

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/281631457>

# Leipzig Corpus Miner – A Text Mining Infrastructure for Qualitative Data Analysis – PRESENTATION

Data · September 2015

CITATION

1

READS

262

3 authors, including:



**Andreas Niekler**

University of Leipzig

64 PUBLICATIONS 400 CITATIONS

[SEE PROFILE](#)



**Gregor Wiedemann**

Hans-Bredow-Institut für Medienforschung

74 PUBLICATIONS 791 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Actionality classes and cross-linguistic coding tendencies. Typological research and development of an analysis software tool [View project](#)



Computing Semantic Change Using Context Volatility [View project](#)

TKE 2014

# **Leipzig Corpus Miner**

## A Text Mining Infrastructure for Qualitative Data Analysis

Andreas Niekler | [aniekler@informatik.uni-leipzig.de](mailto:aniekler@informatik.uni-leipzig.de)  
Gregor Wiedemann | [gregor.wiedemann@uni-leipzig.de](mailto:gregor.wiedemann@uni-leipzig.de)  
Gerhard Heyer | [heyer@informatik.uni-leipzig.de](mailto:heyer@informatik.uni-leipzig.de)

NLP Group | Department of Computer Science  
University of Leipzig  
Augustusplatz 10  
04109 Leipzig

# Introduction

- Humanists, social scientists and media analysts working with text as primary data have been opening up to large scale text analysis procedures
- Users lack a computer science background
- NLP experts lack background knowledge about requirements of social science research
- Leipzig Corpus Miner (LCM) builds a broad set of methods to perform content analysis (CA) tasks in large corpora
  - CA users need to express / formalize their requirements
  - NLP experts need to understand information needs of users
- LCM provides NLP analysis procedures, algorithms and visualizations

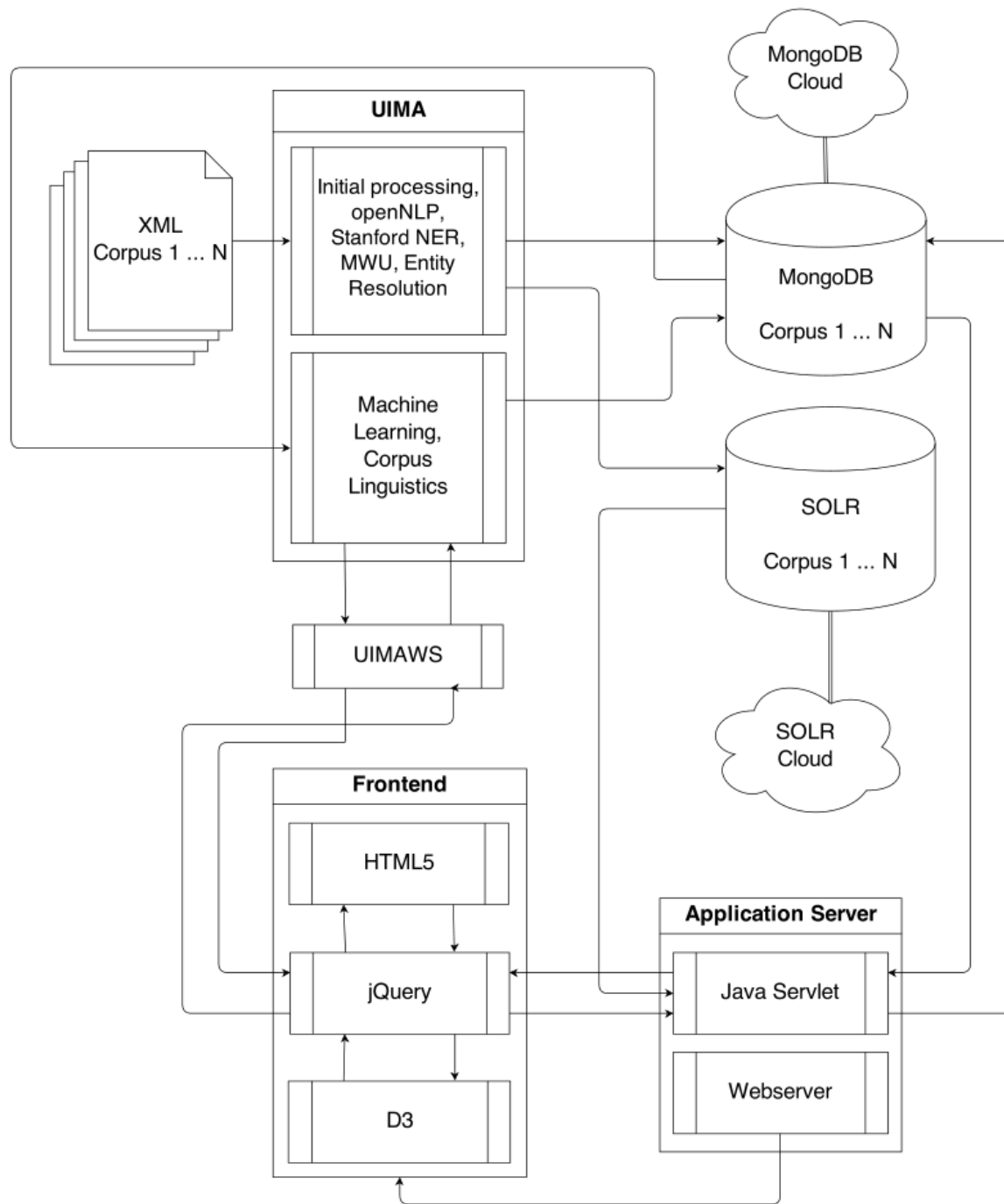
## So where is the difference?

