UNIVERSITÄT BERN

Text- und Datamining in den Sozialwissenschaften

Coffee Lecture Bibliothek vonRoll

Kathi Woitas, Digital Scholarship Services, UB Bern

05.04.2022 kathi.woitas@unibe.ch oder ds.ub@unibe.ch

Digital Scholarship Services

Kathi Woitas

UNIVERSITÄT BERN

Ziel DSS

Unterstützung datengetriebener Forschungs- und Lehrmethoden für alle Disziplinen durch:

- Lizenzierung und Erstellung von Datenbeständen und Tools
- Vermittlung und Beratung
- eigene datenbasierte Projekte

KW

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

Digital Scholarship Services

Was darf man erwarten?

Hat es schon

- Vermittlungsformate zu Data Literacy und Tools
- DS Toolbox mit Jupyter Notebooks
- TDM-Webpage
- Lizenzierung von TDM-Plattformen
- Datenaquise und –aufbereitung on demand

Kommt demnächst

- neue, umfangreiche Webpage mit Übersicht zu Services, Tools, Datenquellen etc. ☺
- weitere vR Coffee Lectures zu OpenRefine und Nexis Data Lab
- Jupyter Notebooks zur Nutzung diverser Daten-APIs

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

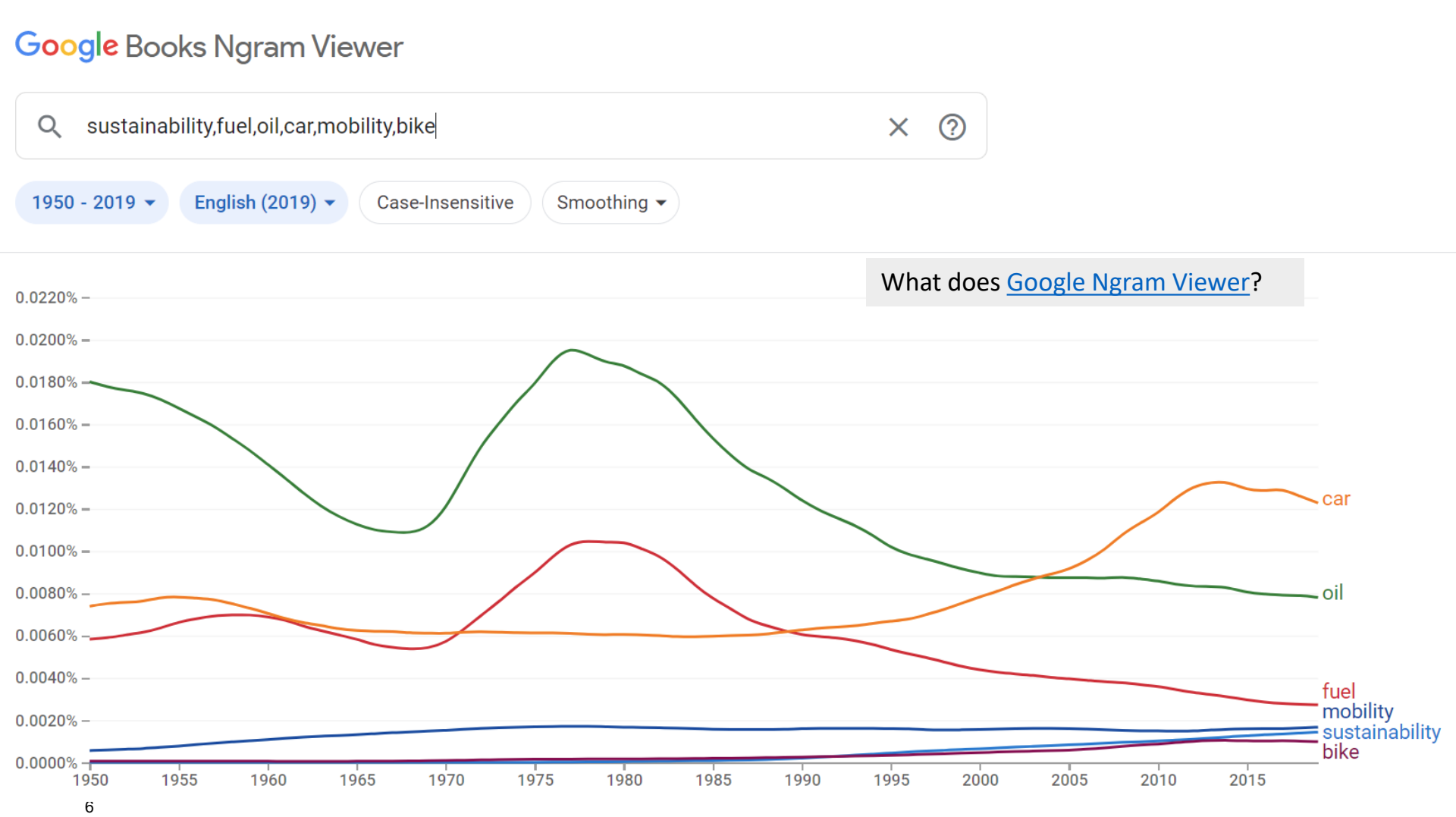
Understanding #WorldEnvironmentDay User Opinions in Twitter: A Topic-Based Sentiment Analysis Approach

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

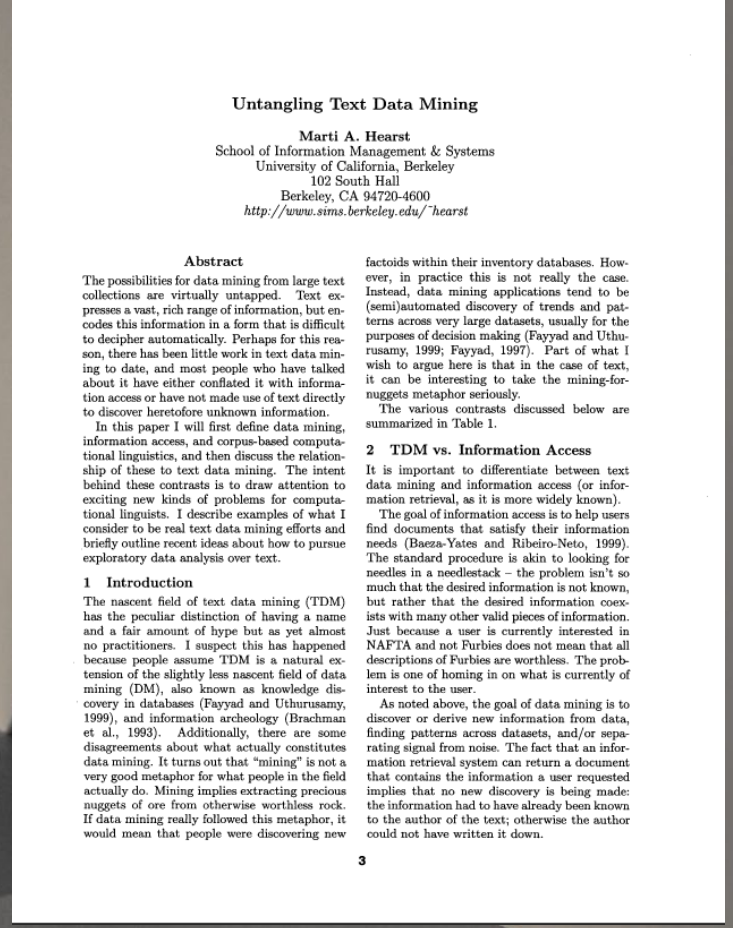
Text- and Data Mining (TDM)

Was ist das?


