Identification and Information Extraction on Scientists Homepages in the Common Crawl Web Archive

Bachelorarbeit Präsentation

Samuel Roth

9. April 2018

Albert-Ludwigs-Universität Freiburg
Technische Fakultät
Institut für Informatik
Lehrstuhl für Algorithmen und Datenstrukturen

Content

Problemstellung

Problemstellung

In einem Satz

Extrahiere strukturierte Daten über Wissenschaftler aus einem Webarchiv.

Problemstellung

In einem Satz

Extrahiere strukturierte Daten über Wissenschaftler aus einem Webarchiv.

In Stichpunkten

- Eine große Menge an Webseiten - *Common Crawl*.
- Identifiziere in dieser Menge *Homepages* von Wissenschaftlern.
- Extrahiere aus den *Homepages* zugehörige Daten wie *Name*, *Geschlecht*, *Institution*, *Beruf*.
- Finde Texte über die gefundenen Personen in den Webseiten.
- Erstelle aus diesen Daten ein Index für die semantische Suchmaschine *Broccoli* [Bast et al., 2012].

Implementierung

Common Crawl als Rohdatenquelle

Offenes Internet Archiv

- Seit 2009 insgesamt 47 Crawls
- Ca. 3 Milliarden Seiten pro Crawl.
- Zuletzt ein Crawl alle zwei Monate.
- Frei zugänglich via Amazon S3 Bucket.

Common Crawl als Rohdatenquelle

Offenes Internet Archiv

- Seit 2009 insgesamt 47 Crawls
- Ca. 3 Milliarden Seiten pro Crawl.
- Zuletzt ein Crawl alle zwei Monate.
- Frei zugänglich via Amazon S3 Bucket.

Speichert HTML Seiten im **WARC** format.

```
1  WARC/1.0
2  WARC-Type: response
3  WARC-Date: 2014-08-02T09:52:13Z
4  Content-Length: 43428
5  Content-Type: application/http; msgtype=response
6  WARC-IP-Address: 212.58.244.61
7  WARC-Target-URI: http://news.bbc.co.uk/2/hi/africa/3414345.stm
8  WARC-Payload-Digest: sha1:M63W6MNGFDWXDSLTHF7GWUPCJUH4JK3J
9  WARC-Block-Digest: sha1:YHKQUSBOS4CLYFEKQDVGJ457OAPD6IJO
10
11 HTTP/1.1 200 OK
12 Vary: X-CDN
13 Cache-Control: max-age=0
14 Content-Type: text/html
15 Date: Sat, 02 Aug 2014 09:52:13 GMT
16 Connection: close
17
18 <!doctype html public "-//W3C//DTD HTML 4.0 Transitional//EN" >
19 <html>
20 <head>
21 <title>
22     BBC NEWS | Africa | Namibia braces for Nujoma exit
23 </title>
24 ...
```

Common Crawl als Rohdatenquelle

Download von relevanten WARC records.

- Aus `http://univ.cc` extrahierte Liste von 9599 Domains für Universitäten.
- Für jede Universität:
  Suchen im Common Crawl Index nach allen im Archiv vorhandenen Seiten.
- Download aus Amazon S3 $\approx 38,5$ Millionen Seiten.

