Ein Verfahren zur Text-Reuse-Extraktion

- Masterverteidigung -

Alexander Kümmel
<alexander.kuemmel@uni-weimar.de>

Web Technologies & Information Systems
Fakultät Medien
Bauhaus-Universität Weimar

2. November 2011

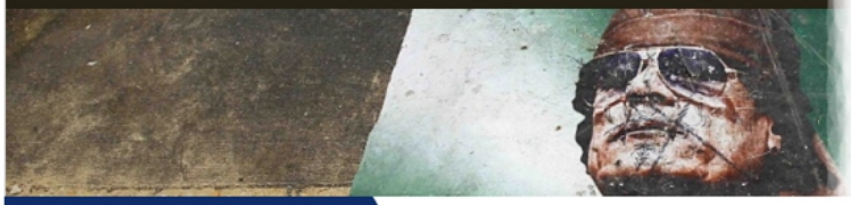
Überblick

Motivation

» Immobilienmarkt » Motormarkt » Stellenmarkt » Abo & Service SZ-Shop E-Paper Tickets

sueddeutsche.de

Politik | Wirtschaft | Geld | Kultur | Sport | Leben | Karriere | München & Region | Bayern



Home > Politik > Krieg in Libyen Gaddafi und die CIA - Souffleure des Despoten

Gaddafi und die CIA

Souffleure des Despoten

Er soll Terrorverdächtige in libysche Foltergefängnisse verfrachtet und Reden für den Despoten formuliert haben. Der US-Geheimdienst CIA soll enger als bislang bekannt mit Gaddafis Regime zusammengearbeitet haben. Das geht Medienberichten zufolge aus Dokumenten hervor, die in Tripolis aufgetaucht sind. Auch der britische MI-6 pflegte demnach ein enges Verhältnis zu dem Diktator.

Twittern 38 | Empfehlen 27 | Senden | +1 1

In Libyen gefundene Dokumente enthüllen nach Informationen der *New York Times* eine enge Kooperation zwischen dem US-Geheimdienst CIA und dem Gaddafi-Regime. So habe die CIA unter anderem acht Mal Terrorverdächtige in das für seine Folterpraxis bekannte Land zur Befragung geschickt, berichtete die Zeitung.

WELT ONLINE

Anmelden

Abo ePaper Shop News

Home | Politik | Meinung | Wirtschaft | Geld | Sport | Wissen | Gesundheit | Panorama | Kultur | Reise | Moto

In den Nachrichten: 9/11 | US Open | Erdbeben in NRW | Gaby Köster

LIBYEN 03.09.2011 Empfehlen 138

CIA soll eng mit Gaddafi zusammengearbeitet haben

Hat der US-Geheimdienst mit dem Gaddafi-Regime kooperiert? Einem Dokument nach hat der CIA Terrorverdächtige in das für seine Folterpraxis bekannte Land geschickt.

FOTO: DPA/DPA

Angeblich hat Muammar al-Gaddafi intensiver als bisher bekannt mit den USA zusammengearbeitet

In Libyen gefundene Dokumente enthüllen nach Informationen der „New York Times" eine enge Kooperation zwischen dem US-Geheimdienst CIA und dem Gaddafi-Regime. So habe die CIA unter anderem acht Mal Terrorverdächtige in das für seine Folterpraxis bekannte Land zur Befragung geschickt, berichtete die Zeitung.

ARTIKEL TEILEN

Empfänger-E-Mail ein

Weiter

WEITERFÜHRENDE LI

- Interpol schreibt Gad
- Über 100 deutsche S
- Diktator Gaddafi und
- Gaddafi verkaufte vor

THEMEN

Motivation

- Texte werden ständig wiederverwendet:
  - Zitate
  - Zusammenfassungen
  - Textbausteine
  - Paraphrasen
  - Übersetzungen
  - Plagiate
- Das Ausmaß wiederverwendeter Texte im Web ist unbekannt.
- Werkzeuge zur Untersuchung dieses Problems fehlen.
- In dieser Arbeit wurde ein neues Verfahren entwickelt, um wiederverwendete Textabschnitte aus zwei gegebenen Dokumenten zu extrahieren.
- Analyse der Texte auf abschnittsweise hohe Ähnlichkeit.

Motivation

- Texte werden ständig wiederverwendet:
  - Zitate
  - Zusammenfassungen
  - Textbausteine
  - Paraphrasen
  - Übersetzungen
  - Plagiate
- Das Ausmaß wiederverwendeter Texte im Web ist unbekannt.
- Werkzeuge zur Untersuchung dieses Problems fehlen.

- In dieser Arbeit wurde ein neues Verfahren entwickelt, um wiederverwendete Textabschnitte aus zwei gegebenen Dokumenten zu extrahieren.
- Analyse der Texte auf abschnittsweise hohe Ähnlichkeit.

