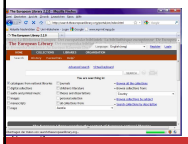
Prof. Dr. Thomas Mandl
Informationswissenschaft
Universität Hildesheim
mandl@uni-hildesheim.de

Was bedeuten Text- und Data-Mining für die Informationsethik?



Berlin, 10. Okt. 2013

Text Mining

Trends

- "Discover useful and previously unknown "gems" of information in large text collections" (Avaquest 2002)
- „Text Mining ist die Extraktion von Wissen aus vielen Texten. .. Das neu gewonnene Wissen kann nicht aus einem einzelnen Text abgelesen werden, sondern nur aus der Gesamtschau auf sehr viele Texten … „ (Mandl 2013 @ KSS)
- „Process of deriving high quality information from text" (Choi 2013 @ LWA)
- *Bergbau-Metapher irreführend*

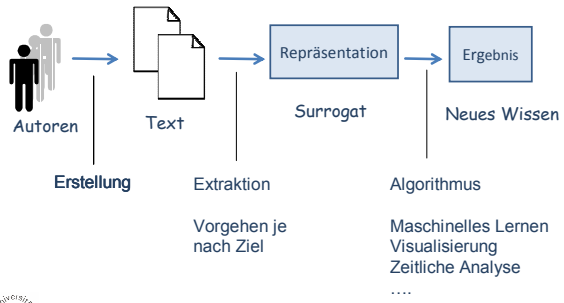
Muster

Assoziationen

Texte als Gegenstand des Text Mining

- Anfragen in Suchmaschinen
- Kurznachrichten (Whatsapp, Twitter, etc.)
- Posts in Sozialen Netzwerken
- Blog-Einträge
- Patientenakten
- Studentische Arbeiten
- Web-Seiten
- Zeitungsartikel
- Patente
- Wissenschaftliche Publikationen
- …

Text Mining



Analyse der Texte

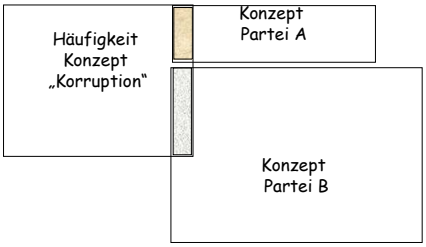
- Häufigkeiten von Begriffen
- Extraktion von Eigennamen etc.
- Linguistische Analyse
- …

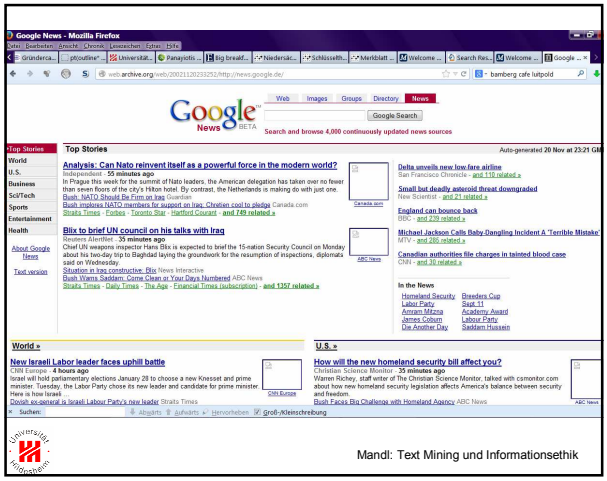
Assoziationsrelationen

Dokument-Term-Matrix	Haus	Bank	Geld	Park
Dok A	0,1	0,6	0,4	0
Dok B	0,9	0,6	0,4	0
Dok C	0	0,6	0	0,6
Dok D	0	0,6	0	0,7