- NLP experts provide infrastructure: set of tools which enable analysts to support certain steps of an applied method and to extend the size of collections under investigation to a degree which could not be handled manually
  - Rather than apply a certain method or static limited workflow the LCM can be used to compile workflows from a set of algorithms and their result presentation
  - Where other computer assisted CA studies only apply one or just a few methods LCM allows for *integrated* application of multiple procedures
  - Separation of the computer science view from the users view
- humanists retain control over procedures
  - control of data selection, workflows and parameters
  - support of theory-led and exploratory research by supervised and unsupervised procedures

# Architecture

- LCM is a combination of different technologies to make qualitative data analysis accessible by an interface for people who are unfamiliar with NLP
- The methodological understanding of the users is focused.
- The technical issues are hidden.
- To solve this requirement we need to address the following tasks:
  - Data storage
  - Retrieval
  - Machine Learning / NLP
  - Presentation

# Architecture



Custom  
ClearTK  
UIMA  
YAML  
OpenNLP  
Solr  
LibSVM  
StanfordNLP  
MongoDB  
jQuery

- Text Mining: Central objective of the LCM is to enable analysts to perform computational text analysis without explicit guidance by NLP experts
  - user-friendly web application
  - analysis workflows specific to methodological requirements
- UIMAWS middleware:
  - processes can be executed, stopped and managed by the users – computational expensive operations in the background
- Data management:
  - Users work on collections which are subselections of all documents stored in a corpus
  - results of analysis procedures stored per collection / user → may be used as input resources for further steps

# Analysis capabilities

- **Information Retrieval**

- Full text index
- Contextualized dictionaries which can be retrieved from a reference corpus of paradigmatic documents
- Retrieved documents can be stored as collections
- Collections can be refined (removing, adding of documents)

- **Lexicometrics**

- Frequency analysis
- Co-occurrence analysis (different significance measures)
- Automatic extraction of key terms



# User-interface

**Search Results** Time Series Document View Facets

Ergebnisdokumente

Seite 1 von 11.144 10 Zeige 1 - 10 von 111.437

Date	Score	Title	Subtitle	Paragraph
21 Feb. 1946 ZEIT DIE ZEIT_GESELLSCHAFT_POLITIK Seite: 0	0.011935912	Brasiliens neuer Präsident		Am 31. Januar wurde General Enrico Gaspar Dutra, der neue Staatspräsident Brasiliens, feierlich in sein hohes Amt eingeführt. Sondermissionen der meisten Staaten und die beflaggten fremden Kriegsschiffe...
21 Feb. 1946 ZEIT POLITIK_POLITIK_KERN Seite: 0	0.01219983	SOPHOKLES: ANTIGONE	Zur Aufführung im Deutschen Schauspielhaus in Hamburg	Das Thema der „Antigone“ ist die „Kollision“ zweier sittlicher Mächte: Familienliebe und Staatsgesetz. So definiert es Hegel, und er deutet den metaphysischen Sinn der Tragödie als die Vernichtung der...
21 Feb. 1946 ZEIT POLITIK_POLITIK_KERN Seite: 3				Die erste Tagung der Vereinten Nationen in London wurde in der Nacht zum Freitag beendet. In der Hauptversammlung sprach der britische Premierminister Attlee die Abschiedsworte. Die Hauptversammlung w...
21 Feb. 1946 ZEIT POLITIK_POLITIK_KERN Seite: 0				Von JAN MOLTOR Manchmal, beim Vorübergehen, blicken sie in erleuchtete Fenster. Sie sehen vielleicht einen runden Tisch, ein Stück Bücherregal, ein Stück Tapete. Sie sehen ein paar Quadratmeter eines...
21 Feb. 1946 ZEIT POLITIK_POLITIK_KERN Seite: 0				Die Vereinten Nationen haben ihre erste Tagung beendet. Aus den Trümmern einer Welt, die nur der Zerstörung und der Vernichtung, dem Kriege und dem Hasse gelebt, erhebt sich das Traumbild einer besser...
21 Feb. 1946 ZEIT POLITIK_POLITIK_KERN Seite: 0				Wie ein anderes, fast nicht minder bewegtes Jahrhundert sich gewissenhaft die Frage nach „dem Berufe unserer Zeit zur Gesetzgebung“ vorlegte, so bangen wir heute vor der Frage nach dem Berufe unserer ...
28 Feb. 1946 ZEIT POLITIK_POLITIK_KERN Seite: 0				„Die Lohn- und Preisfront ist eingedrückt, nicht durchbrochen“, lautet die Formel, die Präsident Truman als Ergebnis des Kompromisses in der Lösung des Stahlarbeiterstreiks geprägt hat. Die Löhne sind...
28 Feb. 1946 ZEIT POLITIK_POLITIK_KERN Seite: 2				Als Bulgarien und Rumänien im September 1944 von der Roten Armee besetzt wurden, erwarteten weite Kreise, daß dies eine Veränderung der staatlichen Struktur bedeuten würde — sogar den Sturz der Throne...
28 Feb. 1946 ZEIT POLITIK_POLITIK_KERN Seite: 0				Von ERNST SAMHABER Vor ihrem Auseinandergehen hat die Hauptversammlung der Vereinten Nationen noch den Beschluß gefaßt, allen Völkern höchste Sparsamkeit zu empfehlen, um die große gemeinsame Not zu L...
28 Feb. 1946 ZEIT POLITIK_POLITIK_KERN Seite: 0				Die Verfassungsgebende französische Nationalversammlung arbeitet eine Erklärung der Menschenrechte aus. Nach Zeiten der Unterdrückung hat man sich schon oft darauf besonnen, daß es ein gewisses Maß von...

Zeige 1 - 10 von 111.437

**Results** Details View DocumentView Task Scheduler

**Standard Parameters**

Collection: 527b6dfe4b04a07ee51bde, FAZ

Analysis: Topic Model

☒ Replace token with multi-word-units and entities.

☒ Transform entities to canonical form.

Baseform reduction/Stemming: Stemming

☒ Remove stopwords.

☒ Transform to lowercase.

N-gram: 1

Prune all words below frequency: 20

Prune all words above frequency: 2000

☐ Use paragraph as document.