Motivation

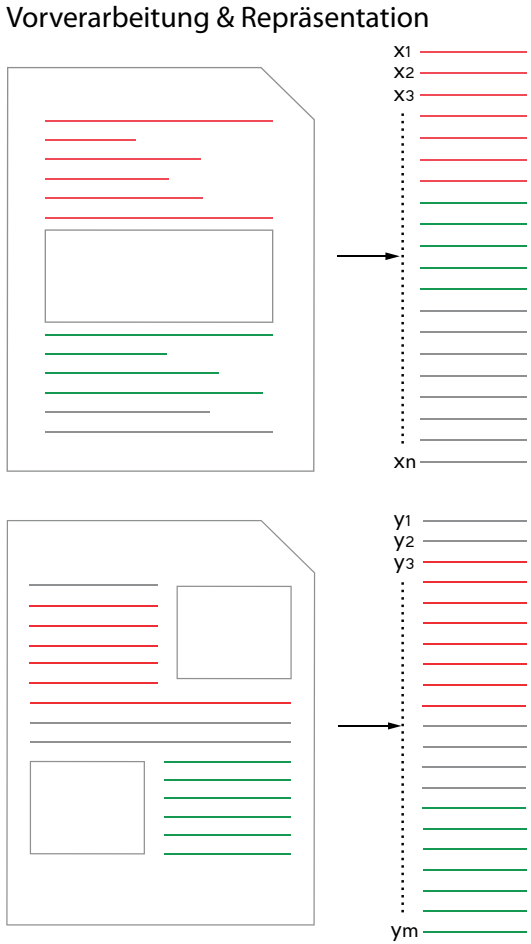
- Texte werden ständig wiederverwendet:
  - Zitate
  - Zusammenfassungen
  - Textbausteine
  - Paraphrasen
  - Übersetzungen
  - Plagiate
- Das Ausmaß wiederverwendeter Texte im Web ist unbekannt.
- Werkzeuge zur Untersuchung dieses Problems fehlen.

- In dieser Arbeit wurde ein neues Verfahren entwickelt, um wiederverwendete Textabschnitte aus zwei gegebenen Dokumenten zu extrahieren.
- Analyse der Texte auf abschnittsweise hohe Ähnlichkeit.