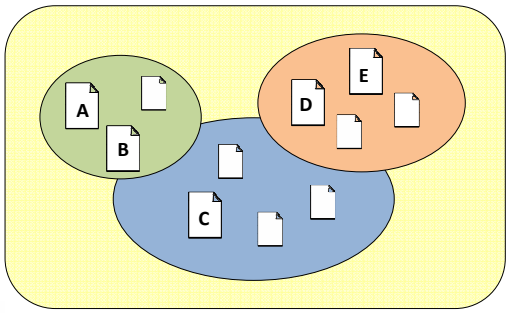
Beispiel Partei und Korruption

- Eine Partei kommt in Texten doppelt so oft vor wie eine andere Partei …





Clustering

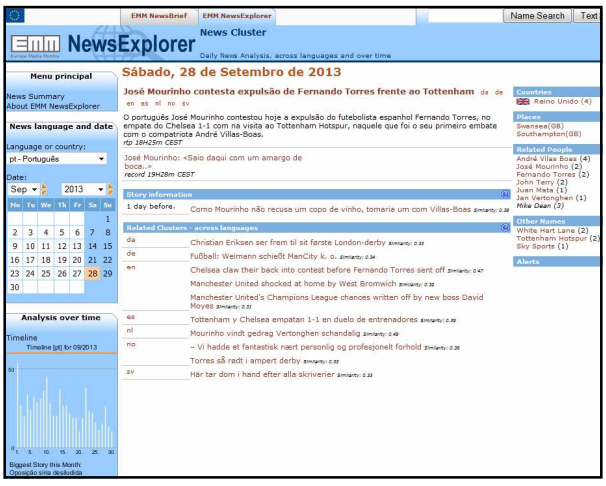


Maschinelles Lernen: Klassifikation

	Eigenschaften	Klasse
Bekannte Items	0 1 2 4	A
	3 5 6 8	B
	0 1 2 4	A
	2 5 7 9	D
	3 6 7 8	C
Neue Items	2 3 5 6	?
	0 1 2 4	?
	3 6 7 9	?

Extraktion eines Modells

Anwendung des Modells



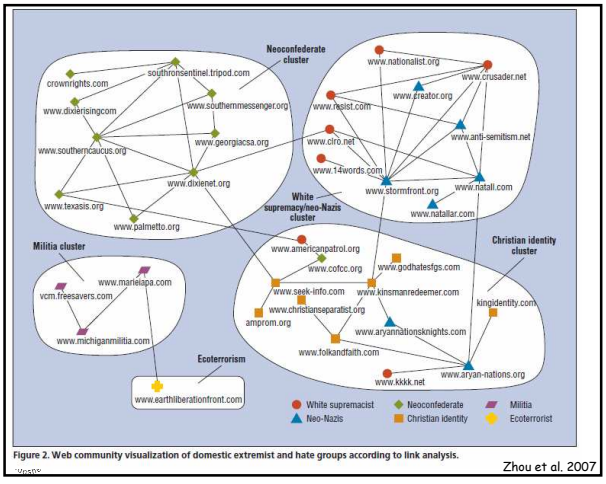


Figure 2. Web community visualization of domestic extremist and hate groups according to link analysis.

Zhou et al. 2007

Suchmaschinen

- Häufig Grundlage für Text Mining oder Werkzeug im Rahmen des Mining Prozesses
- Logs von Anfragen als Gegenstand des Text Mining

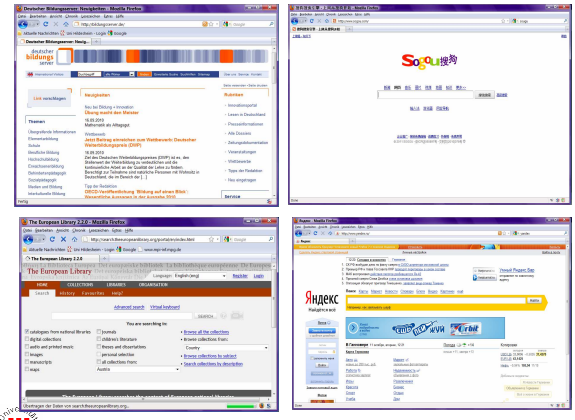
Ethische Probleme Suchmaschinen

- Transparenz (Hinman 2005)
- Objektivität (Blanke 2005)
- Umgang mit Nutzerdaten (Tavani 2005)
- ...