Google Books Ngram Viewer
sustainability,fuel,oil,car,mobility,bike
1950 - 2019
English (2019)
Case-Insensitive
Smoothing
What does Google Ngram Viewer?
0.0220%
0.0200%
0.0180%
0.0160%
0.0140%
0.0120%
0.0100%
0.0080%
0.0060%
0.0040%
0.0020%
0.0000%
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
car
oil
fuel
mobility
sustainability
bike

DATA MINING

"THE GOAL OF DATA MINING IS TO DISCOVER OR DERIVE NEW INFORMATION FROM DATA, FINDING PATTERNS ACROSS DATASETS, AND/OR SEPARATING SIGNAL FROM NOISE."



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).

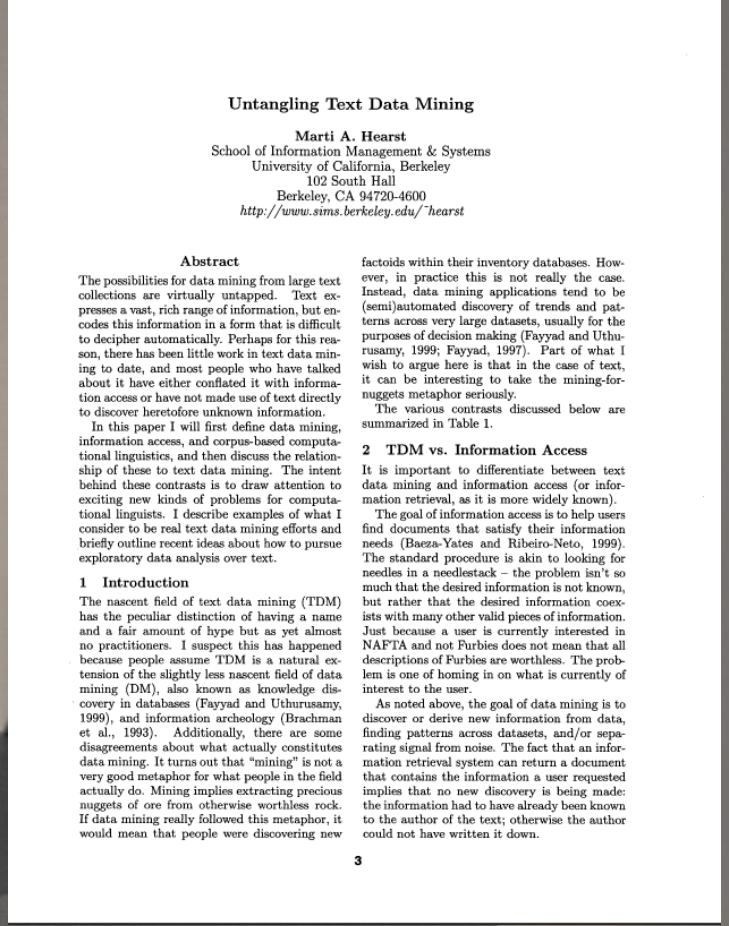
The goal of information access is to help users find documents that satisfy their information needs (Baeza-Yates and Ribeiro-Neto, 1999). The standard procedure is akin to looking for needles in a needlestack – the problem isn't so much that the desired information is not known, but rather that the desired 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.

TEXT MINING

"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!"



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).

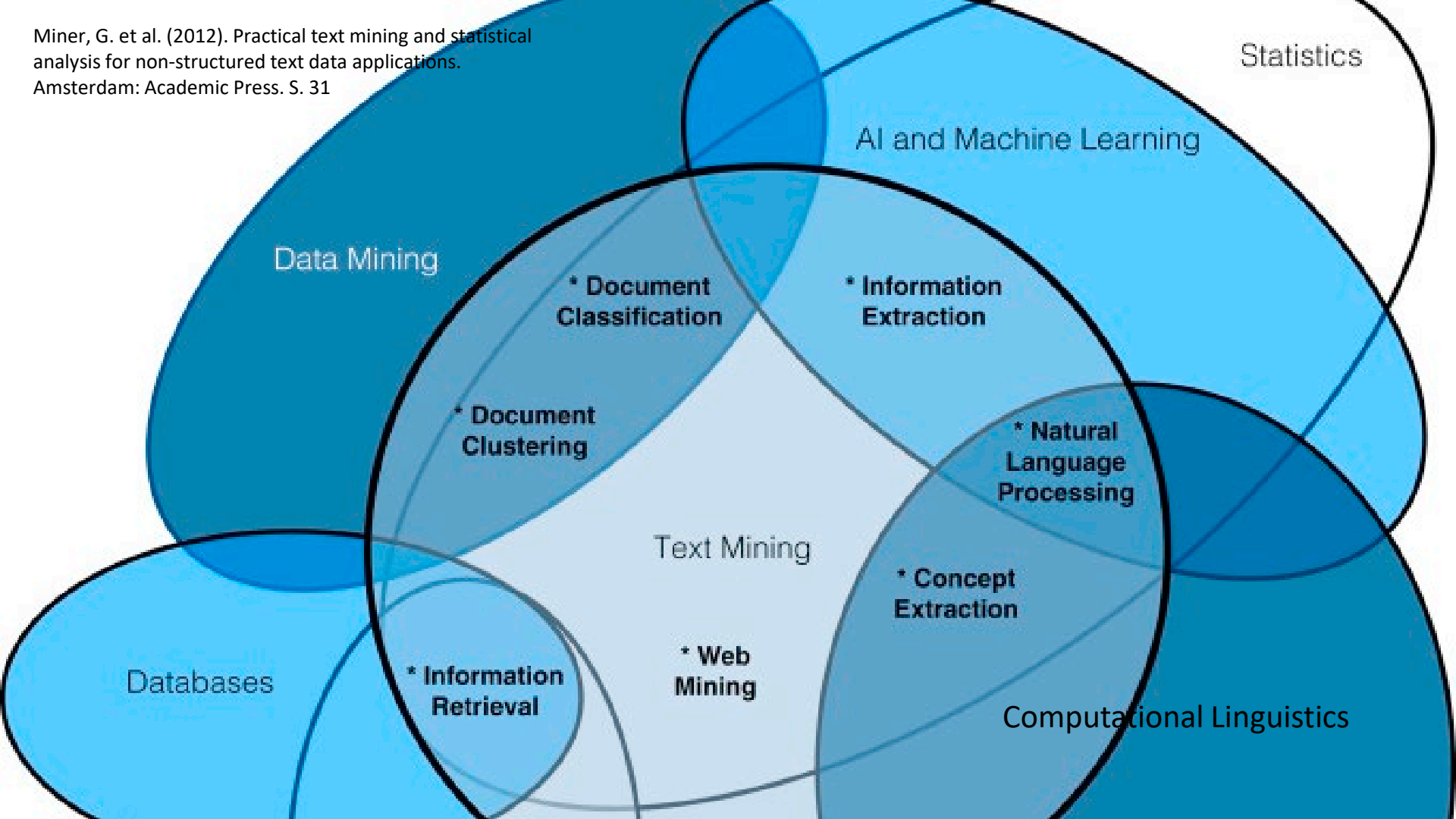
The goal of information access is to help users find documents that satisfy their information needs (Baeza-Yates and Ribeiro-Neto, 1999). The standard procedure is akin to looking for needles in a needlestack – the problem isn't so much that the desired information is not known, but rather that the desired 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.