Überblick

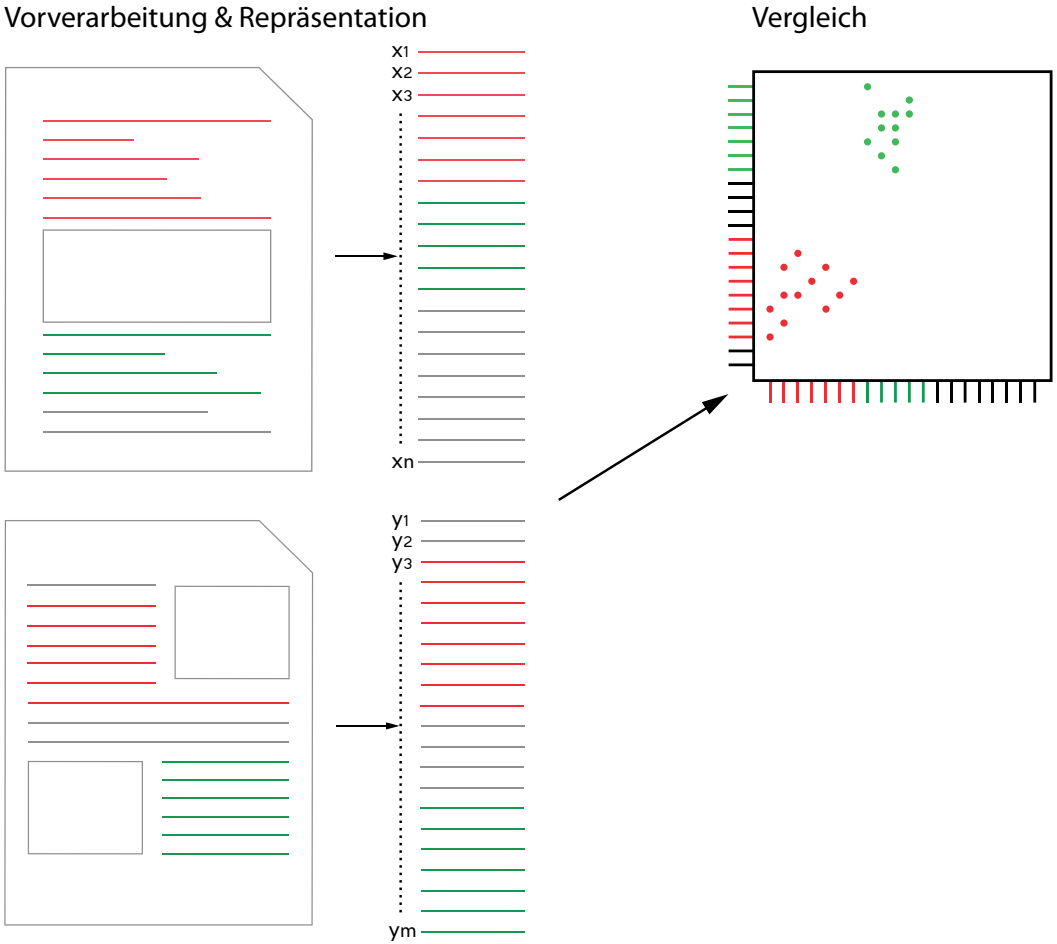
Verfahren zur Text-Reuse-Extraktion

Verarbeitungsschritte



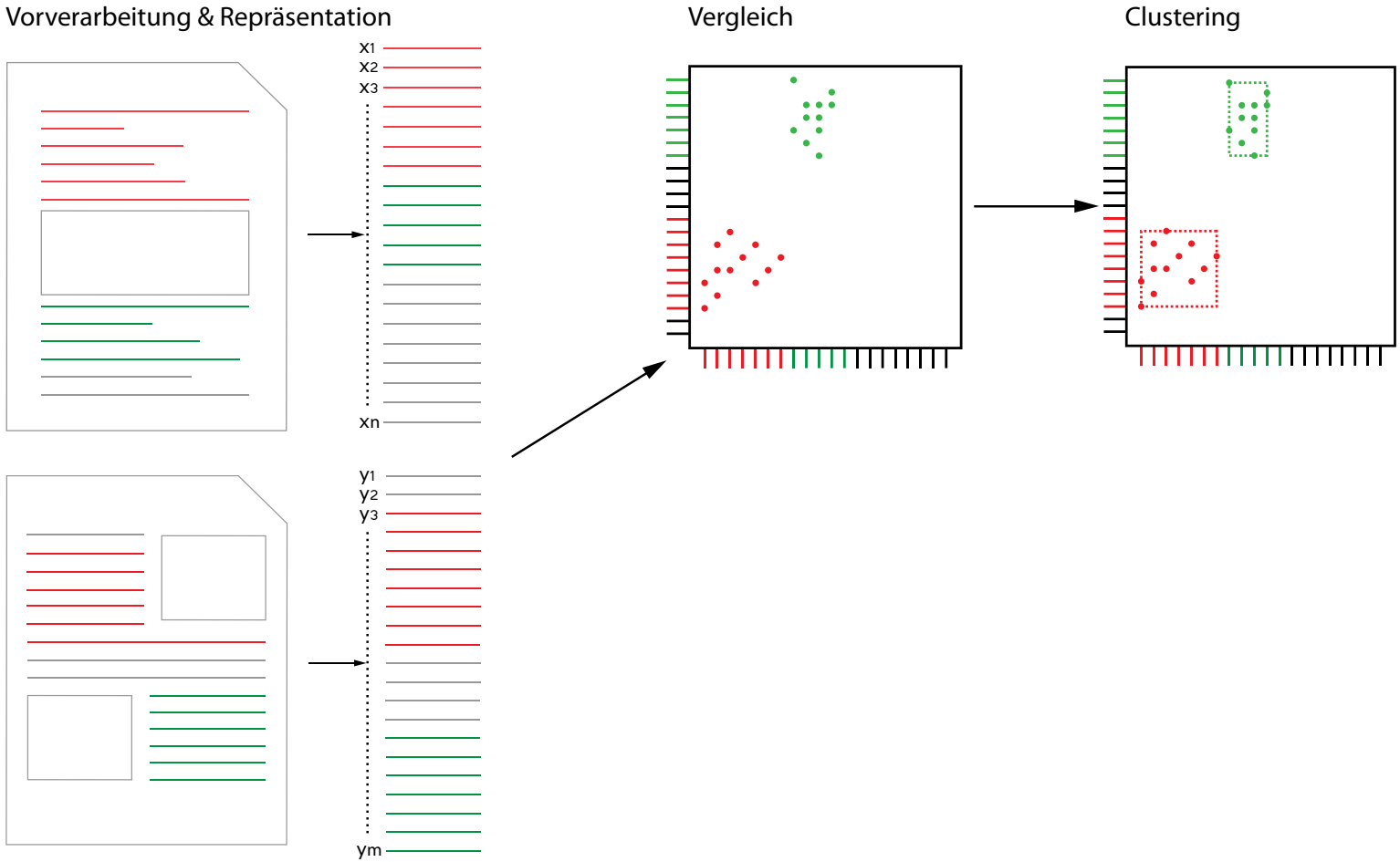
Verfahren zur Text-Reuse-Extraktion

Verarbeitungsschritte



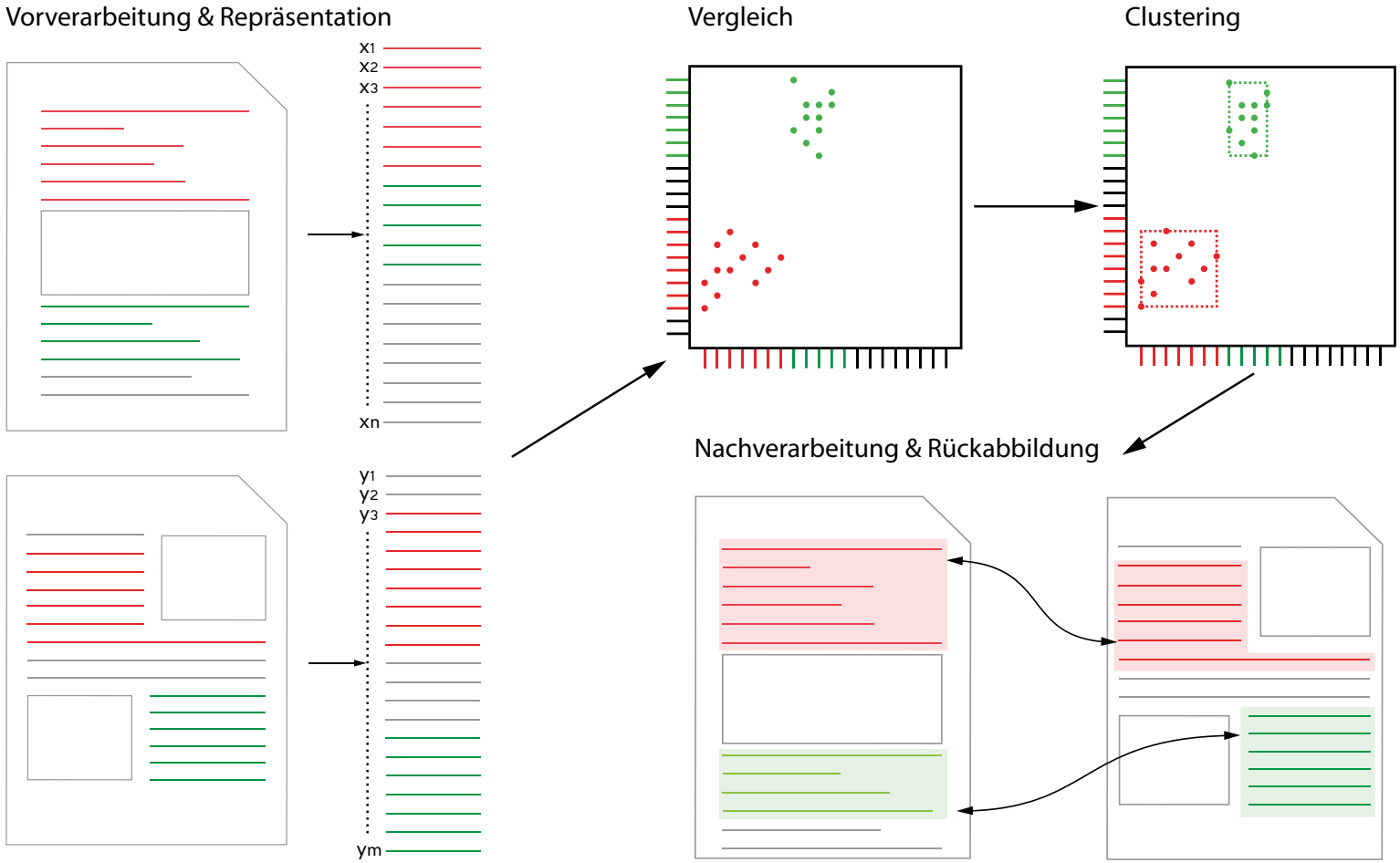
Verfahren zur Text-Reuse-Extraktion

Verarbeitungsschritte



Verfahren zur Text-Reuse-Extraktion

Verarbeitungsschritte



Verfahren zur Text-Reuse-Extraktion

Vorverarbeitung & Repräsentation

"Far far away, behind the word mountains, far from the countries
Vokalia and Consonantia, there live the blind texts."

Verfahren zur Text-Reuse-Extraktion

Vorverarbeitung & Repräsentation

"Far far away, behind the word mountains, far from the countries Vokalia and Consonantia, there live the blind texts."

- Tokenisierung

  (far, far, away, behind, the, word, mountains, far, from, the, countries, vokalia, and, consonantia, there, live, the, blind, texts)