Klassifikation

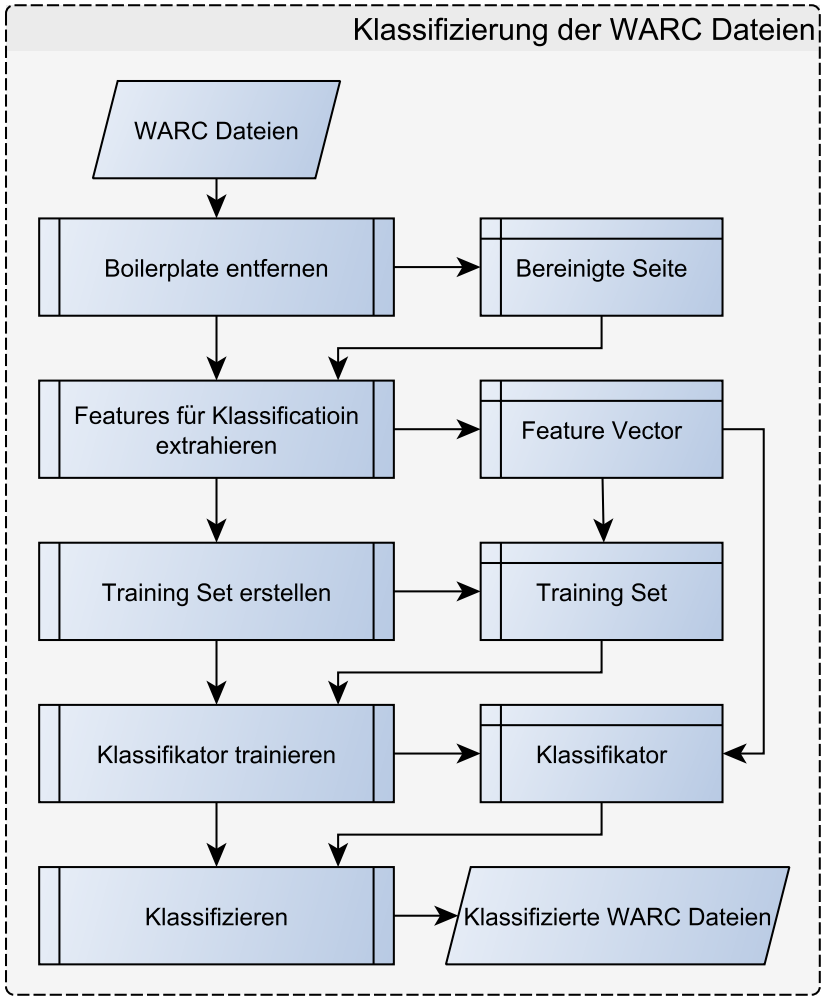


Abbildung 1: Klassifikationsprozess

Klassifikation

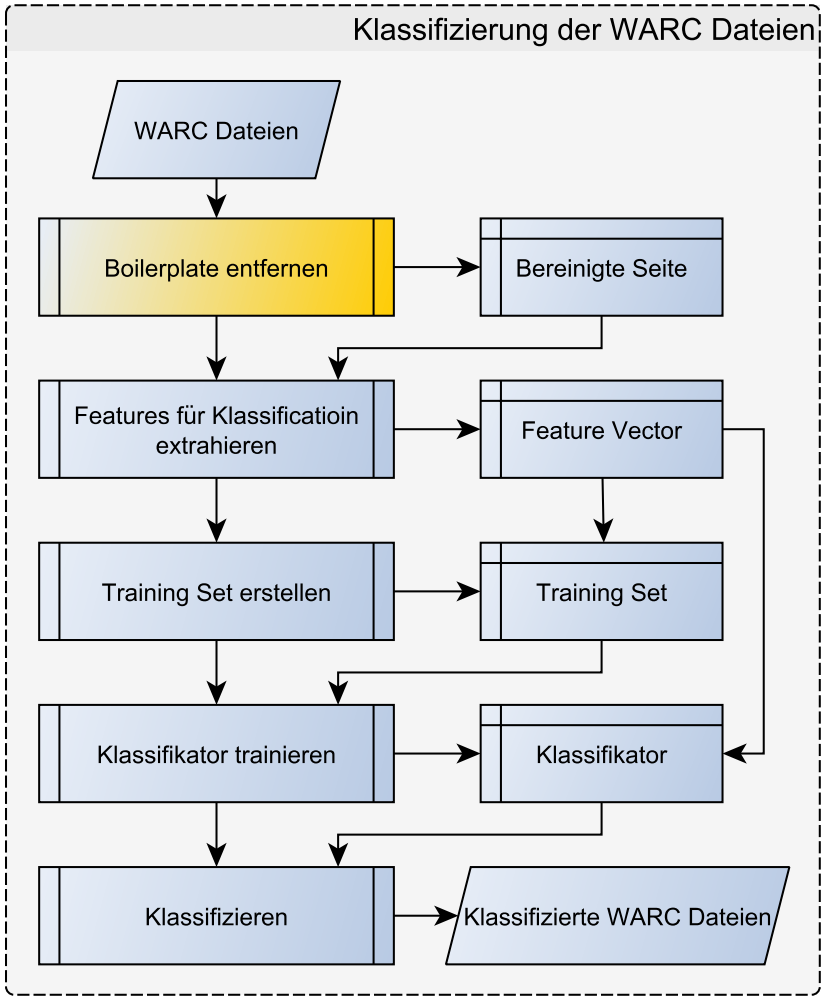


Abbildung 1: Klassifikationsprozess

Klassifikation - Boilerplate

- Relevanten Inhalt Erkennen.
- Entfernen von Navigation, Kopf-/Fußzeile etc.
- Hier *JusText* Algorithmus [Pomikálek, 2011] verwendet.

Klassifikation - Features

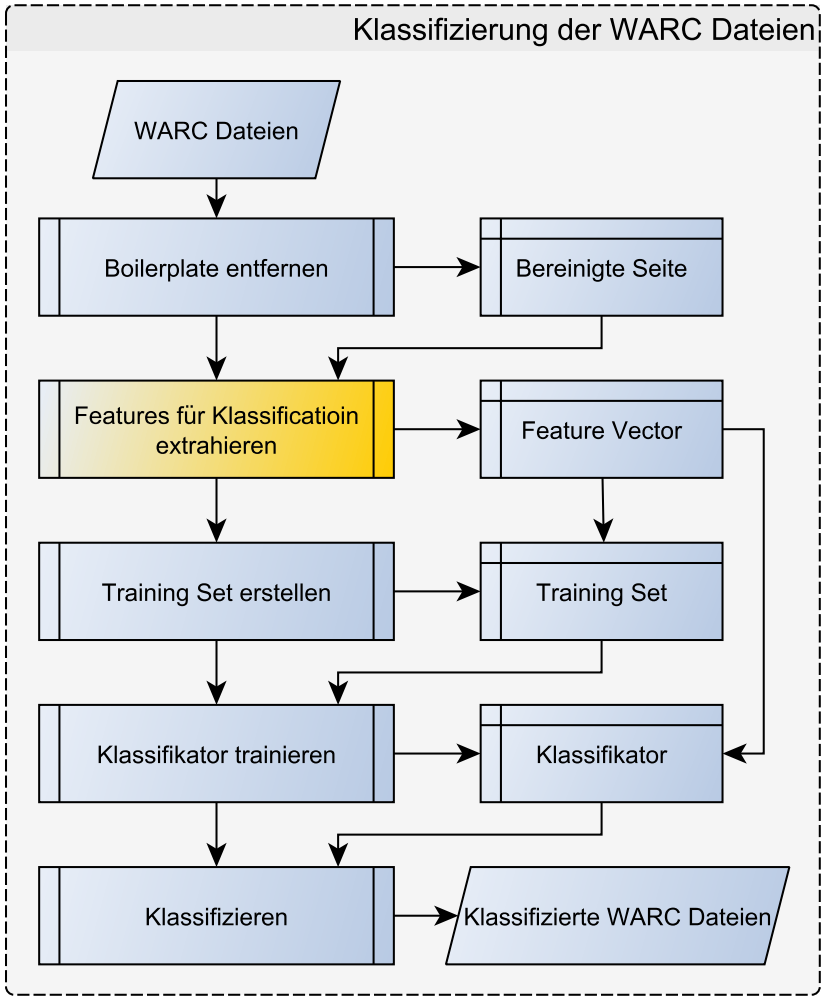


Abbildung 2: Klassifikationsprozess

Klassifikation - Feature Auswahl

Text

- Wort *N*-Gramme.
- Textinhalt, Überschriften, Titel der Seite.
- Prefix für die Worte aus den verschiedenen Ursprüngen
  z.B. *headingparam* *publications*