Benutzerforschung

- LogCLEF
  - Track im Rahmen des Cross Language Evaluation Forums (2009 – 2011)
  - Fokus: Nutzerverhalten und mehrsprachige Suche in Digitalen Bibliotheken wie *The European Library* (TEL) und Web-Suchmaschinen

Mandl et al. @ ECIR 2011
Mandl et al. @ CLEF 2010



LogCLEF

Log Ressourcen bei LogCLEF

Year	Origin	Size	Type
2007	MSN	800.000 queries	Query log
2009	Tumba!	350.000 queries	Query log
2009	TEL	1.870.000 records	Query and activity log
2010	TEL	2.600.000 records	Query and activity log
2010	TEL	1.5 GB (zipped)	Web server log
2010	DIPF.de	5 GB	Web server log
2011	Sogou	19 GB	Query and Click log
2012	Yandex		Query and Click log

Mandl et al. @ WLQA 2010

Queries in LogCLEF

Beispiele aus TEL	BOF	IP	Sprache?
livro do infante dom pedro de portugal	fr	es	pt
faust goethe still image	de	hu	en
cosmographia claudii ptolomaei alexandrini	de	en	it
angus thongs and full frontal	de	bg	en
preparing teachers for styles and strategiesbased instructions	de	es	en
ware eigentliche ordnungh und	fr	el	de
verarbeitung der metalle	fr	uk	de
catone in utica piccinni	de	nl	it

Formale Such-Parameter

Alte Sprachen

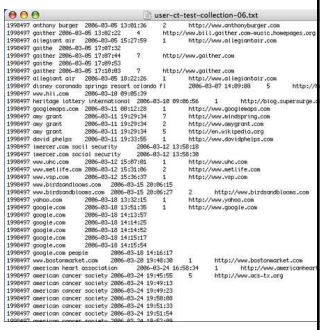
Buchtitel

Tippfehler?

Mehrsprachige Benutzer

Ethische Aspekte

- AOL Skandal (Arrington 2006)
- Anonymisierung (Carpineto & Romano 2013)
- Vernebelungsdienste (Query Obfuscation) Peddinti & Saxena 2010



Smart Metering

- Energiesparen durch das Schaffen von Bewusstsein (Gerwen et al. 2006)
- Häufige Aufzeichnng des Verbrauchs
- Zuordnung zu Tätigkeiten
- Visualisierung
- Aufzeigen von Einsparmöglichkeiten
  - Vergleiche mit anderen (Foster et al. 2010, Petkov et al. 2012)
  - Transformational Systems (Kehr et al. 2011, Kappel et al. 2011)

Ethische Dimension

- Bei Dumb Metering wenig Rückschlüsse möglich
- Smart Metering bedeutet die genaue Beobachtung und Aufzeichnung des Energieverbrauchs
  - Meist auf Server des Energie-Anbieters
- Rückschlüsse auf Verhalten und Charakteristika eines Haushalts möglich (Molina-Markham 2010)
- Z.B. Beckel et al. 2012:
  - eine Woche Energieverbrauch
  - -> Anzahl Personen, Kinder, Alter, Größe Wohnung
  - 80% Genauigkeit
- Verknüpfung mit Wasserverbrauch etc. denkbar

Soziale Netze als Gegenstand

- Persönlichkeit aus Text-Beiträgen und Verhalten in sozialen Netzwerken klassifizieren
  - Hughes et al. 2012, Muscanell & Guadagno 2012, Quercia 2012, Staiano 2012, Moore & McElroy 2012, Bachrach et al. 2012 …

- "the number of friends an individual declares on a web-based social networking service reliably predicted grey matter density in the right superior temporal sulcus, left middle temporal gyrus and entorhinal cortex"
  Kanai 2011: Online social network size is reflected in human brain structure

- Muss die Infrastruktur einen automatischen Zugriff auf private Daten verhindern?
  - Z.B. Crawling in sozialen Netzen, bei Verbrauchsdaten (Rial & Danezis 2012), Bewegungsdaten etc.