- Entfernen der Stoppworte

  (word, mountains, countries, vokalia, consonantia, live, blind, texts)

- Normalisierung der Synonyme

  (word, mountain, country, vokalia, consonantia, live, blind, text)

Verfahren zur Text-Reuse-Extraktion

Vorverarbeitung & Repräsentation

"Far far away, behind the word mountains, far from the countries Vokalia and Consonantia, there live the blind texts."

- Stemming

  (word, mount, count, vokal, conso, live, blind, text)

- Wort-*N*-Gramm-Chunking

  (word, mount, count, vokal, conso),
  (mount, count, vokal, conso, live),
  (count, vokal, conso, live, blind),
  (vokal, conso, live, blind, text)

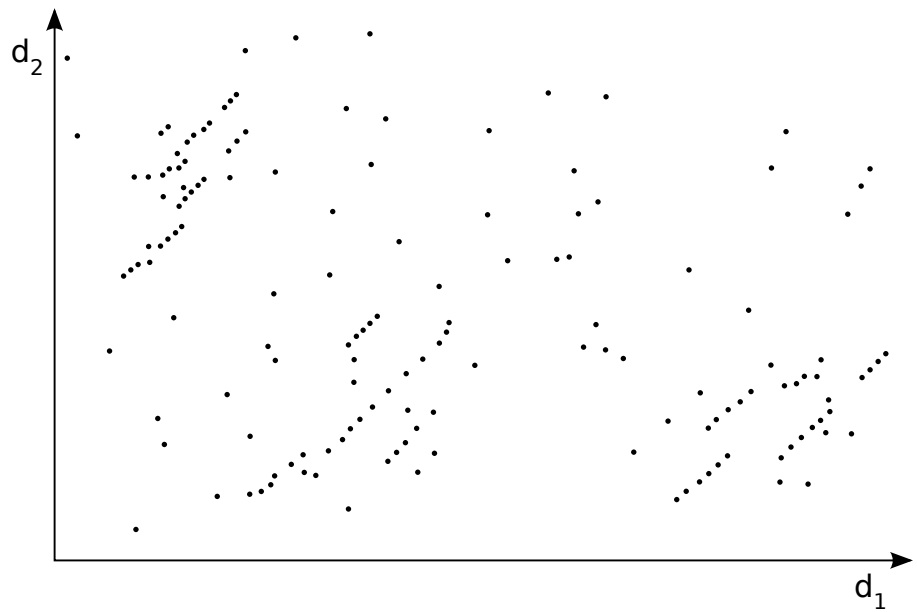
- Wort-*N*-Gramm-Sortierung

  (conso, count, mount, vokal, word),
  (conso, count, live, mount, vokal),
  (blind, conso, count, live, vokal),
  (blind, conso, live, text, vokal)