3

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.

Miner, G. et al. (2012). Practical text mining and statistical analysis for non-structured text data applications. Amsterdam: Academic Press. S. 31



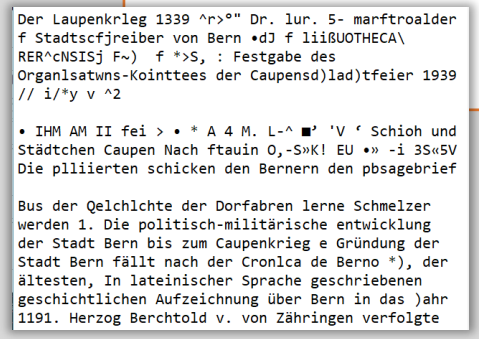
Klassischer linguistischer Ansatz	Statistischer/ML-Ansatz
Erstellung eines regelbasierten Sprachmodells Morphologische, syntaktische… Analyse Nachahmung der menschlichen Sprachverarbeitung Modell enthält Wörter, Grammatik (Syntax) und Bedeutung (Semantik) als Schichten mit zunehmender Komplexität	**Lernen von (hochdimensionalen) Sprachmustern aus einer großen Anzahl von Dokumenten** Sprache = Wiederholung von Mustern Analyse über einen grossen, aber begrenzten Bestand wiederkehrender Muster Auftreten von Mustern löst die wahrscheinlichste Regel/Handlung aus

Datenkontinuum

Unstrukturiert

Text

als Bild, Ton



Semistrukturiert

Markup

(e.g. xml, json)

Graphen

Strukturiert

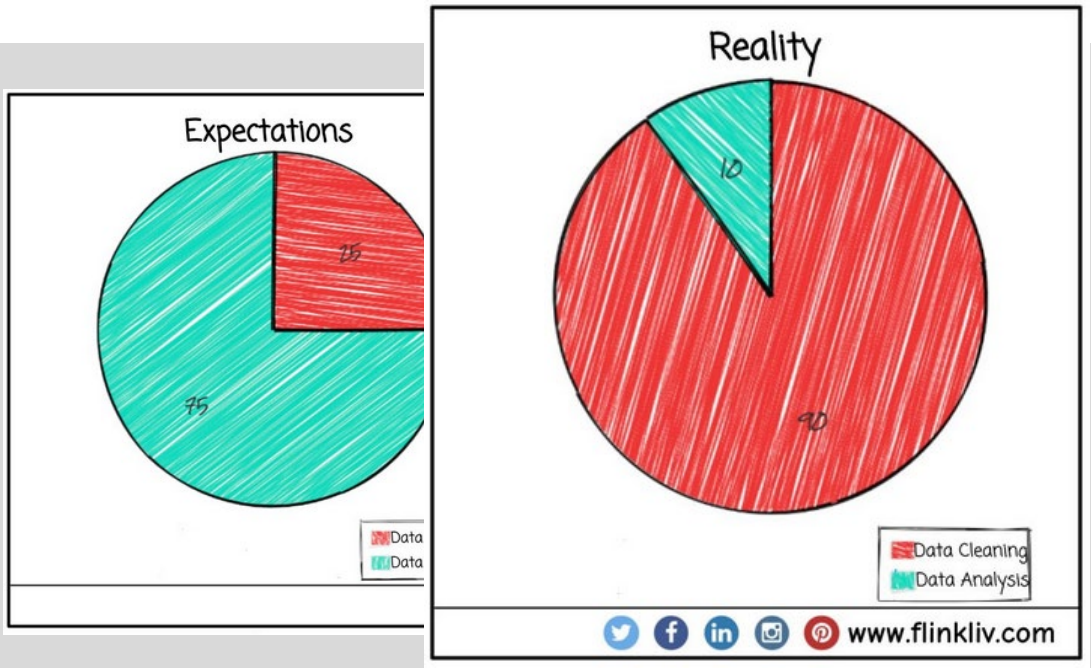
Tabular

Numerisch

E	F	G	J	L
month	day_of_weel	duration	previous	cons.price.id
4	5	1.69019608	0	93.075
8	4	2.10380372	0	92.201
7	5	2.99122608	0	93.918
4	5	2.99913054	1	93.075
7	3	2.76267856	0	93.918
5	3	2.17026172	0	92.893
12	4	2.36735592	0	92.713
3	5	2.25527251	0	92.843
11	3	2.29003461	0	93.2
7	2	2.33041377	0	93.918
10	3	1.98677173	0	92.431
7	1	3.05499586	0	93.918

TDM Workflow

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



1 Datenzugang

Geeignete Ressourcen finden
zumeist Primärquellen

- digitale Sammlungen/Archive, z.B. Zeitungen, Governance-Dokumente, Gebrauchsliteratur
- born digital content, z.B. Social Media
- auf TDM- und/oder Download-Bestimmungen achten!

Rohdaten beziehen

- Wo möglich, Daten-APIs oder Daten-Dumps benutzen
- Scraping von Websites ist aufwändig und fehleranfällig
- Anpassen der Datenstruktur, Zusammenführen von Dateien usw.