Klassifikation - Feature Auswahl

Text

- Wort *N*-Gramme.
- Textinhalt, Überschriften, Titel der Seite.
- Prefix für die Worte aus den verschiedenen Ursprüngen
  z.B. *headingparam* *publications*

URLs

- Ähnlich wie in [Gollapalli et al., 2013] vorgeschlagen.
- `http://www.cs.cmu.edu/~wpdann/index.html`
  ⇒ *urlparam* *tilde* *nodict*, *urlparam* *index*

Klassifikation - Trainingsdaten.

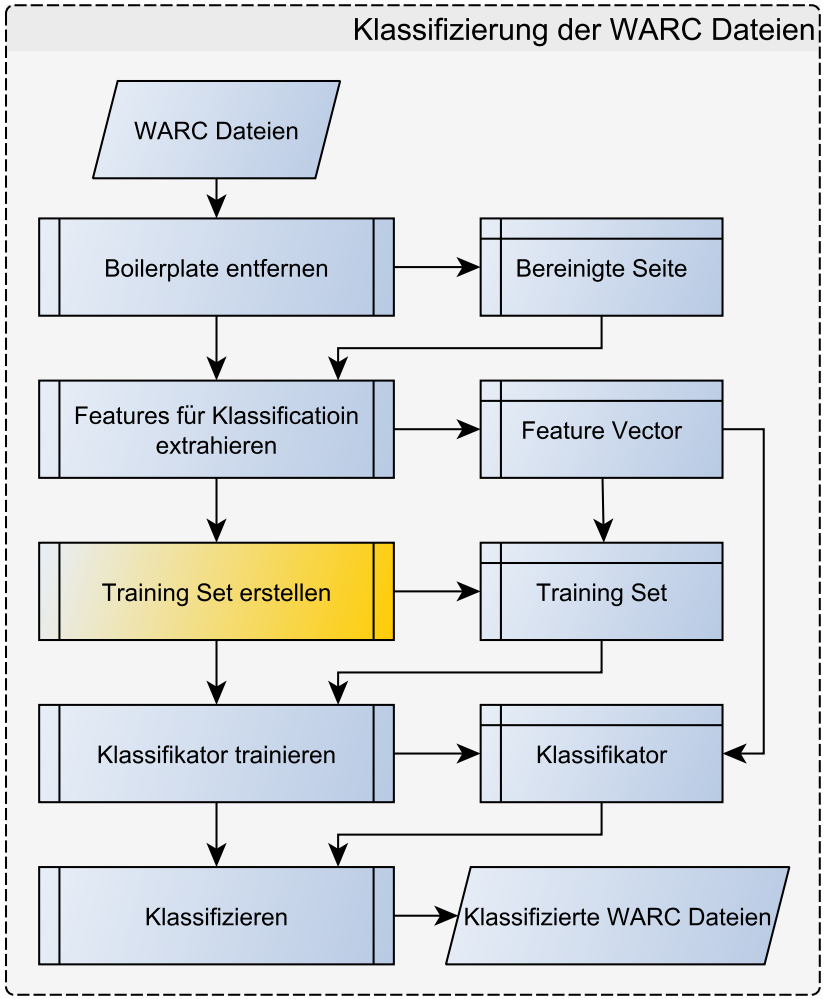


Abbildung 3: Klassifikationsprozess

Klassifikation - Trainingsdaten.

Bestehende Sammlungen klassifizierter Webseiten

- *WebKB*[1] enthält 2.902 Homepages von Wissenschaftlern und 5.214 negative Beispiele.
- *dblp*[2] eine online Bibliothek für Publikationen in der Informatik: Teilweise Links zu Homepages der Autoren $\Rightarrow$ 21.991 Homepages.

Selbst erstellte Datensätze

- 246 Homepages von Wissenschaftlern an der Universität Freiburg.
- 25.139 zufällig aus den vorhandenen Daten ausgesucht und manuell gefilterte negativ Beispiele.

Gesamt: 25.139 *"Homepage"*, 25.139 *"Nicht Homepage"* Beispiele.

[1] `http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/`
[2] `http://dblp.uni-trier.de/`

Klassifikation - Trainieren, ungelabelte Webseiten klassifizieren.