Verfahren zur Text-Reuse-Extraktion

Vergleich & Clustering

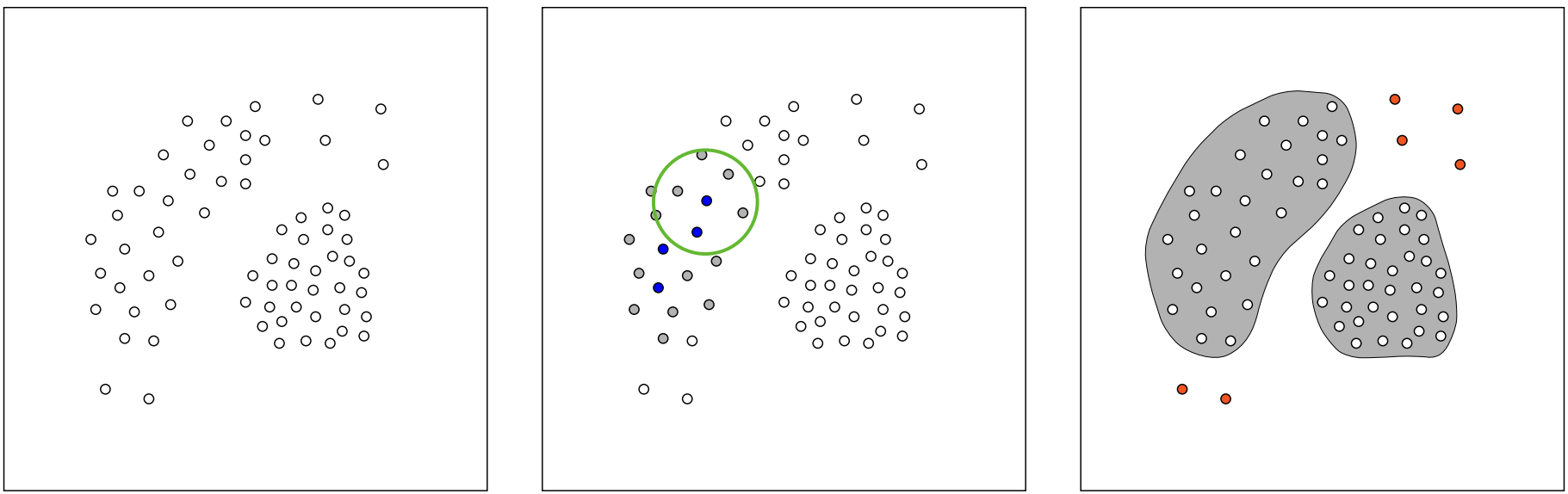
- Extraktion und Repräsentation identischer Fragmente.



Verfahren zur Text-Reuse-Extraktion

Vergleich & Clustering

- Exkurs: Clusteranalyse mit DBScan
  - Unüberwachte Gruppierung dichter Punktbereiche.
  - Automatische Strukturerkennung.

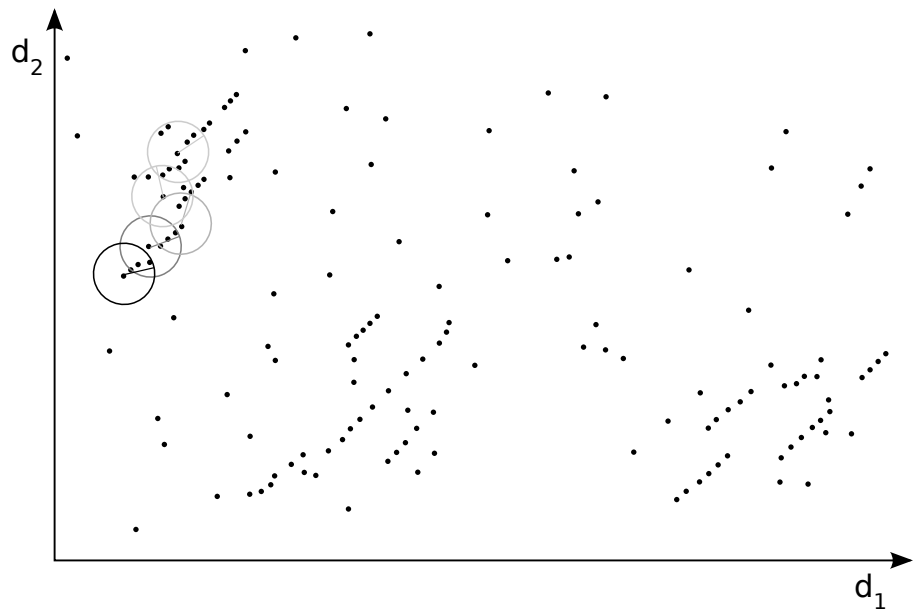


[Quelle: Stein (2011) - Vorlesung Cluster-Analyse]

