UNIVERSITÄT BERN

Nexis Data Lab

Coffee Lecture Bibliothek vonRoll

Kathi Woitas, Digital Scholarship Services, UB Bern

03.05.2022 kathi.woitas@unibe.ch oder ds.ub@unibe.ch

DOI: 10.5281/zenodo.6517542

Digital Scholarship Services

Kathi Woitas

UNIVERSITÄT BERN

KW

- Bibliothekswissenschaft, Europäische Ethnologie (M.A., HU Berlin)
- mehrere CAS in Data Science (Datenanalyse, Statistical Modelling, Practical Machine Learning, Big Data)

- Vermittlungsformate zu Data Literacy und Tools
- DS Toolbox mit Jupyter Notebooks, u.a. zu NLP, Nutzung von Daten-APIs
- DS-Webpage mit TDM-Ressourcen
- Datenaquise und –aufbereitung *on demand*
- Lizenzierung von TDM-Plattformen wie *Nexis Data Lab*

Text- and Data Mining (TDM)

Was ist das?

“The goal of data mining is to **discover or derive new information** from data, **finding patterns** across datasets, and/or separating signal from noise.”
(Hearst, 1999)

“If we extrapolate from data mining … on numerical data to **data mining from text collections**, we discover that there already exists a field engaged in text data mining: computational linguistics!”
(Hearst, 1999)

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.

TDM in den Sozialwissenschaften?

UNIVERSITÄT BERN

Using Word Embeddings to Analyze how Universities Conceptualize “Diversity” in their Online Institutional Presence

David Rozado

Society, 2019, 56, 256–266. DOI:10.1007/s12115-019-00362-9

A new approach to semantic sustainability assessment: text mining via network analysis revealing transition patterns in German municipal climate action plans

Manuel W. Bickel

Energy, Sustainability and Society, 2017, 22(7). DOI: 10.1186/s13705-017-0125-0

Classification of Poverty Condition Using Natural Language Processing

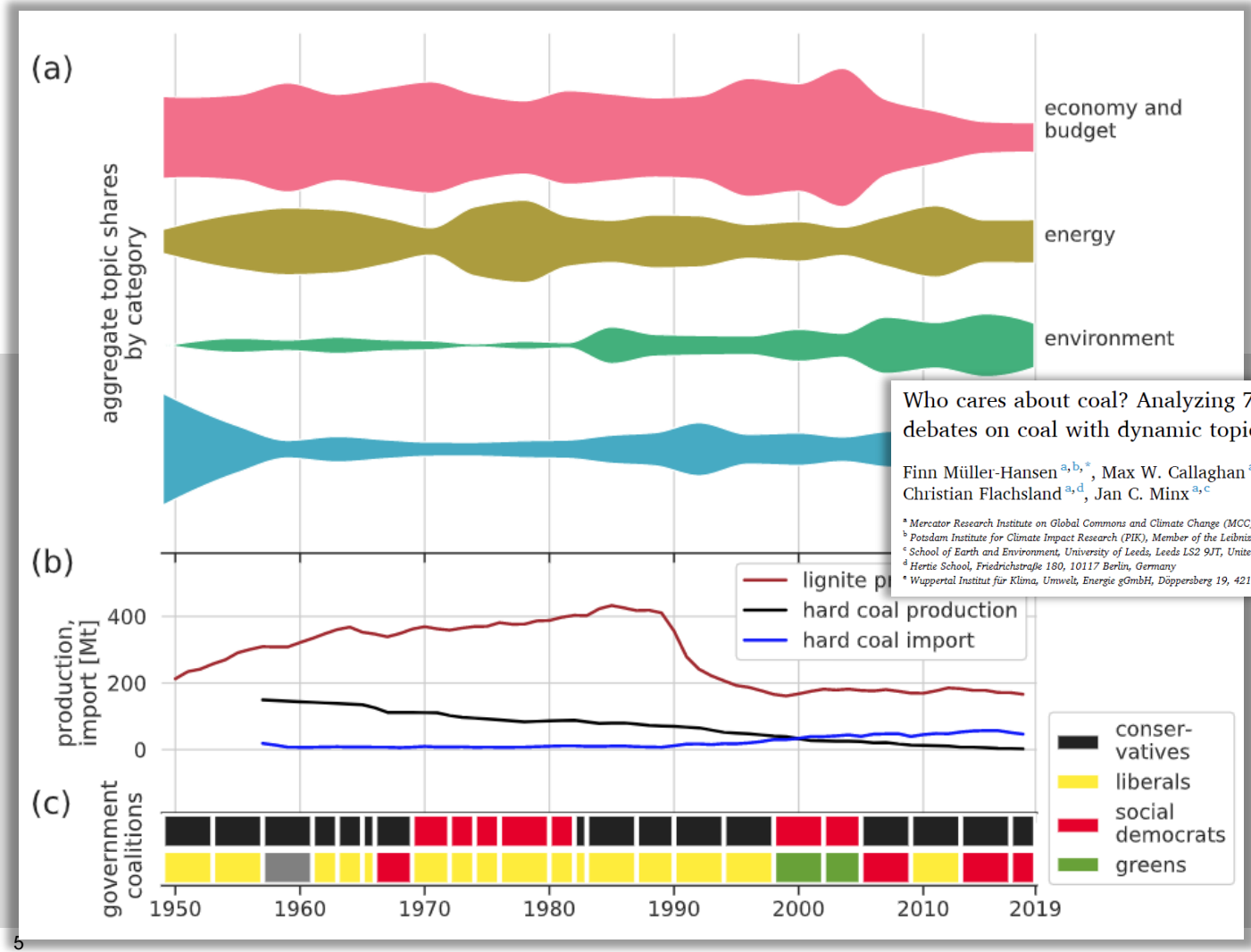
Guberney Muñetón-Santa, Daniel Escobar-Grisales, Felipe Orlando López-Pabón, Paula Andrea Pérez-Toro & Juan Rafael Orozco-Arroyave