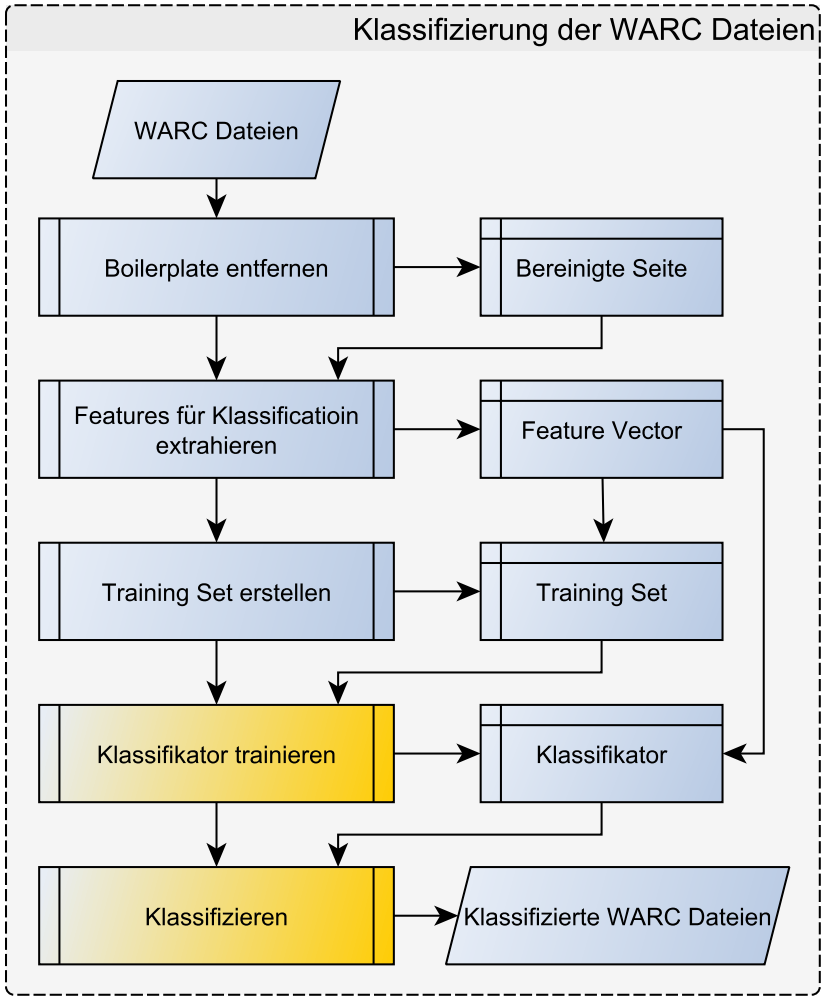


Abbildung 4: Klassifikationsprozess

Klassifikation - Trainieren, ungelabelte Webseiten klassifizieren.

Methode und Implementierung

- SVM (aus *scikit-learn*[3]) mit linearem Kernel.
- Die Features der Webseiten wurden in eine Term-Document Matrix eingetragen.
- Jeder Eintrag durch berechnen von $Tf \cdot Idf$ normalisiert.
- Training der SVM mit 80% des Trainingsset. Rest für Evaluation.
- Klassifizieren der ungelabelten Webseiten

[3] `http://scikit-learn.org`

Extrahieren von Daten - Name.

Namen erkennen

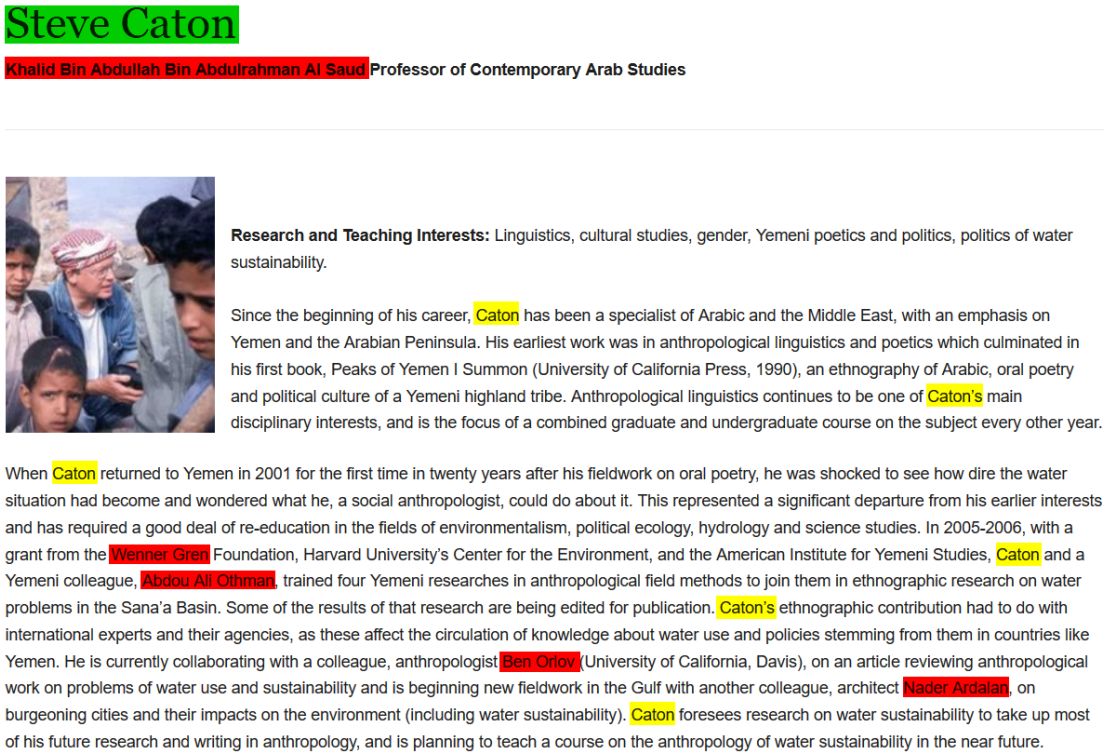
Mittels Stanford NER [Finkel et al., 2005].

Namen verschmelzen?

"*Fuzzy*" Algorithmus der versucht kleine durch den längsten Namen zu ersetzen.

Welcher Name?

SVM mit den Positionen als Features.



Steve Caton

Khalid Bin Abdullah Bin Abdulrahman Al Saud Professor of Contemporary Arab Studies

Research and Teaching Interests: Linguistics, cultural studies, gender, Yemeni poetics and politics, politics of water sustainability.

Since the beginning of his career, Caton has been a specialist of Arabic and the Middle East, with an emphasis on Yemen and the Arabian Peninsula. His earliest work was in anthropological linguistics and poetics which culminated in his first book, Peaks of Yemen I Summon (University of California Press, 1990), an ethnography of Arabic, oral poetry and political culture of a Yemeni highland tribe. Anthropological linguistics continues to be one of Caton's main disciplinary interests, and is the focus of a combined graduate and undergraduate course on the subject every other year.