- Macht "informed consent" angesichts der Vielfalt der Möglichkeiten der Analyse noch Sinn?

Autoren-Erkennung

- Plagiatserkennung
- Wikipedia Vandalismus
  - mutwillig, ideologisch
- Benchmark PAN (Stein @CLEF 2013)

	Eigenschaften	Klasse
Bekannte Items	0 1 2 4	A
	3 5 6 8	B
	0 1 2 4	A
	2 5 7 9	D
	3 6 7 8	C
Neue Items	2 3 5 6	?
	0 1 2 4	?
	3 6 7 9	?

Extraktion eines Modells

⇩

Anwen-dung des Modells

Plagiats-Erkennung

- Stil-Messung
- Unabhängig vom Inhalt
- Beispiele für extrahierte Eigenschaften:
  - Häufigkeit von Wortklassen (POS)
    - insbesondere Pronomen und Stoppwörter
  - Länge von Wörtern und Sätzen
  - Verwendung von Wörten weicht von allgemeines Häufigkeit ab
  - Umfang Wortschatz (Stein @ PAN)

Wikipedia Vandalismus

- Beispiele für extrahierte Eigenschaften
  - Umfang Änderung
  - Zeit für die Änderung
  - Autor eingeloggt
  - Anzahl Revisionen des Artikels
  - Reputation des Landes der IP
  - ...
  - Anzahl Pronomen erste Person hinzugefügt/gelöscht
  - Anzahl beleidigende Wörter hinzugefügt/gelöscht

(Potthast @ PAN)

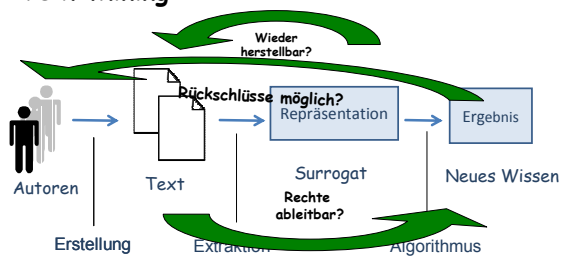
Autoren-Erkennung: ethische Debatte

- Stellungnahmen einiger ASTen zum Einsatz on Software zur Plagiatserkennung
  - Zentrale Speicherung studentischer Arbeiten?
  - Maschinelles Lernen möglich?
  - Wem gehört die Repräsentation, wer darf davon profitieren?
    - Dem Autor, Studierenden, der den Text erstellt hat
    - Dem Unternehmen, das die Repräsentation erstellt hat
- Abwägung zwischen Persönlicheitsrechten und der Notwendigkeit, Plagiate zu identifizieren

Autoren-Erkennung: ethische Debatte

"Zentrale elektronische Datensammlungen bergen grundsätzlich die Gefahr von unerwarteten Datenaggregationen, -manipulationen und -verwertungen, zumal studentische Werke personenbezogene (Name, Matrikelnr.) und personenbeziehbare (Lehrveranstaltung, Schreibstil, inhaltliche Bezüge) Daten enthalten. Auch aus diesem Grund ist eine zentrale Speicherung, und sei sie noch so kurz, von in elektronischer Form abgegebenen studentischen Werken nicht wünschenswert, ob nun an der Universität Hildesheim oder bei Dritten."
(asta-hildesheim.de)

Text Mining

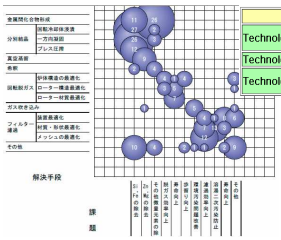


Patente

- ca. 2Mio. neue Patente jährlich
- 80% der Information in Patenten kann nicht in anderen Quellen gefunden werden (Thomson Corp. 2007)
- Patente beschreiben neues Wissen als erstes, lange vor bspw. Lehrbüchern
- Patente erlauben die Beobachtung von Wettbewerbern
- Patent-Beobachtung kann Doppelentwicklungen verwickeln
  - z.B. Sofortbildkamera Kodak vs Polaroid (bustpatents.com)