Minimum length of document: 500

Topic Model: Online Hierarchical Dirichlet Process

Submit Request

## Source corpus selection

Select a Source: Nachrichtenartikel

Simple Detailed Custom

Simple search. Terms may be combined with + and -

Keyword:

☐ Use raw text

From Date:

To Date:

Select a Paper: SZ TAZ ZEIT FAZ

Select a Publication Type: NEWSPAPER MAGAZINE

Select a Section: REGIONALES POLITIK\_KERN GESELLSCHAFT\_POLITIK WIRTSCHAFT

Search

# Analysis capabilities



Postdemokratie und Neoliberalismus

Artikelsuche Collection Worker Kleine Artikelsammlung (taz99) NL-Wörterbuch Kategorien

Results Details View DocumentView Task Scheduler

## Select cooccurrence view:

>Reset Chart<

Display Type

graph

Cooccurrence Type

directed

Cooccurrence Significance

Log-Likelihood

☐ Invert X Initialization.

☐ Invert Y Initialization.

Root max Links:

7

Leaves max Links:

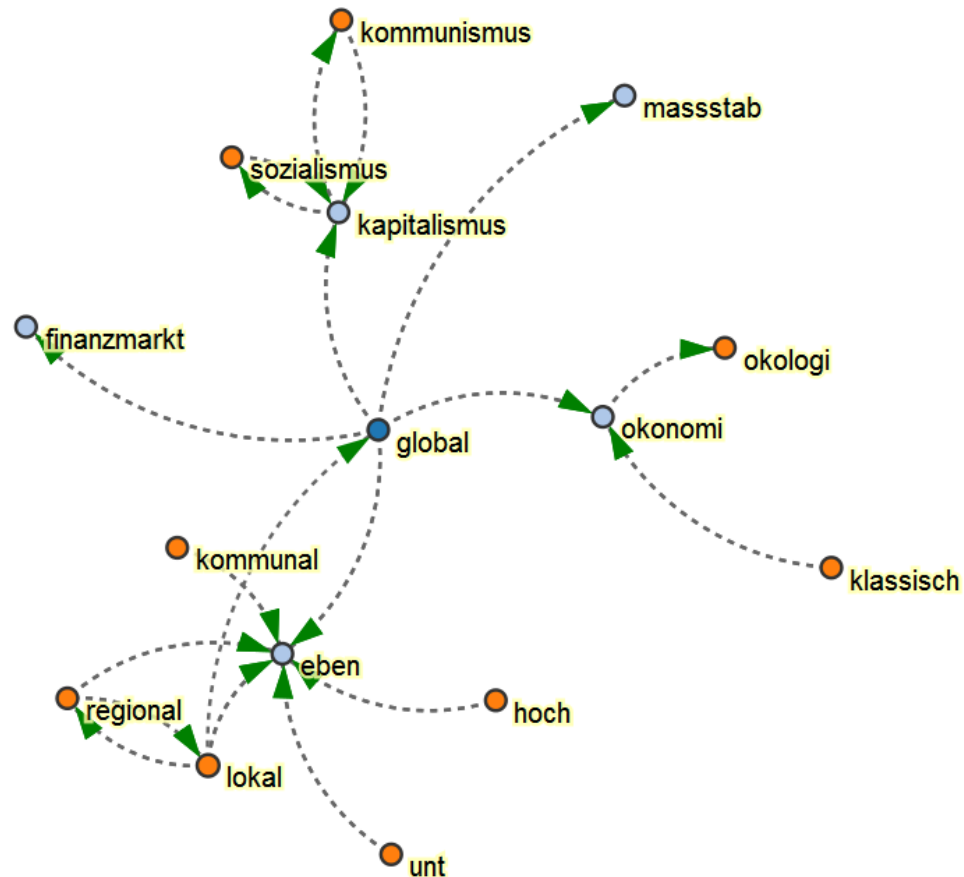
7

Graph Force:

0

>Load Words<

Type to filter



Collecti



Import

(kommunist), 1846  
(organisation), 1843  
(mitglied), 1840  
(best), 1837  
(zusatz), 1832  
(global), 1818  
(konservativ), 1812

# Analysis capabilities

Results Details View DocumentView Task Scheduler

## Term Extraction Result

Seite 1 von 4		25	Zeige 1 - 25 von 96
Wordform	Pitman-Yor Weight		
schwarz	0.05251332927826234		
steuerreform	0.04104232258294118		
armut	0.040581085075405164		
wirtschaftspolitik	0.037536447784544284		
konservativ	0.0890054032650477		
großbritannien	0.07330368722464492		
familie	0.02555741217025264		
stadt	0.0327194282984891		
regierungen	0.04507211646803578		
natur	0.036788306817679105		
euro	0.10359476908453702		
wahlen	0.02126337353003894		
koalition	0.033774624420120615		
frauen	0.04028180934580125		
konjunktur	0.022515109104455075		
theorie	0.05519041007823491		
umwelt	0.06031665466422563		
interesse	0.018841952749104107		
europäischen	0.031574558859018956		
spd	0.09099110584538517		
energie	0.04339872022388537		
china	0.06585250516412187		
arbeitnehmer	0.055637593895247646		
kultur	0.028125689030592517		
...	0.015276187522327211		

Click another term or generate new examples by clicking term again.