Social Indicators Research, 2022. DOI: 10.1007/s11205-022-02883-z

Understanding #WorldEnvironmentDay User Opinions in Twitter: A Topic-Based Sentiment Analysis Approach

by Ana Reyes-Menendez [1], José Ramón Saura [1,*] and Cesar Alvarez-Alonso [2]

International Journal of Environmental Research and Public Health, 2018, 15(11), 2537. DOI:10.3390/ijerph15112537



Who cares about coal? Analyzing 70 years of German parliamentary debates on coal with dynamic topic modeling

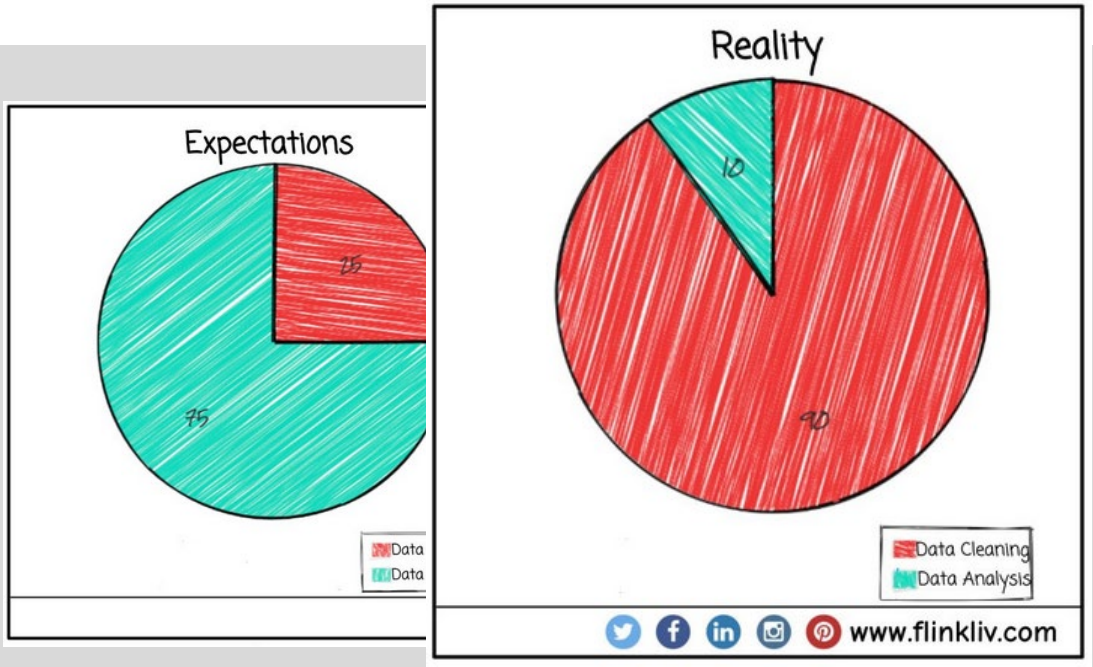
Finn Müller-Hansen [a,b,*], Max W. Callaghan [a,c], Yuan Ting Lee [a,d], Anna Leipprand [e], Christian Flachsland [a,d], Jan C. Minx [a,c]