When Caton returned to Yemen in 2001 for the first time in twenty years after his fieldwork on oral poetry, he was shocked to see how dire the water situation had become and wondered what he, a social anthropologist, could do about it. This represented a significant departure from his earlier interests and has required a good deal of re-education in the fields of environmentalism, political ecology, hydrology and science studies. In 2005-2006, with a grant from the Wenner Gren Foundation, Harvard University's Center for the Environment, and the American Institute for Yemeni Studies, Caton and a Yemeni colleague, Abdou Ali Othman, trained four Yemeni researches in anthropological field methods to join them in ethnographic research on water problems in the Sana'a Basin. Some of the results of that research are being edited for publication. Caton's ethnographic contribution had to do with international experts and their agencies, as these affect the circulation of knowledge about water use and policies stemming from them in countries like Yemen. He is currently collaborating with a colleague, anthropologist Ben Orlov (University of California, Davis), on an article reviewing anthropological work on problems of water use and sustainability and is beginning new fieldwork in the Gulf with another colleague, architect Nader Ardalan, on burgeoning cities and their impacts on the environment (including water sustainability). Caton foresees research on water sustainability to take up most of his future research and writing in anthropology, and is planning to teach a course on the anthropology of water sustainability in the near future.

Abbildung 5: Mehrdeutigkeit der Namen

Extrahieren von Daten - Institution/Geschlecht/Beruf.

Institution

Mapping der Domain Hosts denen die Homepage zugeordnet ist.

Geschlecht

- Bestimmung über den Namen mittels statistischer Daten.[4]
- Bestimmung über Textfeatures via SVM als Backup, falls 1. Variante fehlschlägt.

Beruf

SVM mit den gleichen Features wie Klassifikation, ohne URL-Features.

`Professor, PostDoc, Graduate, Other`

[4] `https://pypi.python.org/pypi/namegender`

Erstellen des Broccoli Index.

Wissensdatenbank

Bestehend aus

- den erhaltenen Personen zusammen mit den extrahierten Daten.
- Universitäten.

Texte

- Suche nach Namen der Personen in allen Webseiten.
- Assoziiere Pronomen in Texten auf der Homepage.
  `''I am interrested in ...''`
  `''After my PhD ...''`

Ergebnisse

Ergebnisse - Abdeckung in Common Crawl.