1. , Wirtschaftssprecher der Labour-Fraktion, stieß denn auch in die verwundbarste Stelle, nämlich die Fähigkeit der Thatcher-Regierung, die **Konjunktur** sicher und stetig zu steuern. „Das wirkliche Wirtschaftswunder“, höhnte | **Document: 50e4129be4b09f954b7984a9** | **Source:articles**

2. Dagegen die Europäische Union: Regierungen und Ökonomen bezweifeln, dass sich Europas **Konjunktur** aus eigener Kraft erholt. Sie setzen ausgerechnet wieder auf die USA, deren Aufschwung | **Document: 50ddb82ae4b09f954b73d9b9** | **Source:articles**

3. Heute muss die **Konjunktur** als Entschuldigung für politische Passivität herhalten. Ein Aufschwung wie der aktuelle ist willkommenes Valium für die Bundesregierung. Mehr als zwei Prozent | **Document: 51d4082fe4b09a55c10807d6** | **Source:articles**

4. Andererseits hat die Diskussion um die Gemeingüter gerade **Konjunktur**: Der Bankrott des neoliberalen Denkmodells führt zur Besinnung auf traditionelle Werte, das ungeahnte Tempo | **Document: 5093d914e4b04fe27ff295a3** | **Source:articles**

5. vorgesehen. Außerdem hat die Bank von Japan schon im Frühjahr die Leitzinsen gesenkt und damit auch von der monetären Seite her die Voraussetzungen für ein Wiederanziehen der **Konjunktur** geschaffen. Rechtzeitig | **Document: 50a751d4e4b079434e09f8a3** | **Source:articles**

6. In der Politik hat das Wort „Zumutung“ ja gerade **Konjunktur**. Was würden Sie Ausländern zumuten, damit sie ihren Teil zur Integration beitragen? | **Document: 50929adee4b04fe27fef9ee5** | **Source:articles**

# Analysis capabilities

- **Topic Models** (Topic Models are statistical models which infer probability distributions over latent variables, assumed to represent topics, in text collections as well as in single documents - e.g. LDA, Blei et. al 2003)
  - Unsupervised clusters of topics
  - Retrieve documents for different topics
  - Topic time series analysis
- **Classification**
  - Creation of hierarchical category systems
  - Annotation of training data of different granularity (Paragraph, Sentence, Document).
  - Multiple Labels can be assigned to a annotation
  - Usage of different classification algorithms (libSVM, Naive Bayes)
  - Features on lexical and syntactical level are used

# Analysis capabilities



Postdemokratie und Neoliberalismus

Artikelsuche [Collection Worker](#) [Kleine Artikelsammlung \(taz99\)](#) [NL-Wörterbuch](#) [Kategorien](#)

[Results](#) [Details](#) [View](#) [DocumentView](#) [Task Scheduler](#)

## Select graph resolution

>Reset Chart<

Year

Document Count

relative (Collection)

minProbability: 0.5



Probability: 0.1278



**moralisch mensch kriegem**  
wort ... meinung geschehen wahrh  
eit schreiben intellektuell



Probability: 0.1273



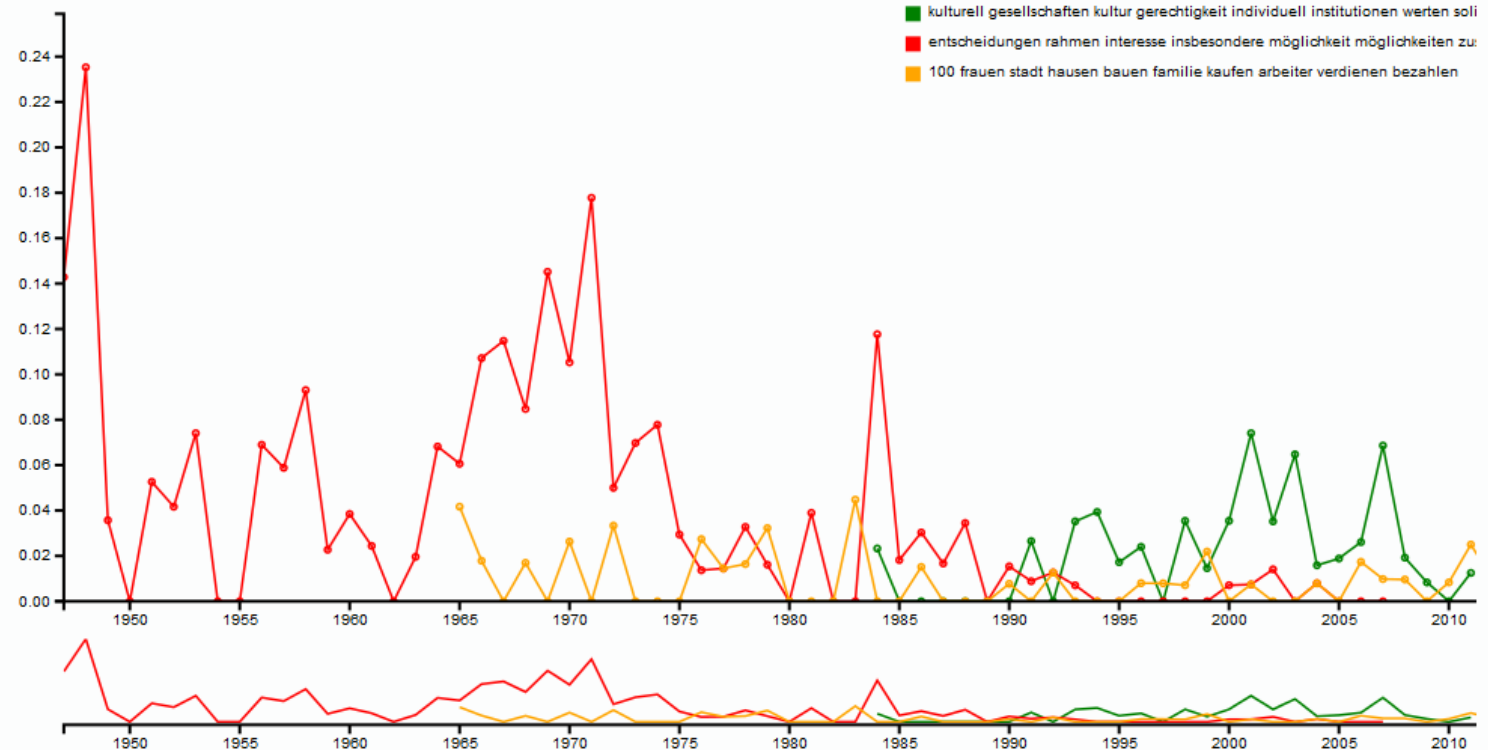
**entscheidungen rahmen in**  
teresse insbesondere möglick  
eit möglichkeiten zusammenhan  
g entscheidung beteiligen ergeben