2 Pre-Processing

OCR, Speech-to-text und Bereinigung

- Text auslesen oder erkennen, z.B. aus PDFs
- Nicht benötigte Teile entfernen, z.B. Dateikopf mit Metadaten, weiteres Markup, z.B. aus XML

Basis-Textverarbeitung

- *Tokenisierung*: Aufteilung des Textes in Sätze und Wörter
- Kleinschreibung aller Wörter (sprachspezifisch)
- Entfernen von Interpunktion und/oder Stoppwörtern (= häufige Wörter mit geringer Semantik wie „ein“ oder „zu“)

2 Pre-Processing

Der Koch kocht das Menü und der Kellner serviert das Menue.

der, Koch, kocht, das, Menü, und, der, Kellner, serviert, das, Menue

Basis-Textverarbeitung

- Normalisierung von Rechtschreibvarianten, Abkürzungen
- *Stemming* (= alle Wörter auf ihre Stämme kürzen) **oder**
- *Lemmatisierung* (= Rückführung auf gramm. Grundformen)

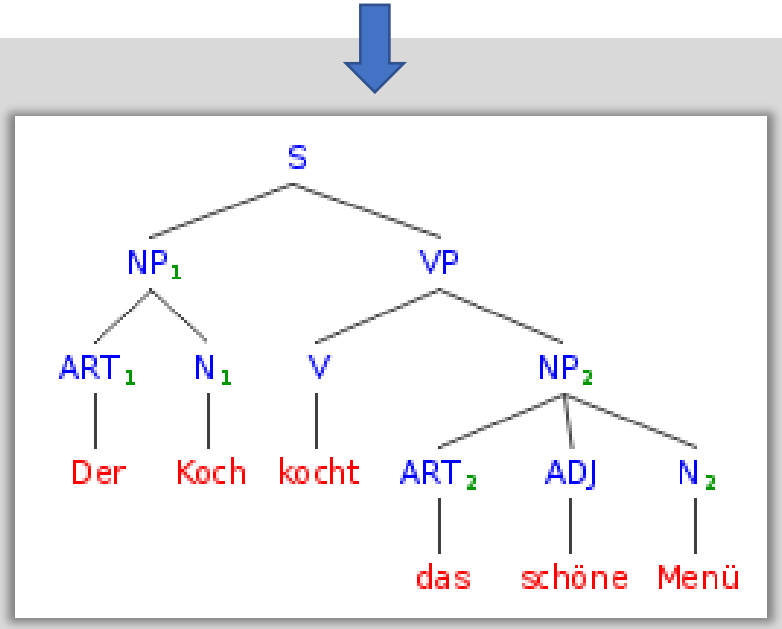
koch	2
menü	2
kelln	1
servier	1

koch	1
kocht	1
menü	2
kellner	1
serviert	1

koch	2
kochen	1
menü	1
kellner	1
servieren	1

2 Pre-Processing

Der Koch kocht das schöne Menü.



Tiefere Textverarbeitung:
Morphologische und syntaktische Analyse

- Ermittlung der Wortarten (POS) von Token
- Ermittlung der syntaktischen Merkmale von Token (Syntaxbaum/parse tree)

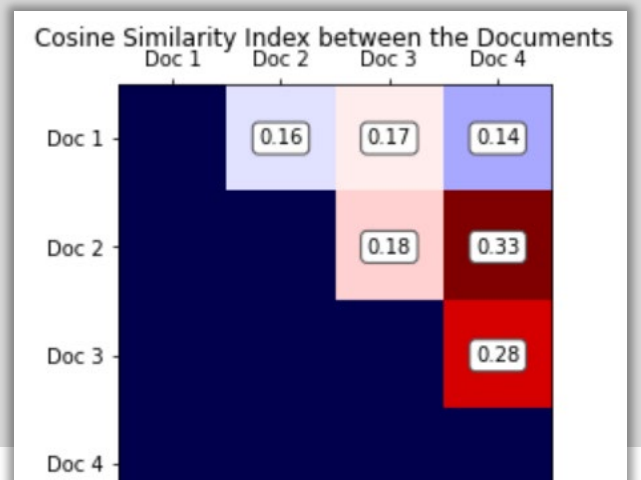
3 Statistische Analyse

- Datenübersicht
- Häufigkeiten von Token oder N-Grammen im Zeitverlauf (Frequenzanalyse)
- gemeinsames Vorkommen von Token (Kookurrenzanalyse)
- Gewichtung von Wörtern, z.B. Vorkommenshäufigkeit - Inverse Dokumenthäufigkeit (TF-IDF)
- Analysen unter Beizug von Metadaten

4 Vektorisierung

- Umwandlung der Token in numerische „Features“ = *Vektorisierung*
- Repräsentation von Texten als Art und Anzahl ihrer Tokens → *word count vectors* dienen zur Anwendung von ML-Algorithmen
- weitere Vektorisierungsarten, z.B. Embeddings (word2vec)

	stadt	schaffhausen	ist	see	liegt	blausee	heute
Doc 1	0	0	0	4	2	4	1
Doc 2	0	0	3	0	1	0	1
Doc 3	2	4	0	0	0	0	1
Doc 4	2	0	1	0	1	0	1



5 Training/Modellentwicklung

UNIVERSITÄT BERN

- Ziel: Entwicklung eines adäquaten Modells
- verschiedene Algorithmen ausprobieren
- *Tuning* des bevorzugten Algorithmus durch iteratives Anpassen von Hyperparameter, Feature Engineering
- stete Evaluation des Tuning-Modelle durch Messung von Güte- und Performance-Kriterien, Vermeidung von Overfitting
- Publikation des endgültigen Modells

5 Training

UNIVERSITÄT BERN

Unüberwachte Lernmethoden

- Ziel: unbekannte Muster in Texten entdecken
- keine Ground Truth („Lösungsdaten“) zum Trainieren notwendig

Anwendungsfälle

- Clustering von Dokumenten: Gruppen von ähnlichen Texten finden
- *Topic Modeling*: Suche nach Gruppen ähnlicher Dokumente und Ermittlung ihrer gemeinsamen Themen

5 Training

UNIVERSITÄT BERN

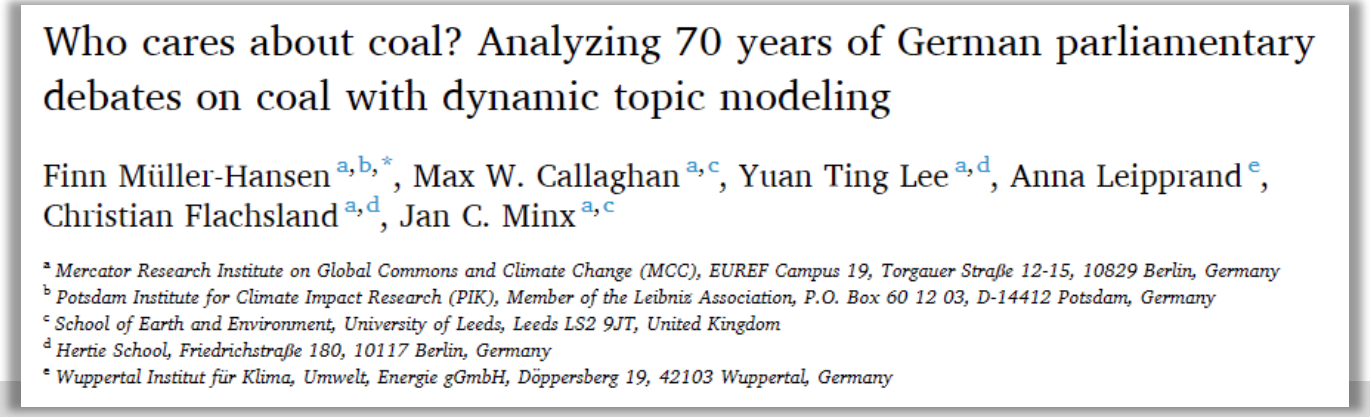
Methoden des überwachten Lernens

- Ziel: Den mathematischen Zusammenhang zwischen bestimmten Datenwerten finden.
- benötigt Ground Truth („Lösungsdaten“) zum Trainieren
- wird für Vorhersage benutzt

Anwendungsfälle

- *Sentiment Analysis*: Erkennen der Stimmung eines Textes
- *Text-Klassifizierung*: Texte in bekannte Gruppen einordnen
- *Named Entity Recognition*: Erkennen von benannten Entitäten wie Orten, Personen, Organisationen usw.