[a] Mercator Research Institute on Global Commons and Climate Change (MCC), EUREF Campus 19, Torgauer Straße 12-15, 10829 Berlin, Germany
[b] Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, P.O. Box 60 12 03, D-14412 Potsdam, Germany
[c] School of Earth and Environment, University of Leeds, Leeds LS2 9JT, United Kingdom
[d] Hertie School, Friedrichstraße 180, 10117 Berlin, Germany
[e] Wuppertal Institut für Klima, Umwelt, Energie gGmbH, Döppersberg 19, 42103 Wuppertal, Germany

Energy Research & Social Science (2021), 72, 101869. DOI: 10.1016/j.erss.2020.101869

TDM Workflow

1 Datenzugang
2 Pre-Processing
3 Statistische Analyse
4 Vektorisierung
5 Training/Modellentwicklung
6 Auswertung/Interpretation



Vorteile von TDM-Plattformen

UNIVERSITÄT BERN

Datenzugang

- einfacher Datenzugang mit gesicherter Datenqualität
- komfortable Suche + individuelle Korpus-Erstellung
- geklärte Rechte
- oft Inhalte, die sonst nicht frei zugänglich sind

Pre-Processing usw.

- Bereinigung z.T. schon während Suche/Korpus-Erstellung möglich
- oft Online-Analyse-Umgebung, (z.B. JupyterHub)
- z.T. vorgefertigte Analyse-“Notebooks“
- z.T. vorgefertigte Datenpakete mit spezifischen Features

TDM-Plattformen

UNIVERSITÄT BERN

Aktuelles Angebot

- *HathiTrust Research Center* – digitalisierte Publ. aus wiss. US-Bibliotheken
- *Swissdox@LiRI* – Schweizerische Berichterstattung/Medieninhalte
- Nexis Data Lab – internationale Berichterstattung/Medieninhalte

Nexis Data Lab

- News: 20k Quellen, >100 Länder
- Downloadbar:
  - Daten (< 10MB je File)
  - Analyseergebnisse
  - Analyse-Code (Notebooks)
- Uploadbar: eigener Code
- Single-User-Account → auf Zeit, bei Interesse: ds.ub@unibe.ch

Nexis Data Lab

UNIVERSITÄT BERN

Einstieg

Suchoberfläche zur Erstellung von Korpora

→ Datenpaketierung

→ max. 6 *Workspaces* (je bis 100k Dokumente)

→ Laden eines Datensets in eigentliches Data Lab: JupyterHub (Python, R kommt)

→ Start im Ordner *WORK*

TDM + Analyse

- Hilfe und Video zum Einstieg
- Analyse-Notebooks für:
  - Datenset-Struktur verstehen und basale Auswertungen machen
  - Daten in Dataframe laden und als CSV/ZIP downloaden
  - Einstieg in das Text Processing + Topic Modelling

Live Demo: Nexis Data Lab

Vielen Dank für Ihre Aufmerksamkeit!

Fragen? Anmerkungen? Hinweise?

Kathi Woitas, Digital Scholarship Services, UB Bern

03.05.2022 kathi.woitas@unibe.ch oder ds.ub@unibe.ch