Probability: 0.0734



**kulturell gesellschaften kul**  
tur gerechtigkeit individuell in  
stitutionen werten solidarität al





# Analysis capabilities

Categories Annotations Document View Classifications

## Metadata

Date: 09 Oct. 1982

Publisher: FAZ

Token: 2460

Section: Wirtschaft\_wirtschaft

Subsection:

Page: 13

Subject:

Type: newspaper

Language: de

Author: Professor Dr. Michael Zöllner

File: /data/postdemokratie/rohdaten  
/FAZ/2002/1982/821009\_FAZ\_0013\_13\_0001

ID: 51d4044de4b09a55c107cef5

## Allgemein

- ☐ ROOFTITLE
- ☐ SENTENCE
- ☐ TOKEN
- ☐ TITLE
- ☐ SUBTITLE
- ☐ PARAGRAPH

## Nomen

- ☒ POS\_NN
- ☐ POS\_NE

## Adjektive

- ☐ POS\_ADJA
- ☐ POS\_ADJD

## Verben

- ☐ POS\_VVFIN

## Wirtschaft

### Die Sozialpolitik braucht neue Methoden

Das Prinzip der Bevormundung ist mit der freien Ordnung unvereinbar / Von Professor Dr. Michael Zöllner

Langsam spricht sich herum, daß der Betreuungs- und Versorgungsstaat nicht mehr finanzierbar ist, daß die sozialstaatlichen Wohltaten letztlich doch von den vermeintlichen Begünstigten selbst zu bezahlen sind. So ist der Sozialstaat ins Gerede gekommen. Doch, obwohl vieles dafür spricht, den anatomischen Ort der Vernunft im Portemonnaie zu vermuten, wäre es verfehlt, die fällige Debatte um den Sozialstaat und seine Nebenwirkungen nur unter dem Aspekt der Finanzierung zu führen. Es geht längst um grundsätzlichere Fragen, nämlich um die Ziele und die Methoden der Sozialpolitik.

Fragt man aber nach den Zielen der gegenwärtigen Sozialpolitik, dann wird sich eine Vorstellung von sozialer Verantwortung oder Solidarität mittlerweile veranlassen. Unter vielen sind die folgenden Äußerungen des Abgeordneten Dr. Michael Zöllner ein Beispiel: "Die Belastung durch steigende Krankenversicherungsbeiträge trifft alle Versicherten. Die Einführung der Selbstbeteiligung trifft aber nur einen Teil der Versicherten, die Krankheit Leistungen der Krankenversicherung in Anspruch nehmen müssen. Sozialdemokraten das - gegen das Prinzip der Solidarität." Folgt man dem Abgeordneten Zöllner, so ist die gegenseitige Absicherung von Belastungen, die der einzelne nicht mehr der solidarischen Vorsorge trifft die Vergesellschaftung aller Risiken, und das Spiel dabei keine Rolle mehr.

Hand in Hand mit dieser Veränderung des Bewußtseins verwandelt sich jedoch die Herrschaft. Dabei geht es vor allem um zwei Entwicklungen, die den Charakter der Herrschaft empfindlich verändert haben. Die Methoden, mit denen der moderne Sozialstaat die Herrschaft erstens eine Vielzahl falscher Anreize, die sich in ihrer Summe zu einer Lasten vereinigen, und sie führen zweitens in Gestalt einer alle Lebensbereiche neuen Form paternalistischer Herrschaft.

Das Gesundheitswesen als ein lehrreiches Beispiel

Für die erste Gruppe von Symptomen, also die Abwälzungsmechanismen, bietet das Gesundheitswesen Fülle von Beispielen. Dieses Gesundheitswesen eignet sich, um angeliehenden Politikwissenschaftlern - sozusagen am Krankenbett - zu zeigen, welche Anreize volkswirtschaftliche Rationalität und individuelles Nutzenkalkül miteinander in Institutionen so einrichten kann, daß den Beteiligten die erwünschten Verhaltensweisen wie man schließlich politische Entscheidungsverfahren so gestalten kann, daß die Verantwortlichen sich vermengen. Zu den Krankheiten des Gesundheitswesens und des gesamten Sozialsystems gehört also zunächst einmal seine politische Struktur, wobei zwischen der Schwäche der Selbstverwaltungen, und der politischen Manipulierbarkeit der

## Projects

- ☒ Epol Argumentation
- ☒ Andreas Taggt sich blöde
- ☒ Evaluation ANGW
- [Add new project](#)

## Annotations

Epol Argumentation



### 1. Absatz mit Argumentationszusammenhang

Leiht sich die öffentliche Hand Geld, um damit zum Beispiel Konjunkturprogramme zu finanzieren, wird Kapital gebunden. Der Zins ist der Preis des Kapitals? und der Preis eines Gutes steigt, wenn es knapper wird. Darunter würden die Unternehmen und die Haushalte leiden, weil auch für sie die Finanzierungskosten steigen. Schlimmstenfalls verdrängen die staatlichen Ausgaben die privaten Ausgaben, die Wirtschaft kommt nicht vom Fleck.

view document certainty: 0.96

approve deny ignore

Wer Massenarbeitslosigkeit beseitigen will - das zeigen internationale Vergleiche -, der muss vor allem im unteren Bereich sozialer Hierarchie ansetzen. Wenn auf den Preisverfall geringqualifizierter Arbeit die Senkung der Arbeitseinkommen insgesamt folgt, so das kleine Einmaleins des Neoliberalismus, dann kann die Massenarbeitslosigkeit zügig abgebaut werden. Die Folge: Die öffentlichen Haushalte blühen auf. Angewendet auf die Weltwohlfahrtsnische Deutschland heißt das: Der Reißwolfkapitalismus frisst die Regelsysteme der Tarifautonomie und des Sozialstaates, er verursacht Macht- und Lebensstandardseinbrüche und gefährdet so die Grundlagen der Freiheit.