Beispiel



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

Müller-Hansen, F. et al. (2021): Who cares about coal? Analyzing 70 years of German parliamentary debates on coal with dynamic topic modeling. In: Energy Research & Social Science, 72. Jg., S. 101869.

Datenzugang

- 870k Bundestagsreden 1949-2019 + Daten zu Politikern (Rednern) via Open Data Service API des Dt. Bundestags

Pre-Processing

- Parsing der PDF- und XML-Rohdaten + Überführung in relationale DB
- Suche: *9167 Reden*, die “Kohle” erwähnen = Dokumente
- Tokenisierung, Stemming, Eliminierung von Stopwords + allgemeinen Wörtern (Python) → Vokabular = 20’000 häufigste verbliebene Token

Beispiel

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

Vektorisierung:

- Berechnung der Word-Count-Vektoren für Reden (Python)

Analyse/Modellentwicklung: *Dynamic Topic Modeling*

- Basis: Annahme, dass wiederholt gemeinsam auftretende Wörter semantische Nähe implizieren
- *Topics* = verschiedene Kombinationen von gemeinsam auftretenden Wörtern auf Grundlage ihrer Häufigkeit in Dokumenten (Verteilungen)
- Reden erhalten scores für die einzelnen topics → *document-topic scores*
- Topics erhalten scores für die verschiedenen Wörter → *word scores*

Beispiel

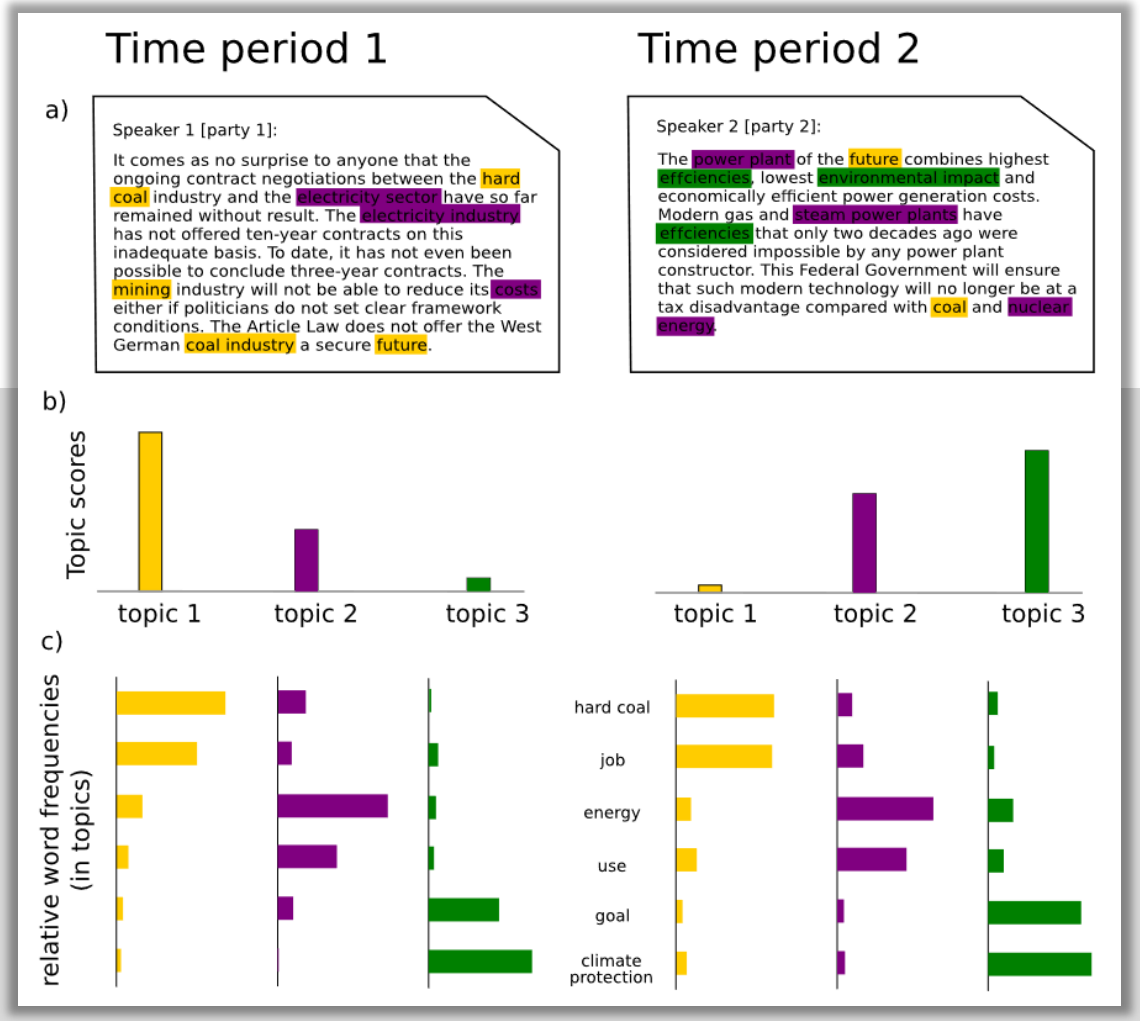
UNIVERSITÄT BERN

List of topics grouped by category with labels, top words and topic scores (in percent of total score).