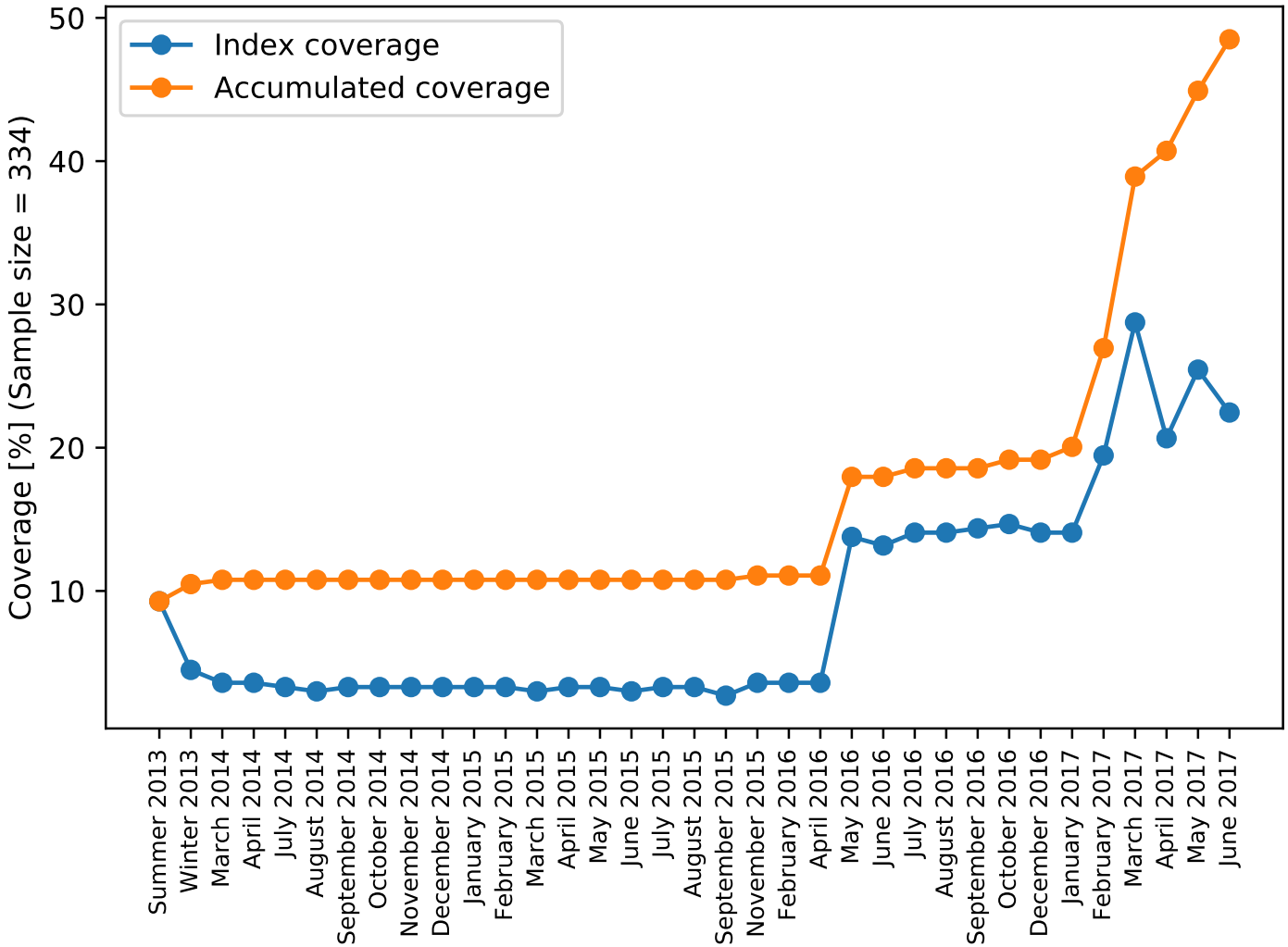


Abbildung 6: Abdeckung Homepages der Uni Freiburg

Ergebnisse - Performance der Klassifikatoren

Klassifikator	Precision $\frac{tp}{tp+fp}$	Recall $\frac{tp}{tp+fn}$	F_1-Score $\frac{2}{\frac{1}{Precision}+\frac{1}{Recall}}$	n Testset
Page	0.89	0.89	0.89	1613
Name	0.90	0.88	0.89	46
Gender	0.81	0.74	0.75	19
Profession	0.58	0.59	0.58	22

Tabelle 1: Klassifikatoren Performance Werte sind gemittelt über alle Klassen. 20% Test, 80% Trainingsdaten

Ergebnisse - Top Features nach Gewichten

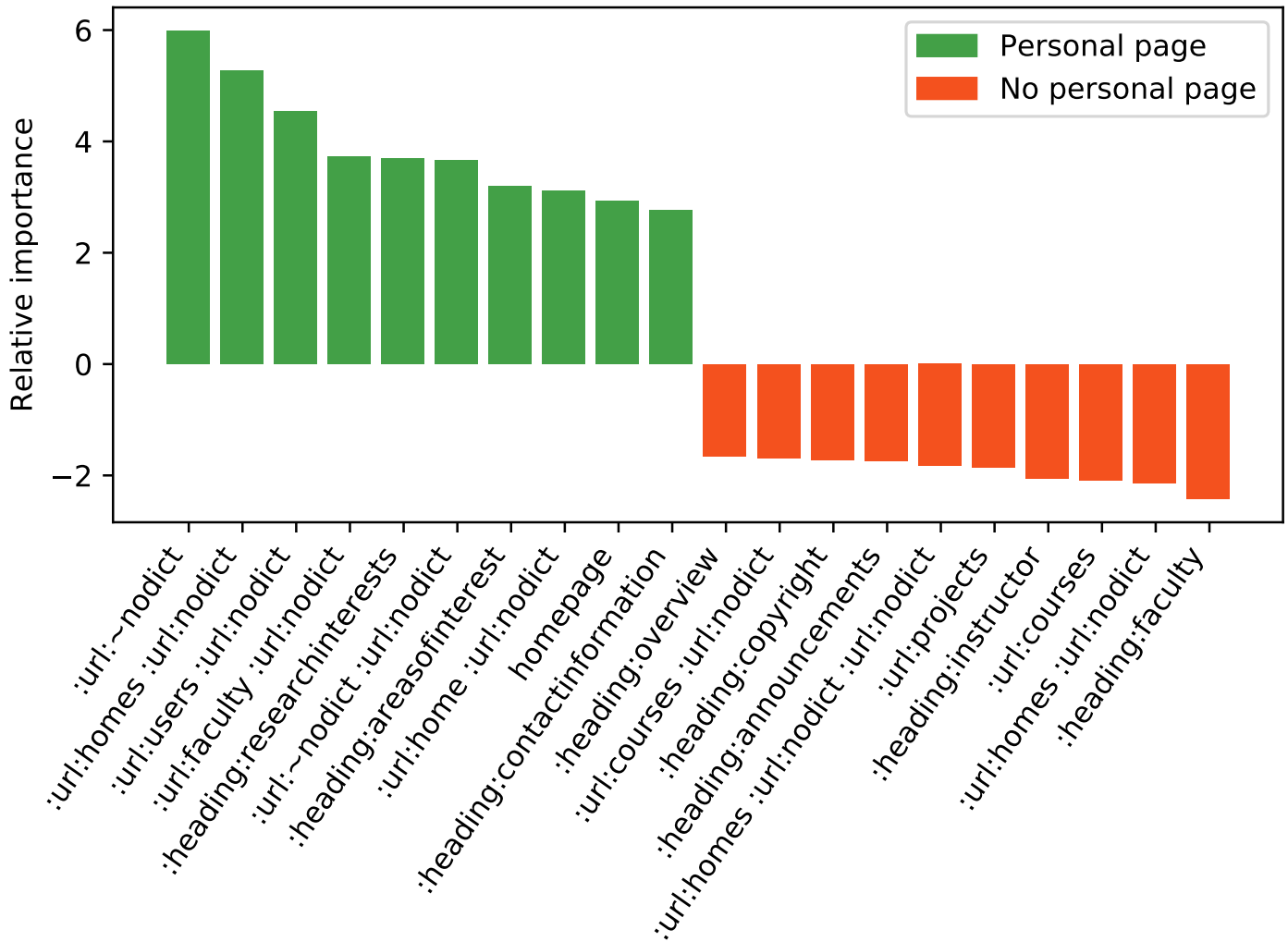


Abbildung 7: Top 20 Features nach Gewicht

Ergebnisse - Broccoli Index Recall

Quelle	Total	Common Crawl		Broccoli Index		
	n_t	n_c	$p_c := \frac{n_c}{n_t}$	n_b	$p_{cb} := \frac{n_b}{n_c}$	$p_b := \frac{n_b}{n_t}$
`informatica`[5]	113	26	23,01%	5	19,23%	4,42%
`uni-freiburg.de`[6]	334	96	28,74%	21	21,88%	6,29%
`Top 100 H-Index`[7]	100	59	59,00%	14	23,73%	14,00%

Tabelle 2: Recall für ausgewählte listen an Homepages

[5] informatica-feminale.de
`https://www.inforamatica-feminale.de/Professorinnen/Uni/listeuni.html`

[6] Homepages von unter der Domain `http://www.uni-freiburg.de`

[7] Liste der Top 100 Wissenschaftler nach *H*-index
`http://www.guide2research.com/scientists/`

Ergebnisse - Broccoli Index Precision

Suchanfrage	n	Is name	Is personal page	Is correct name
Person mit *"robotics"*	185	92.97%	79.46%	94.56%

Tabelle 3: Qualität des Ergebnis einer Suchanfrage an den erstellten Index.