Europäisches Patentamt
European Patent Office
Office européen des brevets
Intellectual Property Office
Deutsches Patent- und Markenamt

Patent-"Landkarten"



	Effect 1	Effect 2	Effect 3
Technology 1	[AAA 1993] [US Pat. XX/XXXX]		[BBB 2002]
Technology 2	[CCC 2000]		
Technology 3		[US Pat. YY/YYYY]	[US Pat. ZZ/ZZZ [USP WW/WWW

Through <Technology>closed-loop feedback control</Technology>, the system could <Effect><Value>minimize</Value> the <Attribution>power loss<Attribution> </Effect>.

<Technology>CRF-based approach</Technology> obtained a <Effect><Attribute>precision</Attribute> o <Value>0.935</Value></Effect>.

Suzuki et al. 2010; Nanba et al. 2010 @NTCIR

Patent Trend Mining

- Analyse des Informationsverhaltens von Patent-Experten
- Entwicklung von semantischen und statistischen Verfahren zur Analyse von Trends
- Systeme zur Beobachtung von Veränderungen in Themenfeldern
  - Topic Detection and Tracking (TDT)
  - Technologien und ihre Elemente
- Kooperation:

KISTI Korea Institute of Science and Technology Information

FIZ Karlsruhe Advancing Science

Wissenschaftliche Texte

- Text Mining Verfahren können die Produktivität wissenschaftlicher Arbeit erhöhen
- Gilt Text Mining als "permissible use"?
- Ist das generierte Wissen als "derivative work" zu betrachten?
  - BioMed Licence

BioMed Central copyright and Open Access license agreement

In submitting a research article ('article') to any of the journals published by BioMed Central I certify that:
1. I am authorized by my co-authors to enter into these arrangements.
2. I warrant, on behalf of myself and my co-authors, that:
the article is original, has not been formally published in any other peer-reviewed journal, is not under consideration by any other journal and does not infringe any existing copyright or any other third party rights;
I am/we are the sole author(s) of the article and have full authority to enter into this agreement and in granting rights to BioMed Central are not in breach of any other

"Derivative Work" means a work based upon the Work or upon the Work and other pre-existing works, such as … , condensation, or any other form in which the Work may be … transformed

I agree to BioMed Central's Open Data policy
And I agree to the following license agreement:
BioMed Central Open Access license agreement
Brief summary of the agreement
Anyone is free:
to copy, distribute, and display the work;
to make derivative works;
to make commercial use of the work;
Under the following conditions: Attribution
the original author must be given credit;
for any reuse or distribution, it must be made clear to others the license terms of this work are;

Wissenschaftliche Texte

- *Hargreaves Review*
  - *"a text- and data-mining exception to copyright should be created"*
  - *"non-consumptive" use*
- Verlage blockieren den Zugang zu Texten für Text Mining Verfahren
  - für UK: McDonald & Kelly 2012
- Generierung neuen Wissens wird verhindert

Fair Use Policy of Carnegie Mellon University

exceptions generally, and fair use in particular.

fair use doctrine are to preserve free speech and to promote creativity. Codified in the Copyright Act of 1976 at 17 U.S.C. § 107, the preamble to the fair use doctrine lists six favored purposes: criticism, comment, teaching, scholarship, and research. For this reason, the Supreme Court has recognized that "the fair use defense affords considerable latitude for scholarship and comment."

The Fair Use Factors

The ultimate test of fair use "is whether copyright's goal of promoting the Progress of Science and useful Arts would be better served by allowing the use than by preventing it."

…

Digitization of works to provide access to the print disabled, to enable indexing, and to enable "non-consumptive research" (e.g. text mining) has been considered transformative.

Members of the University community are required by Section 107 to consider and balance the following factors to determine if a use qualifies as a fair use. The factors should not be balanced mechanically, but weighed together "in light of the purposes of copyright." The ultimate test