Category	Label	Weighted top words	Topic share
economy and budget	economic policy	Wirtschaft, unternehm, entwickl, deutsch, markt, deutschland, wirtschaftspolit, staat, stark, marktwirtschaft	5.49
	budget	milliard, million, bund, haushalt, hoh, rund, ausgab, bundeshaushalt, bundesregier, offent	3.78
	budget 2	haushalt, milliard, bundesfinanzminist, investition, regier, finanzpolit, finanzminist, schuld, offent, hoh	2.87
	job market	arbeit, arbeitslos, arbeitsplatz, sozial, arbeitnehm, arbeitsmarkt, unternehm, beschaft, zahl, wirtschaft	2.83
	fiscal reform	spd, prozent, deutschland, euro, koalition, steu, haushalt, milliard, hoh, reform	2.75
	economic policy 2	wachstum, wirtschaft, bundesregier, wirtschaftspolit, offent, stabilitat, seit, bundeswirtschaftsminist, konjunkturpolit, aufschwung	2.17
	housing & social security	sozial, gesetz, wohnung, wohnungsbau, rent, hoh, fall, arbeit, mittel, alt	2.09
	economic policy & participation	gesetz, unternehm, gesetzentwurf, entwurf, betrieb, mitbestimm, offent, arbeitnehm, wirtschaft, gewerkschaft	2.07
	subsidy reduction	euro, subvention, milliard, prozent, mittelstand, deutschland, geld, wirtschaft, handwerk, unternehm	1.85
	tax policy	steu, belast, kommun, erhoh, hoh, gemeind, entlast, gesetz, bundesrat, vorschlag	1.78
	research & development	forschung, million, bereich, haushalt, mittel, forder, programm, technologi, entwickl, wissenschaft	1.77
energy	energy supply mix	energi, erneuerbar, energiepolit, kohl, bundesregier, energietrag, energieversorg, kernenergi, nutzung, fossil	3.42

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

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



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

Beispiel

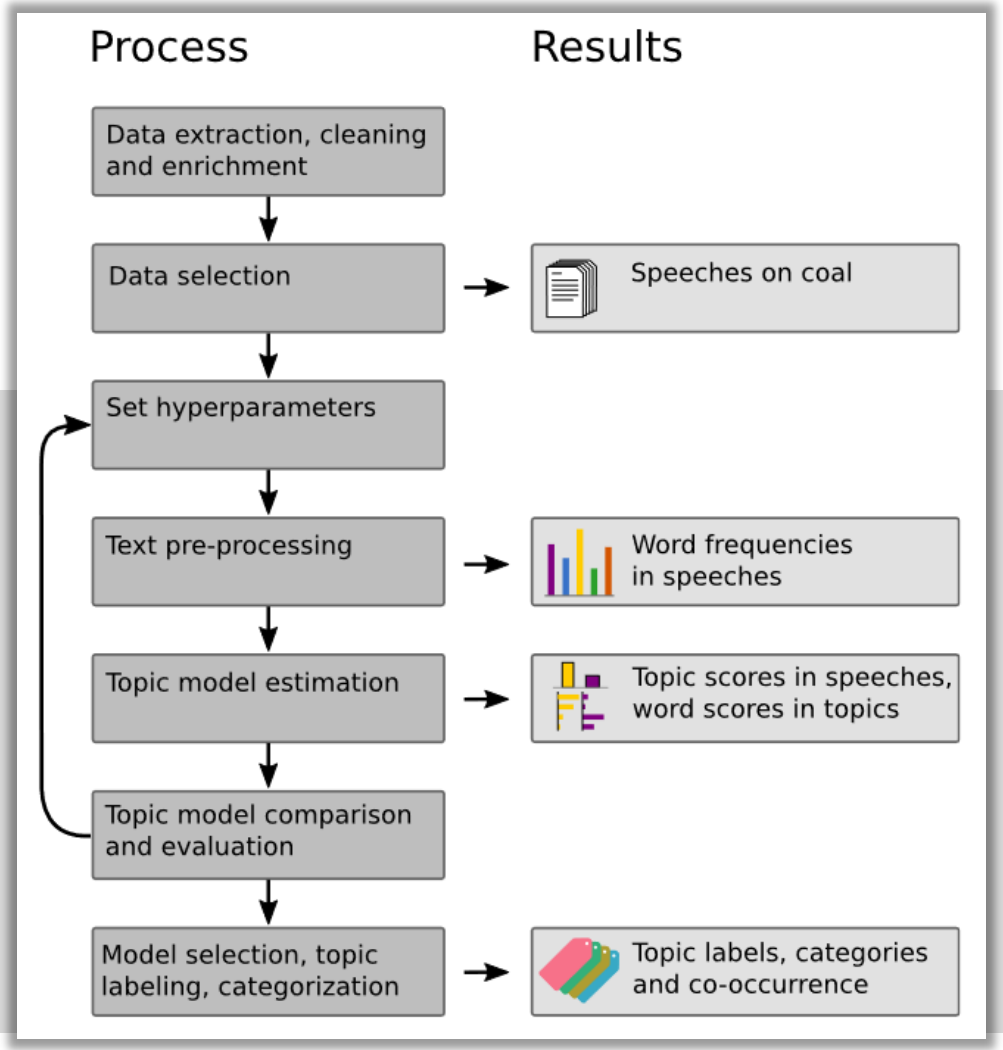
Who cares about coal? Analyzing 70 years of German parliamentary debates on coal with dynamic topic modeling

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

Analyse/Modellentwicklung: *Dynamic Topic Modeling*

- *Dynamic* TM erlaubt Änderungen der word scores für die topics über einzelne Zeitabschnitte
- Model Tuning & Evaluation: Iterationen des TM-Algorithmus mit verschiedenen Hyperparametern, intellektuelle Begutachtung + Gütemasse der entstehenden Topics



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

Beispiel

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]

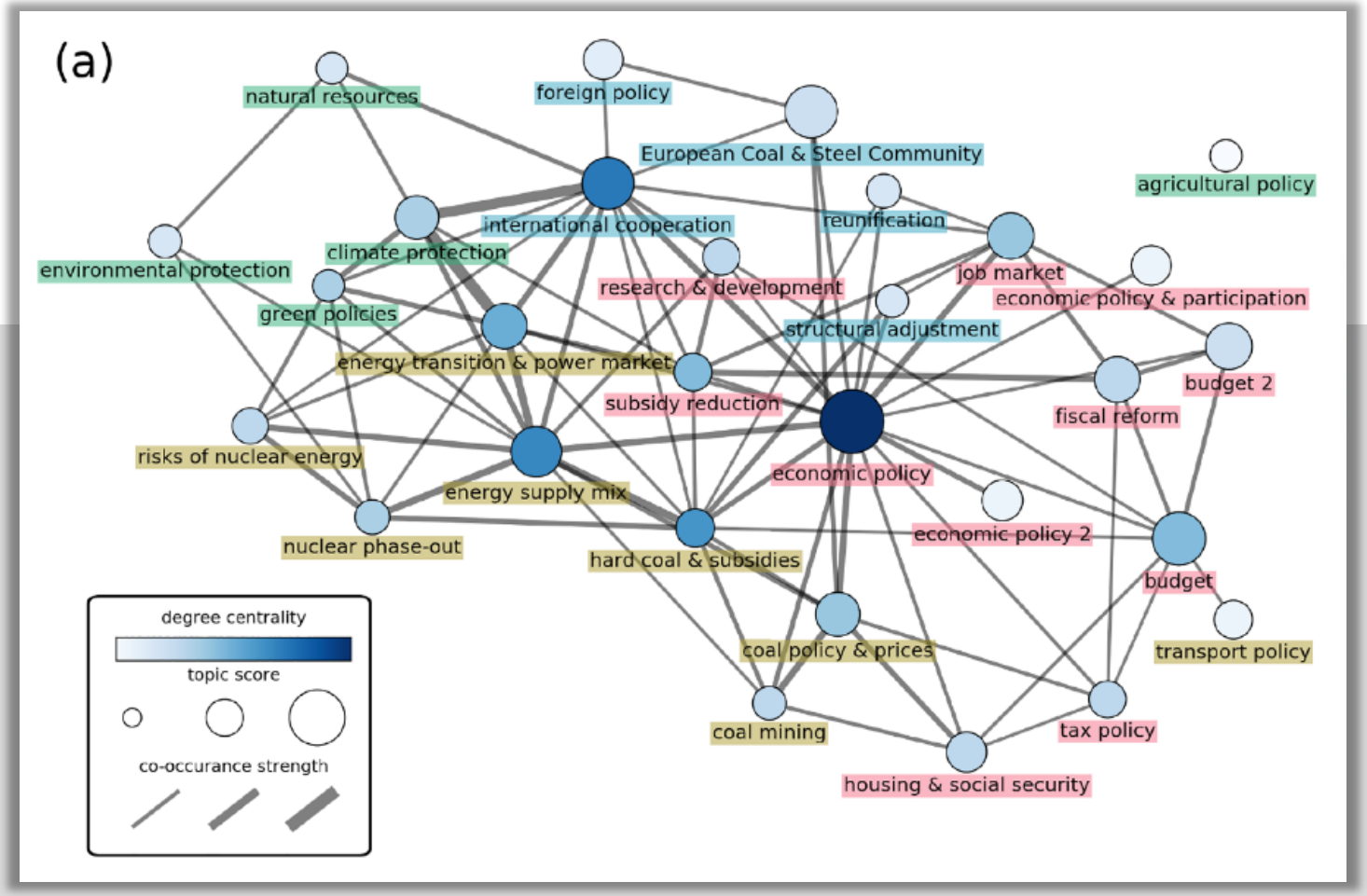
[a] Mercator Research Institute on Global Commons and Climate Change (MCC), EUREF Campus 19, Torgauer Straße 12-15, 10829 Berlin, Germany
[b] Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, P.O. Box 60 12 03, D-14412 Potsdam, Germany
[c] School of Earth and Environment, 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