Demo

Danke

Literatur i


Bast, H., Bäurle, F., Buchhold, B., and Haussmann, E. (2012).
Broccoli: Semantic full-text search at your fingertips.
arXiv preprint arXiv:1207.2615.


Finkel, J. R., Grenager, T., and Manning, C. (2005).
Incorporating non-local information into information extraction systems by gibbs sampling.
In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370.
Available at
`http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf`.


Gollapalli, S. D., Caragea, C., Mitra, P., and Giles, C. L. (2013).
Researcher homepage classification using unlabeled data.
In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 471–482, New York, NY, USA. ACM.

Literatur ii


Pomikálek, J. (2011).
Removing boilerplate and duplicate content from web corpora.
PhD thesis, Masaryk university, Faculty of informatics, Brno, Czech Republic.

Appendix

Appendix - Feature Vector für - Welcher Name

$$v_{ner}(name) := \begin{pmatrix} name \texttt{ in } title, \\ name \texttt{ in } heading^0, \\ \cdots, \\ name \texttt{ in } heading^9, \\ name \texttt{ in } paragraph^0, \\ \cdots, \\ name \texttt{ in } paragraph^9 \end{pmatrix} \in \{0,1\}^{21} \tag{1}$$

Appendix - Algorithmus für Name merging.

available_names ← *getNamesUsingNer*(*page*)
function NAMEMERGEALIAS(*name*)
    new_name ← *name*
    if *name* substring of any *available_names* **then**
        new_name ← matched name in *available_names*
    end if
    if *new_name* ≠ *name* **then**
        return NAMEMERGEALIAS(*new_name*)
    end if
    return *new_name*
end function

Appendix - Hardware Spezifikationen

Hardware Spezifikationen	
Processor	Intel(R) Core(TM) i7-3770 CPU @ 3.40GHz
RAM	16GB DDR3 1333 MHz
HDD	2 x Western Digital RE4 (3TB) using RAID 0
Internet connection	1 GBit/s synchron

Tabelle 4: Hardware Spezifikationen.

Appendix - Laufzeiten nach Aufgaben

Process	Task	n Tasks	Speed	Runtime	Bottleneck
Download WARC	Get locations	9599	$1.67\frac{loc}{s}$	$4.5h$	Internet Connection
	Download from S3	38,517,248	$37.65\frac{WARC}{s}$	$11.8d$	
Extract WARC	Get Html	38,517,248	$517.6\frac{WARC}{s}$	$20.7h$	CPU
	Rem. Boilerplate	38,517,248	$98.44\frac{WARC}{s}$	$4.5d$	CPU
Train	Train models	40.222		$74.7s$	CPU
Classify	Page	38,517,248	$448.39\frac{PPE}{s}$	$23.9h$	CPU
	NER	39,017	$126.45\frac{PPE}{s}$	$308s$	CPU
	Predict attributes	39,017	$452.2\frac{PPE}{s}$	$86s$	CPU
Broccoli Index	NER	38,478,231	$131.42\frac{Texts}{s}$	$3.4d$	CPU
	Generate KB	39,017	$452.2\frac{PPE}{s}$	$86s$	CPU
	Add texts	181,258	$71.9\frac{Text}{s}$	$0.7h$	CPU
	Score Entities	181,258	$43.6\frac{Text}{s}$	$1.1h$	CPU

Tabelle 5: Laufzeiten nach Aufgaben

Appendix - $Tf * Idf$

$N :=$ Anz. Dokumente, $df_t :=$ Anz. Dokumente mit $t \in$ Dok., $Idf_t := log\left(\frac{N}{df_t}\right)$

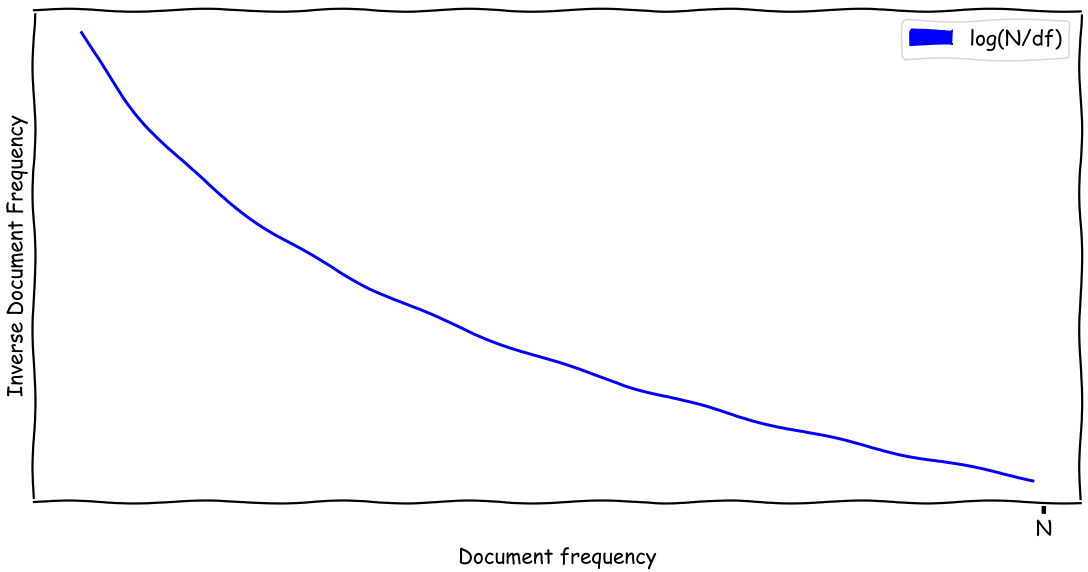


Abbildung 8: *Idf* Verlauf.

