.

Dieser Bereich ist im Aufbau, wir freuen uns über Hinweise zu Tools, die Sie für unsere Daten einsetzen möchten.

Kollektionserstellung

...lle können Sie Bilder
...tionen erstellen

Graphenerstellung

Erstellen Sie mit Plotly interaktive Graphen zu unseren Datensets.

...QL

...e Anfragen an unsere
...und sie mit anderen
...verbinden

Annolyzer

Erstellen Sie Collections und analysieren Sie Zeitschriften mit diesem Tool aus dem NewsEye Projekt

ACCESSING GHENT UNIVERSITY LIBRARY IIIF API

This notebook extracts a dataset as a CSV file based on la Russie illustrée which is a periodical with 15 volumes and 748 issues. The digital content can be retrieved at UGent libraries.

Binder Preview

BL LABS DIGITAL RESEARCH SPACE

This notebook is an example of Topic Modeling based on Digitised Volumes of theatrical English, Scottish, and Irish playbills between 1600-1902 from data.bl.uk.

Binder Preview

LC COLORS

This notebook analyses ONB Labs' historic postcards and generates individual color swatches from the images available via IIIF. It's based on a Jupyter Notebook by Laura Wrubel colour clustering images of the Library of Congress and hence a good example that openly licensed code helps your peers!

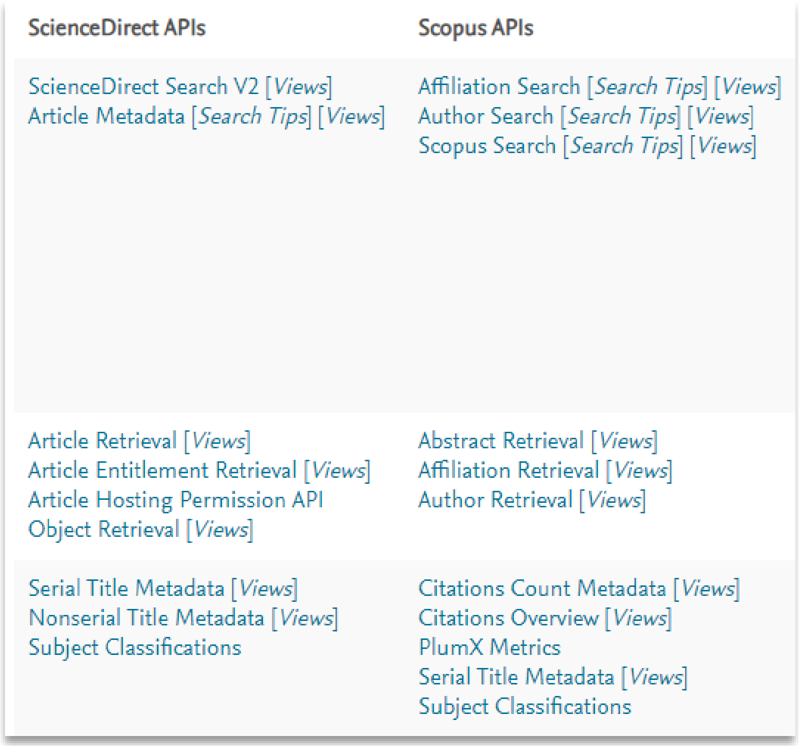
Caveats von Datenzugängen

API-Zugang XY ≠ API-Zugang XY

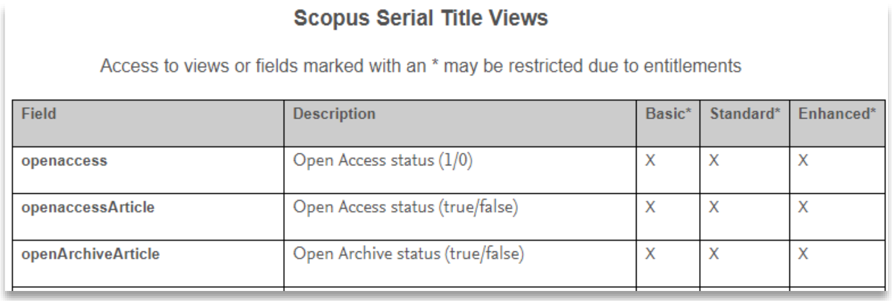
- versch. Endpoints, Versionen, Output-Formate
- Einschränkungen durch Extra-Lizenzen, z.B. WoS, Scopus „Views“)

Bedingungen, Design/Usability, Inhalte von APIs

- sehr divers
- im steten Wandel



ScienceDirect APIs	Scopus APIs
ScienceDirect Search V2 [*Views*] Article Metadata [*Search Tips*] [*Views*]	Affiliation Search [*Search Tips*] [*Views*] Author Search [*Search Tips*] [*Views*] Scopus Search [*Search Tips*] [*Views*]
Article Retrieval [*Views*] Article Entitlement Retrieval [*Views*] Article Hosting Permission API Object Retrieval [*Views*]	Abstract Retrieval [*Views*] Affiliation Retrieval [*Views*] Author Retrieval [*Views*]
Serial Title Metadata [*Views*] Nonserial Title Metadata [*Views*] Subject Classifications	Citations Count Metadata [*Views*] Citations Overview [*Views*] PlumX Metrics Serial Title Metadata [*Views*] Subject Classifications



Scopus Serial Title Views

Access to views or fields marked with an * may be restricted due to entitlements

Field	Description	Basic*	Standard*	Enhanced*
openaccess	Open Access status (1/0)	X	X	X
openaccessArticle	Open Access status (true/false)	X	X	X
openArchiveArticle	Open Archive status (true/false)	X	X	X

Rechtliches zu Datenzugängen

Nutzung

Für wissenschaftliche Zwecke sind die mit TDM-Methoden verbundenen **Vervielfältigungen und Speicherungen** von rechtmässig zugänglichen Inhalten durch das Schweizerische Urheberrechtsgesetz erlaubt.

ABER: Lizenzbedingungen greifen.

Zugang

Die Ressourcen und ihre Zugänge unterliegen verschiedenen **rechtlichen und technischen Nutzungsbedingungen**. **Konsultieren Sie diese** vor einem automatisierten Zugriff.

Insbesondere für hier nicht aufgeführte lizenzierte Inhalte ist ein automatisierter Zugriff oft ausgeschlossen und **kann zur Sperrung des Zugriffs** auf die Datenbank durch den Anbieter führen.

Kontaktieren Sie uns, wenn Sie unsicher sind, ob ein Zugriff rechtmässig ist.

Do & Don‘ts für den TDM-Zugang

Erste Regel: Daten-APIs oder **Daten-Dumps des Anbieters benutzen**!

→ Scraping von Websites ist aufwändig, fehleranfällig – und oft nicht erlaubt.

→ Haben Sie Zugriff auf API oder Dump auf der Anbieter-Seite, ist das i.d.R. rechtens.

→ Immer Anbieter-Website auf TDM-Bedingungen und Anleitungen checken.

Zweite Regel: Bei der Uni-Bibliothek nachfragen.

→ Wenn keine TDM-Info beim Anbieter zu finden ist oder Zugriff auf API/Dump nicht funktioniert.

→ Berner TDM-Bedingungen aus Lizenzierungen können von den allgemeinen Infos abweichen.

→ Hilfestellung bei der API-Nutzung (z.T. Code vorhanden, auf ds-pytools und intern).

E-Library: esupport.ub@unibe.ch oder Digital Scholarship: ds.ub@unibe.ch

Merci für Ihre Aufmerksamkeit!
Fragen, Anregungen?

Kathi Woitas

kathi.woitas@unibe.ch

Digital Scholarship Services