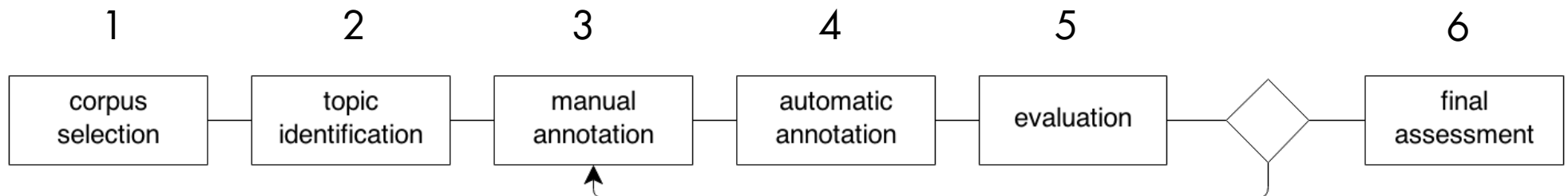
view document certainty: 0.95

approve deny ignore

- ☒ affirmativer Text
- ☒ Thematischer Text
- ☒ kritischer Text

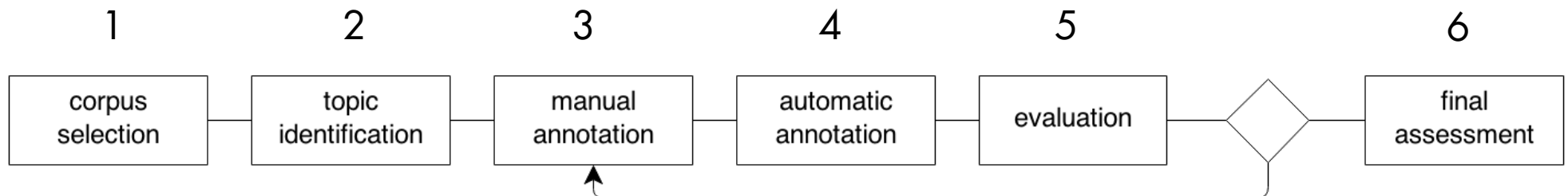
# Use Case

- Content Analysis study on “Neoliberalism” by political theorists
- Social science objective: identify changes in discursive patterns of policy justifications in public media
- a corpus consisting of 3.5 million German newspaper articles from 1946 to 2012 is investigated
- (simplified) workflow realized with the LCM



# Use Case

- 1) Select a subset of relevant documents
  - Using documents of themes and modes that reflect the research question to formulate key terms and co-occurrence patterns
  - On this basis a set of 10.000 documents was retrieved from the data source to formulate an initial collection (Wiedemann/Niekler 2014)
- 2) Topic Model processing on the set of 10.000 documents
  - Identification of thematic clusters (policy fields)
  - Identification of irrelevant topics and documents



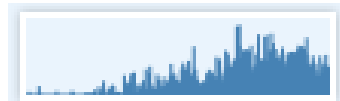


# Use Case

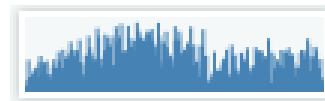
score	length	year	title	
347,22	685	1977	Pro und kontra Mehrwertsteuer	[Pro's and con's of VAT]
321,81	662	1973	Oelkrise und Konjunktur	[Oil crisis and economy]
290,48	705	1966	Energie muß billig sein	[Energy has to be cheap]
289,34	687	1977	Die Steuern senken	[Lower the taxes]
287,26	845	1964	Korrektur der Einkommensteuer	[Correction of VAT]
281,07	687	1971	Die Bauern im Nacken	[The farmers at the neck]
279,74	884	1965	Was ist uns die Mark wert?	[What is the „Mark“ worth to us?]
272,75	682	1970	Steuern mit der Steuer	[Governing with taxes]
264,82	719	1971	Ohne Abkühlung keine Stabilität	[No stability without slowdown]
262,81	671	1973	Das sicherste Mittel	[The most secure instrument]
261,33	707	1972	Entlastung – wovon?	[Relief – of what?]
254,97	676	1979	Das Fernsehen und die Angst	[Television and fear]
254,93	704	2011	Nicht ernst gemeint: die Quote	[Quotas not meant serious]
251,53	457	1977	Eine Konfliktstrategie der Union	[A conflict strategy of the EU]
250,46	638	2010	Die neue Auflehnungsbereitschaft	[The new willingness for rebellion]

# Use Case

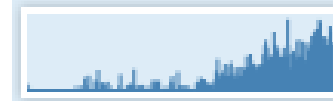
kulturell, 0.0048  
gesellschaften, 0.0040  
kultur, 0.0038  
gerechtigkeit, 0.0037  
individuell, 0.0034  
institutionen, 0.0032  
werten, 0.0031  
solidarität, 0.0030  
globalisierung, 0.0029  
kollektiv, 0.0029  
gleichheit, 0.0025  
moralisch, 0.0024  
alternativ, 0.0023



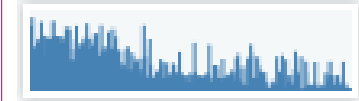
nachfragen, 0.0051  
wirtschaftspolitik, 0.0051  
löhnen, 0.0048  
bundesbank, 0.0036  
konjunktur, 0.0034  
zinsen, 0.0033  
bundesregierung, 0.0030  
steigend, 0.0030  
unternehmer, 0.0029  
rezession, 0.0030  
stabilität, 0.0029  
währung, 0.0029  
real, 0.0028



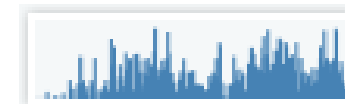
märkten, 0.0072  
global, 0.0069  
euro, 0.0058  
banken, 0.0057  
globalisierung, 0.0056  
regeln, 0.0045  
weltweit, 0.0042  
ökonomien, 0.0041  
kapital, 0.0039  
regierungen, 0.0036  
schulden, 0.003454  
kapitalismus, 0.003497  
finanzmärkte, 0.0031



landwirtschaft, 0.0067  
industrie, 0.0041  
bauern, 0.0039  
verbraucher, 0.0031  
landwirtschaftlich, 0.0028  
erheblich, 0.0028  
industriell, 0.0027  
gesamt, 0.0025  
kaufen, 0.0025  
betragen, 0.0023  
agrarpolitik, 0.0022  
höhen, 0.0020  
nachfragen, 0.0020