$$Tf_t := \frac{\text{Anz. } t \text{ in } D}{\text{Anz. Terme in } D}$$

Appendix - Abdeckung in Common Crawl.

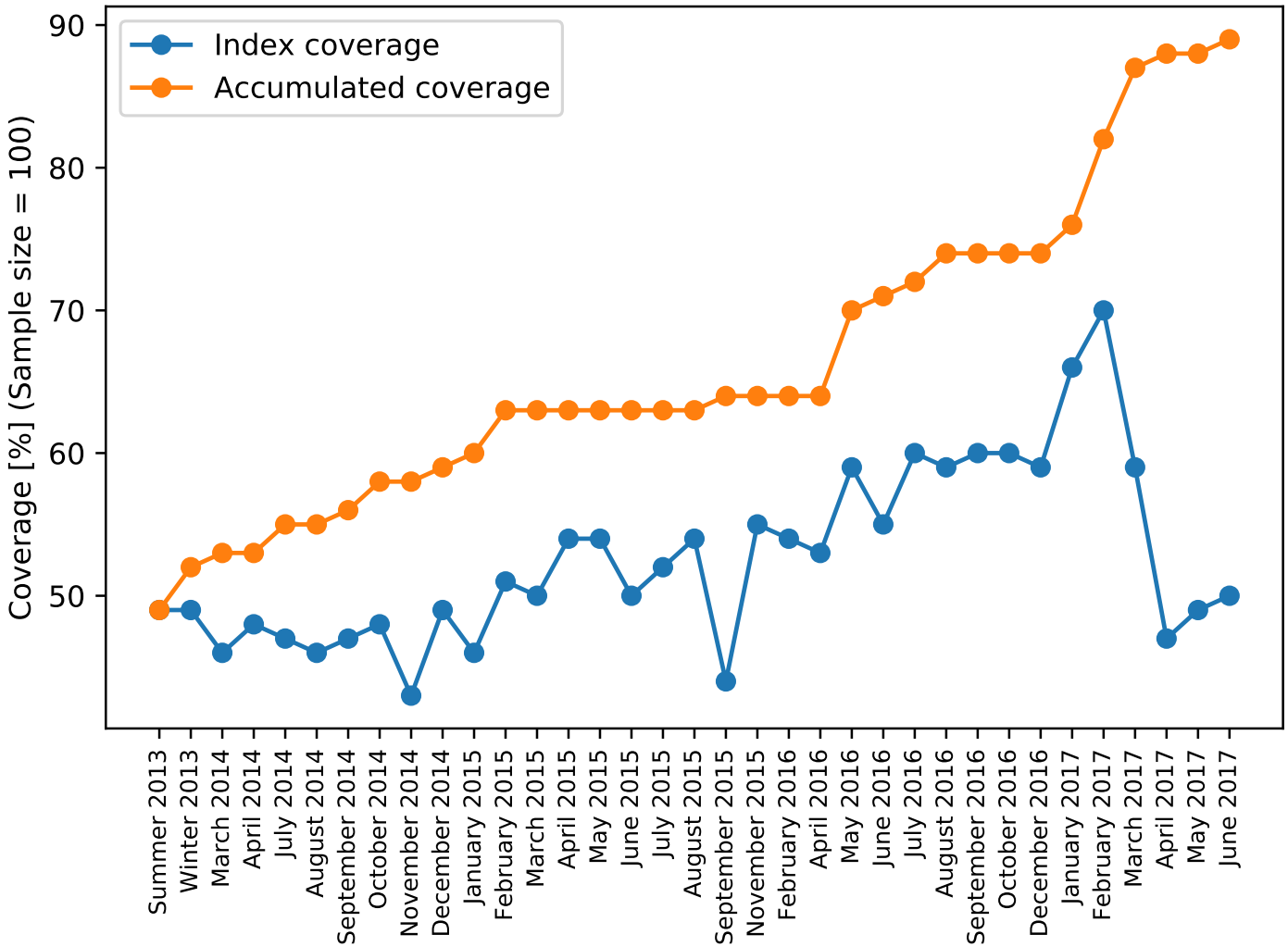


Abbildung 9: Abdeckung Top 100 Wissenschaftler nach *H*-Index

H-Index (Hirschfaktor)

Definition

Der Index h eines Wissenschaftlers wurde definiert als die größtmögliche Anzahl der Publikationen dieses Wissenschaftlers, die mindestens h-mal zitiert wurden.

Beispiel

5 Publikationen mit den Zitathäufigkeiten $h\_Index(5, 4, 3, 2, 1) = 3$
5 Publikationen mit den Zitathäufigkeiten $h\_Index(100, 100, 100, 3, 3) = 3$
5 Publikationen mit den Zitathäufigkeiten $h\_Index(4, 4, 4, 4, 1) = 4$

Zahlen

In Zahlen	
Erhaltene Personen	39.017
Text Dokumente	181.310
Lines of code	10.315
Commits	326

Tabelle 6: Die Arbeit in Zahlen.