Auswertung/Interpretation:

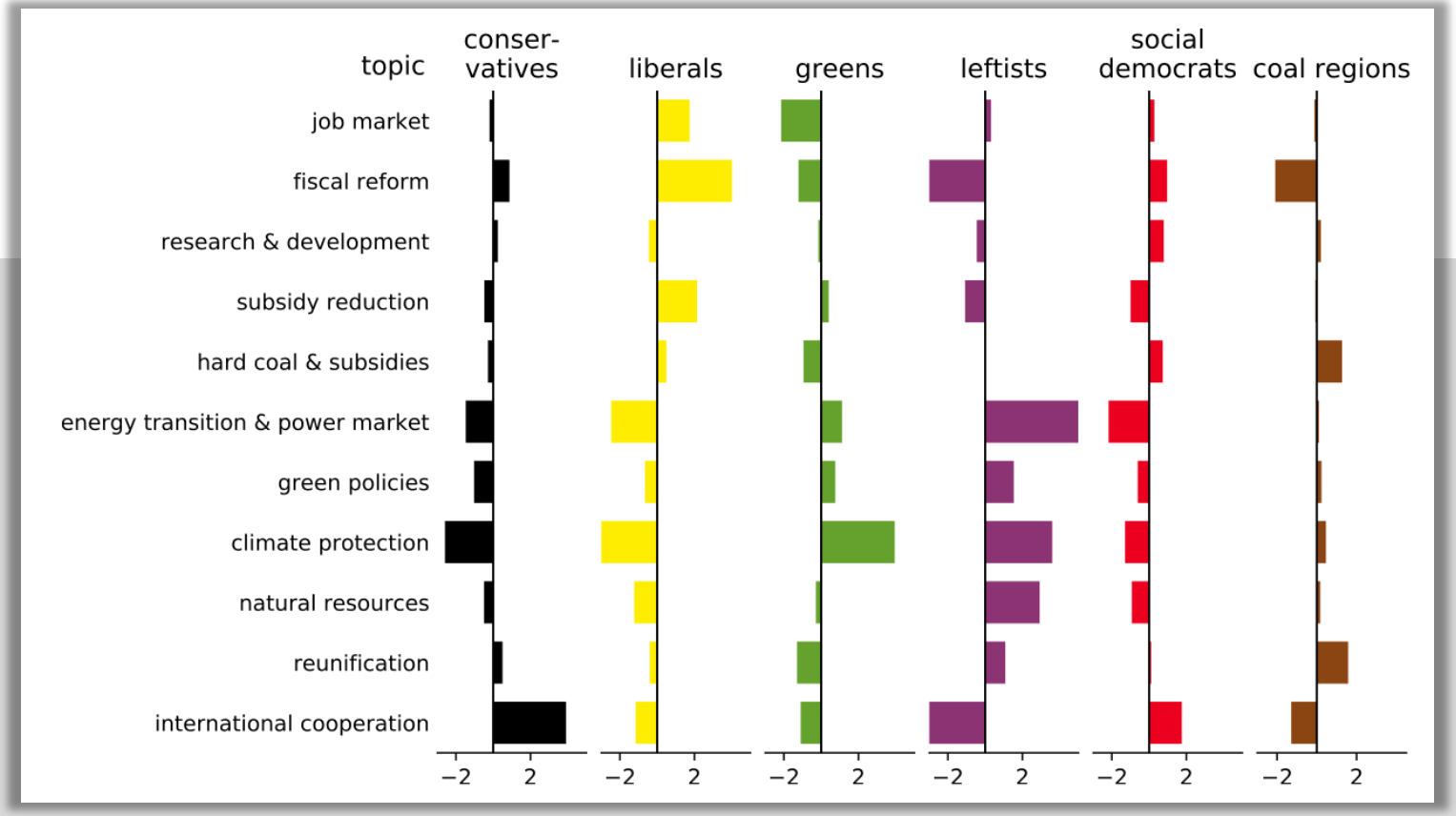
- Interpretation der topics anhand der Reden mit den höchsten scores
- Auswertung nach Regionen und Parteien der Redner
- Topic-Kookurrenz-Analyse zur Untersuchung der Framings

Dokumentation:

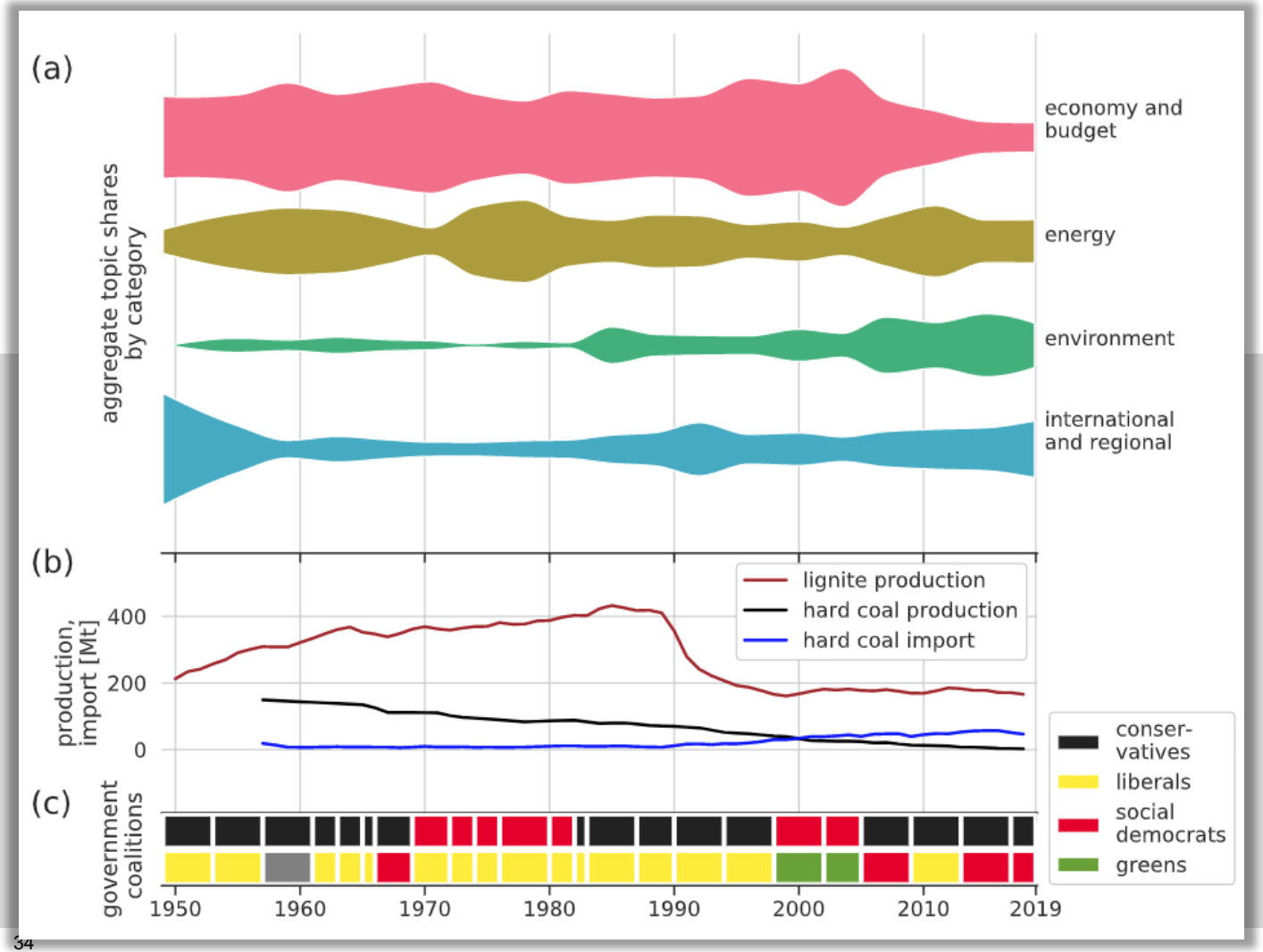
- GitHub Repository mit Code (PDF- und XML-Parser, Analyse-Notebooks) und Daten



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



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



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

TDM: Pros & Cons

UNIVERSITÄT BERN

Vorteile

- Ansatz für grosse Textmengen (oder: Bild, Ton…), die manuelle Kodierung übersteigen
- keine strenge Auswahl, Stichproben nötig
- exploratives Vorgehen möglich
- Basis für qualitative Methoden
- breit gefächerte Ansätze, in steter Entwicklung

Nachteile

- Einarbeitung in NLP-Methoden und ML-Techniken
- benötigt grossen Datenkorpus
- z.T. manuelle Erstellung von Trainingsdaten erforderlich
- lässt Kontextinformationen der Texte ausser Acht
- Interpretationen/Aussagekraft noch in Verhandlung?

Tools

- zum Ausprobieren von TDM braucht es wenig…
- fertige Software (z. B. Voyant, Mallet, Gate, Module in SPSS, SAS)
- TDM-Plattformen von Content-Anbietern: HTRC, *Nexis Data Lab*
- eigener Code (R, Python): Freiheit, Reproduzierbarkeit (online ohne lokale Installation: noto.epfl.ch, Google Colab, zur «Publikation»: notebooks.gesis.org/binder)
- (für big data: High Performance Cluster UBELIX der UNIBE)

Hinweise & Tipps

- Sie möchten lizenzierte Inhalte scrapen/downloaden? Wenden Sie sich an uns! (Es kann zu uniweiten Sperrungen kommen.)
- Sie benötigen spezifische Daten, die lizenziert oder noch digitalisiert werden müssen? Bitte melden Sie uns den Bedarf!
- *Transferable Skills Program:* Einstiegskurse in Datenanalyse mit R und Python

- Schauen Sie auf die neue DSS-Webpage im Mai ☺

Tutorials

- GESIS Materialien und Vortragsreihe zu *Computational Social Science and Digital Behavioral Data*
- Constellate Tutorials & Jupyter Notebooks for Python and TDM
- Eduardo Muñoz NLP Blog: Getting started with NLP
- Programming Historian Tutorials
- Digital Toolbox der UB Bern: einfache Jupyter Notebook Tutorials
- …eine Websuche lohnt sich immer…

Literaturempfehlungen

UNIVERSITÄT BERN

Einstieg/Überblick:

Manderscheid, K. (2019): Text Mining. In: Baur, N./Blasius, J. (Hrsg.): Handbuch Methoden der empirischen Sozialforschung. Wiesbaden: Springer VS. S. 1103–1116. DOI: 10.1007/978-3-658-21308-4_79 (*online zugänglich*)

Ignatow, G./Mihalcea, R. (2017): Text mining: A guidebook for the social sciences. Thousand Oaks: Sage Publications. (*UB vRoll: VRF_MR_2800 72*)

Umfassend:

Anandarajan, M. et al. (2019): Practical text analytics: maximizing the value of text data. Cham: Springer International Publishing. (*online zugänglich*)

Grimmer, J. et al. (2022): Text as data: A guide for using computational text analysis to learn about the social world. New Jersey. (*wird erworben, Verlagsseite*)

Foster, I. et al. (2020): Big data and social science: Data science methods and tools for research and practice. Boca Raton: Chapman and Hall/CRC. (*wird erworben, Verlagsseite*)

Zu Paradigmen von CSS und Data Science:

Hofman, J. M. et al. (2021): Integrating explanation and prediction in computational social science. In: Nature, 595. Jg., H. 7866, S. 181–188. DOI: 10.1038/s41586-021-03659-0 (*online zugänglich*)

Beipielstudie:

Müller-Hansen, F. et al. (2021): Who cares about coal? Analyzing 70 years of German parliamentary debates on coal with dynamic topic modeling. In: Energy Research & Social Science, 72. Jg., 101869. DOI: 10.1016/j.erss.2020.101869