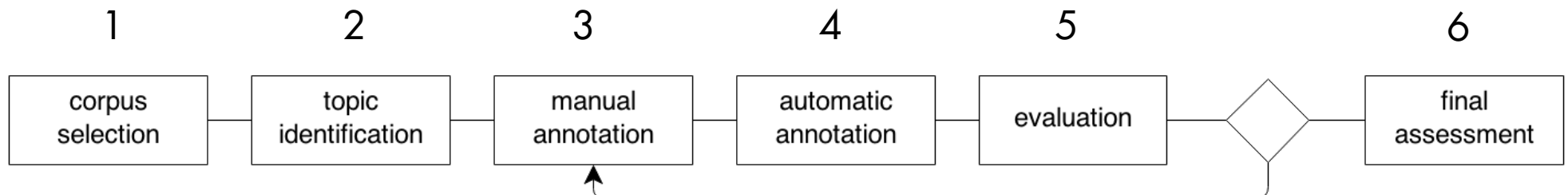


wissenschaft, 0.0083	wissenschaftler, 0.0034
wissenschaftlich, 0.0068	Information, 0.0034
medien, 0.0045	genetisch, 0.0033
fernsehen, 0.0041	kulturell, 0.0033
informationen, 0.0037	internet, 0.0033
computer, 0.0037	mensch, 0.0027
forschung, 0.0035	kultur, 0.0025



# Use Case

- 3) Annotation of training data
  - The document set of 10.000 is a list ranked by relevancy
  - → select most relevant articles per topic for manual annotation
  - political scientists argumentative structures in newspaper articles (annotation with hierarchical category systems developed for this task)
  - Relevant parts in the documents are annotated
- 4) Automatic classification
  - An automatic classification procedure is invoked on the unlabeled data to identify more text parts containing argumentative structures. The NLP group supports the analysts by identifying discriminating linguistic features.



# Use Case

- Example category schema
  - Paragraph with argumentation
    - Claim
    - Argument
      - Date
      - Reasons
      - Marker
      - Conclusion
  - Bias
    - critically
    - Affirmative
  - Type
    - Not relevant
    - Meta
    - Text with topic

## Examples for category Bias - affirmative

Wer vom Ertrag seiner Arbeit nicht leben kann, sollte eine Steuergutschrift bekommen – das hilft mehr als ein Mindestlohn

Those who can not live off the proceeds of his labor, should get a tax credit - which helps more than a minimum wage

Die Sozialpolitik braucht neue Methoden Das Prinzip der Bevormundung ist mit der freien Ordnung unvereinbar

Social policy needs new methods, the principle of paternalism is incompatible with the free order

Weitere Steuererhöhungen scheinen kaum noch tragbar, viel eher müßte die Steuerbürde vermindert werden. Besonders notwendig wäre eine Milderung offenkundiger Härten bei der Lohnsteuer und bei der Besteuerung kleiner Unternehmen.

More tax increases seem hardly portable, much more the tax burden would be reduced. Particular need would be a mitigation of manifest hardship on the payroll tax and the taxation of small businesses.

# Use Case

## Examples for automatic classification

Unter reinen Marktbedingungen sind Bildungs-, Gesundheits-, und Pflegedienstleistungen für viele nicht bezahlbar.

Under pure market conditions educational, health, and care services is not affordable for many.

... dass der Staat stets der schlechtere Akteur am Markt und deshalb außerstande ist, die Geschicke privater Unternehmen, namentlich von Banken zu leiten?

...that the state is always the worse player on the market and therefore is incapable of the fortunes of private enterprises, particularly to lead by banks?

Klar ist auch, daß die Chancen für den selbständigen Unternehmer abnehmen, wenn sich der Staat im Unternehmenssektor auf Kosten, der Privaten ausdehnt.

It is also clear that the opportunities for self-employed entrepreneurs decrease as the state expands in the corporate sector at the expense of the entrepreneurs

### 1. Absatz mit Argumentationszusammenhang

Ablösung sehr unterschiedlicher und damit Wettbewerbs- und freizügigkeitshemmender nationaler Regelungen durch europäische Regeln, die in den meisten Fällen kein Produkt eines Harmonisierungswahns sind, sondern die Voraussetzung dafür, daß Grenzen verschwinden.

[view document](#) certainty: 0.99

[approve](#) [deny](#) [ignore](#)

Da hier große Mengen Geld im Spiel sind (siehe den Artikel von Daniel Baudru und Bernard Maris), ist die Versuchung groß, die Erfordernisse der medizinischen Versorgung als zweitrangig zu betrachten. Die Internationale Gewerkschaft des öffentlichen Dienstes (IÖD) vertritt daher die Position, daß ?die Wasserversorgungsunternehmen, ob es sich um staatliche, halbstaatliche oder private Unternehmen handelt, verpflichtet werden müssen, dieses Gut zu sozial akzeptablen Preisen zu liefern. Die beste Lösung besteht darin, die Bereitstellung und Verwertung staatlichen Institutionen zu übertragen.? Denn ?ein unbeschränkter Wettbewerb im Bereich der Wasserversorgung und ?aufbereitung liegt nicht im öffentlichen Interesse?16.