Verfahren zur Text-Reuse-Extraktion

Vergleich & Clustering

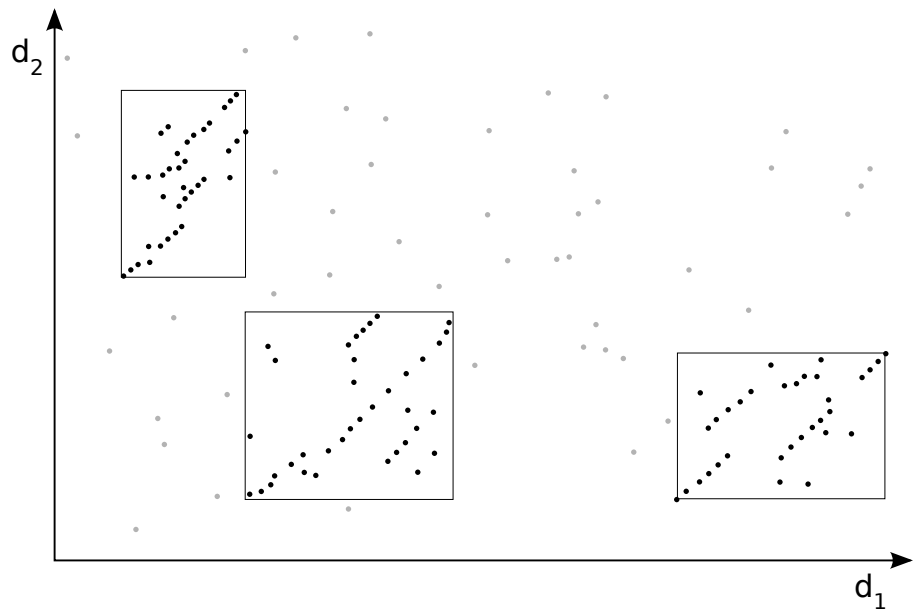
- Clusteranalyse lokaler Bereiche.
  - Dichtebasierte Clusteranalyse mit DBScan.
  - Analyse und Vereinigung erkannter Cluster.
  - Künstliche Fragmentierung.



Verfahren zur Text-Reuse-Extraktion

Vergleich & Clustering

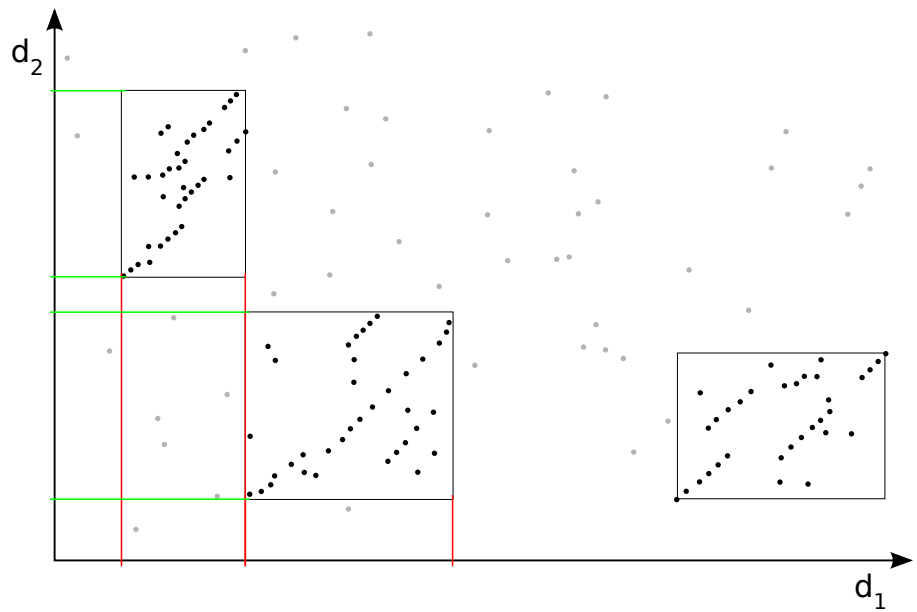
- Clusteranalyse lokaler Bereiche.
  - Dichtebasierte Clusteranalyse mit DBScan.
  - Analyse und Vereinigung erkannter Cluster.
  - Künstliche Fragmentierung.



Verfahren zur Text-Reuse-Extraktion

Vergleich & Clustering

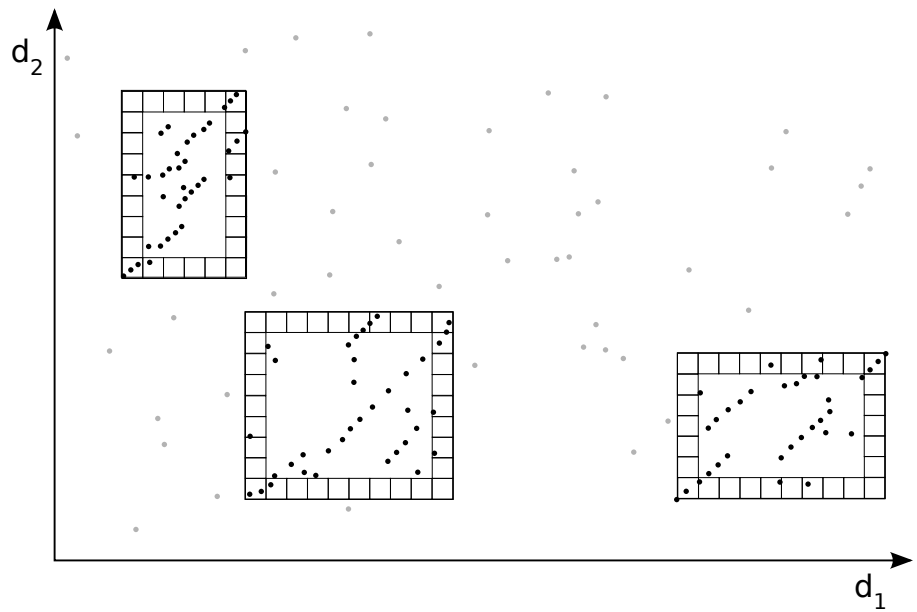
- Clusteranalyse lokaler Bereiche.
  - Dichtebasierte Clusteranalyse mit DBScan.
  - Analyse und Vereinigung erkannter Cluster.
  - Künstliche Fragmentierung.