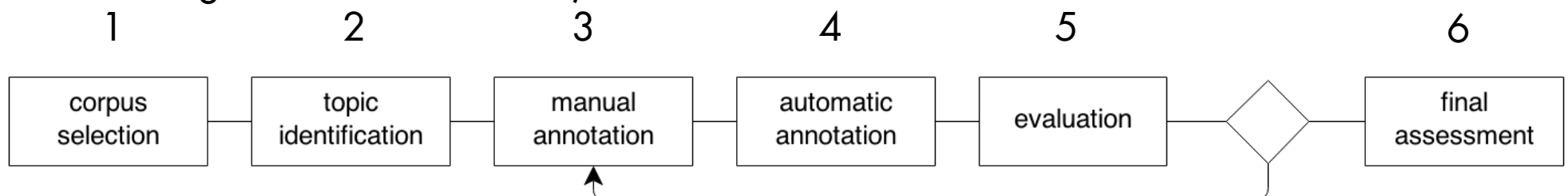
[view document](#) certainty: 0.99

[approve](#) [deny](#) [ignore](#)

Darüber hinaus kann ein moderner Sozialstaat den Aufbau qualitativ

# Use Case

- 5) Classification review
  - Text snippets identified in the previous step (supposedly containing arguments of interest) are presented to the analysts ranked by certainty of label assignments.
  - Analysts can verify or reject the results manually.
  - In this active learning paradigm we calculate internal precision / recall measures while the analysts are evaluating the process qualitatively. If those ongoing evaluations show satisfactory results, the process of creating training data is concluded.
- 6) Rerun classification task with corrected results on entire collection
  - The classification procedure is run on the entire collection under investigation.
  - Results can be described qualitatively and quantitatively (e.g. proportions of categories over time slices).



# Conclusion

- The LCM supports manual Content Analysis (CA) via basic corpus linguistic procedures as well as supervised state-of-the-art Text Mining techniques
- The methods can be used separately and independently
- It reflects the diversity of CA instead of providing a static predefined processing
- Extensions
  - Our target is to integrate more methods for semantic analysis and changes over time
  - For the classification process we need to integrate more feature engineering methods
  - Scalability

# References

1. Biemann, C., Heyer, G., Quasthoff, U., Richter, M.: *The leipzig corpora collection - monolingual corpora of standard size*. In: Proceedings of Corpus Linguistics Conference (2007)
2. Blei, D.M.: *Probabilistic topic models: Surveying a suite of algorithms that offer a solution to managing large document archives*. Communications of the ACM 55(4), 77–84 (2012)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: *Latent dirichlet allocation*. Journal of Machine Learning Research 3, 993–1022 (2003)
4. Bostock, M., Ogievetsky, V., Heer, J.: *D3: Data-driven documents*. IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis) 17(4), 2301 – 2309 (2011)
5. Büchler, M.: *Medusa: Performante Textstatistiken auf großen Textmengen: Kookkurrenzanalyse in Theorie und Anwendung* (2008)
6. Burghardt, M., Wolff, C.: *Stand off-Annotation für Textdokumente: Vom Konzept zur Implementierung (zur Standardisierung?)*. In: Proceedings of the Biennial GSCL Conference. pp. 53–59 (2009)
7. Ferrucci, D., Lally, A.: *Uima: An architectural approach to unstructured information processing in the corporate research environment*. Nat. Lang. Eng. 10(3-4), 327–348 (2004)
8. Finkel, J.R., Grenager, T., Manning, C.: *Incorporating non-local information into information extraction systems by gibbs sampling*. In: Proceedings of the 43<sup>rd</sup> Annual Meeting on Association for Computational Linguistics. pp. 363–370 (2005)
9. Glaser, B.G., Strauss, A.L.: *Grounded theory: Strategien qualitativer Forschung* (2005)
10. Heyer, G., Holz, F., Teresniak, S.: *Change of topics over time and tracking topics by their change of meaning*. In: Proceedings of the International Conference on Knowledge Discovery and Information Retrieval. pp. 223–228 (2009)
11. Hoffman, M.D., Blei, D.M., Bach, F.R.: *Online learning for latent dirichlet allocation*. In: Neural Information Processing Systems (NIPS). pp. 856–864 (2010)
12. Krippendorff, K.: *Content analysis: An introduction to its methodology*. SAGE, 3 edn. (2013)
13. Laclau, E., Mouffe, C.: *Hegemony and socialist strategy*. Verso, 2 edn. (2001)
14. Lemke, M.: *Die Okonomisierung des Politischen: Entdifferenzierungen in kollektiven Entscheidungsprozessen*. Discussion Paper Nr. 2. Hamburg and Leipzig (2012), <http://www.epol-projekt.de/discussion-paper/discussion-paper-2/>
15. Lughofer, E.: *Hybrid active learning (hal) for reducing the annotation efforts of operators in classification systems*. Pattern Recognition 45(2), 884–896 (2012)
16. Mayring, P.: *Qualitative Inhaltsanalyse: Grundlagen und Techniken*. Beltz, 11 edn. (2010)
17. Niekler, A., Jähnichen, P., Heyer, G.: *ASV Monitor: Creating comparability of machine learning methods for content analysis*. In: Proceedings of the 2012 European Conference on Machine Learning and Knowledge Discovery in Databases – Volume Part II. pp. 812–815 (2012)
18. Ogren, P.V., Wetzler, P.G., Bethard, S.: *ClearTK: A UIMA toolkit for statistical natural language processing*. In: Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP workshop at Language Resources and Evaluation Conference (LREC). pp. 865–869 (2008)
19. Scharnow, M.: *Automatische Inhaltsanalyse und maschinelles Lernen*. Epubli, 1 edn. (2012)
20. Steinberger, R., Pouliquen, B., Kabadjov, M., Belyaeva, J., van der Goot, E.: *JRC-NAMES: A freely available, highly multilingual named entity resource*. In: Proceedings of the International Conference Recent Advances in Natural Language Processing. pp. 104–110 (2011)
21. Teh, Y.W., Jordan, M.I.: *Hierarchical Bayesian nonparametric models with applications*. In: Bayesian Nonparametrics. 1 edn. (2010)
22. Wiedemann, G.: *Opening up to big data: Computer-assisted analysis of textual data in social sciences*. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research 14(2) (2013)
23. Zhou, X., Zhang, X., Hu, X.: *Semantic smoothing for bayesian text classification with small training data*. In: Proceedings of the SIAM International Conference on Data Mining. pp. 289–300 (2008)