Verfahren zur Text-Reuse-Extraktion

Vergleich & Clustering

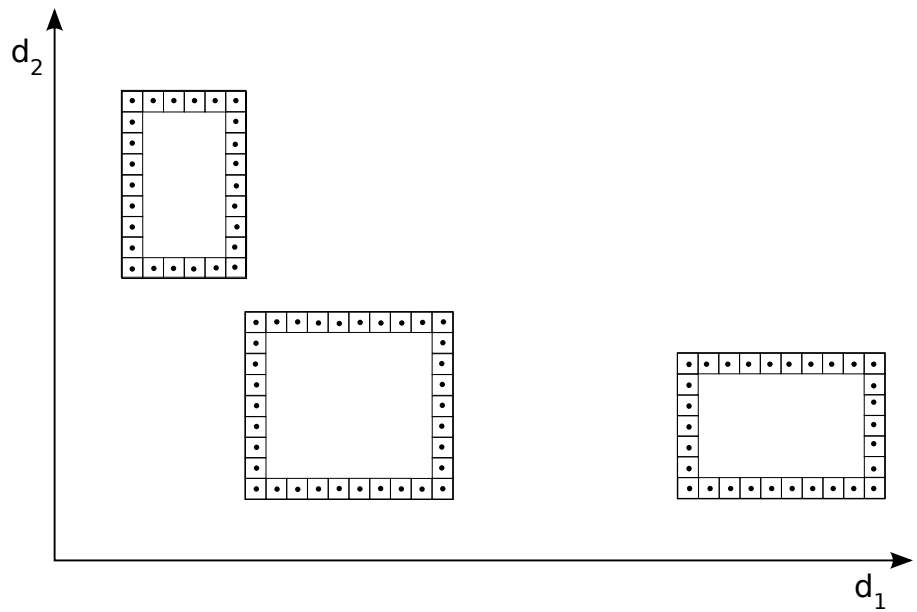
- Clusteranalyse lokaler Bereiche.
  - Dichtebasierte Clusteranalyse mit DBScan.
  - Analyse und Vereinigung erkannter Cluster.
  - Künstliche Fragmentierung.



Verfahren zur Text-Reuse-Extraktion

Vergleich & Clustering

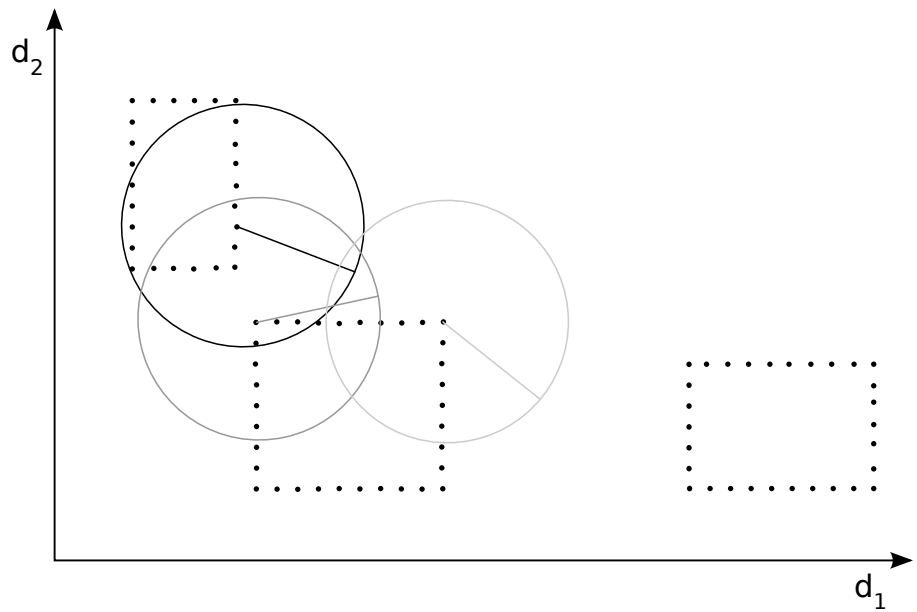
- Clusteranalyse lokaler Bereiche.
  - Dichtebasierte Clusteranalyse mit DBScan.
  - Analyse und Vereinigung erkannter Cluster.
  - Künstliche Fragmentierung.



Verfahren zur Text-Reuse-Extraktion

Vergleich & Clustering

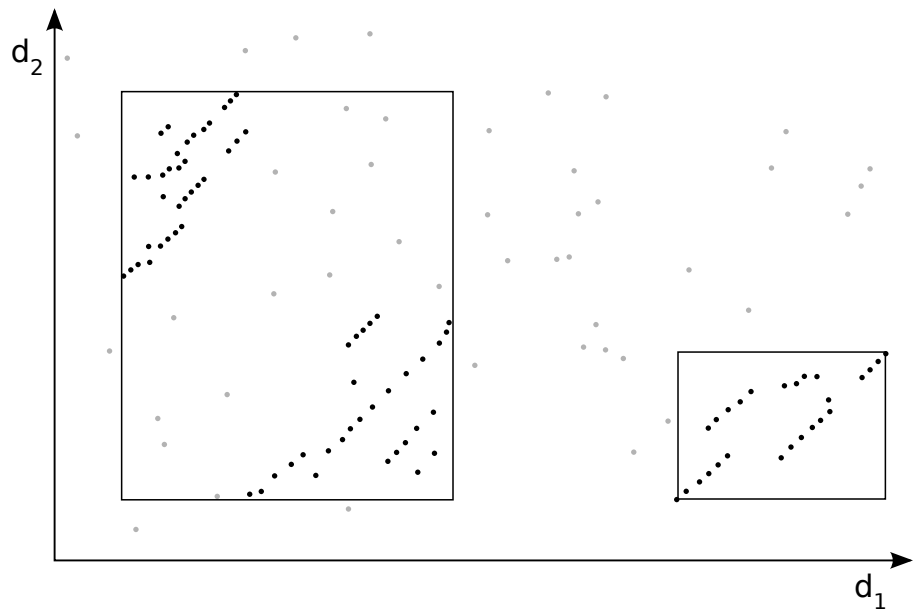
- Clusteranalyse globale Bereiche.
  - Dichtebasierte Clusteranalyse mit DBScan.
  - Analyse und Vereinigung erkannter Cluster.



Verfahren zur Text-Reuse-Extraktion

Vergleich & Clustering

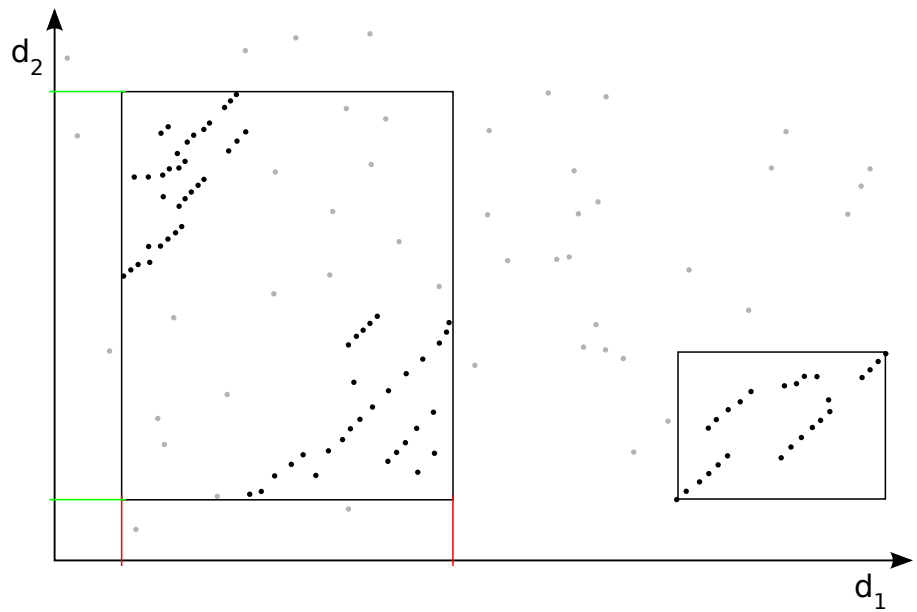
- Clusteranalyse globale Bereiche.
  - Dichtebasierte Clusteranalyse mit DBScan.
  - Analyse und Vereinigung erkannter Cluster.



Verfahren zur Text-Reuse-Extraktion

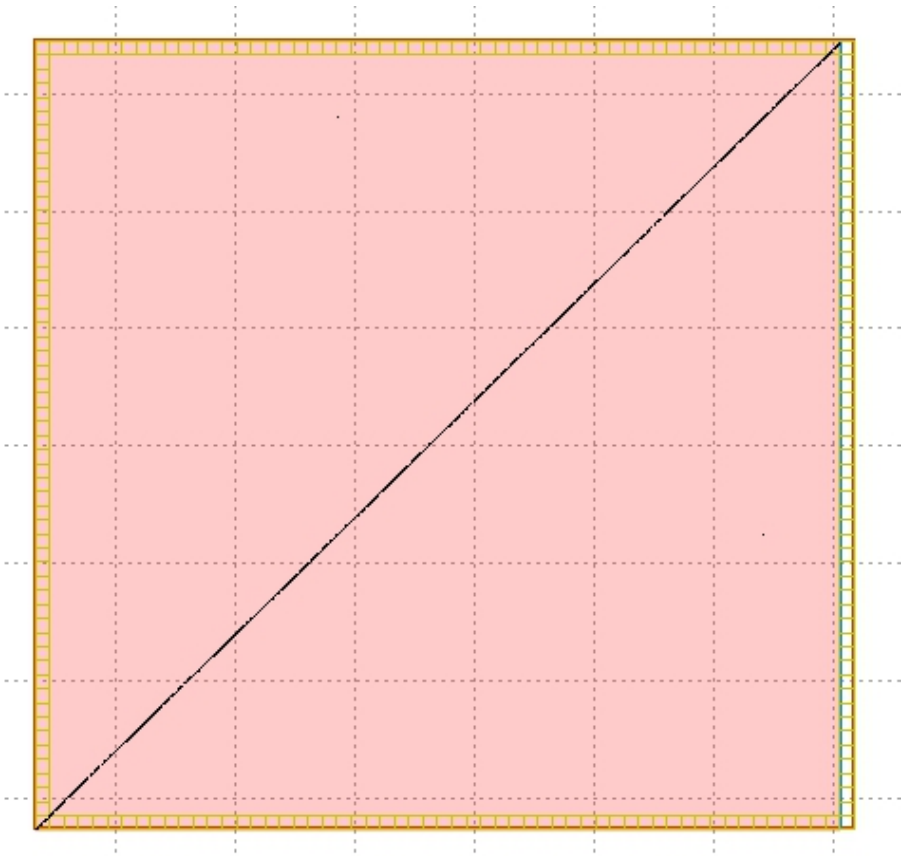
Vergleich & Clustering

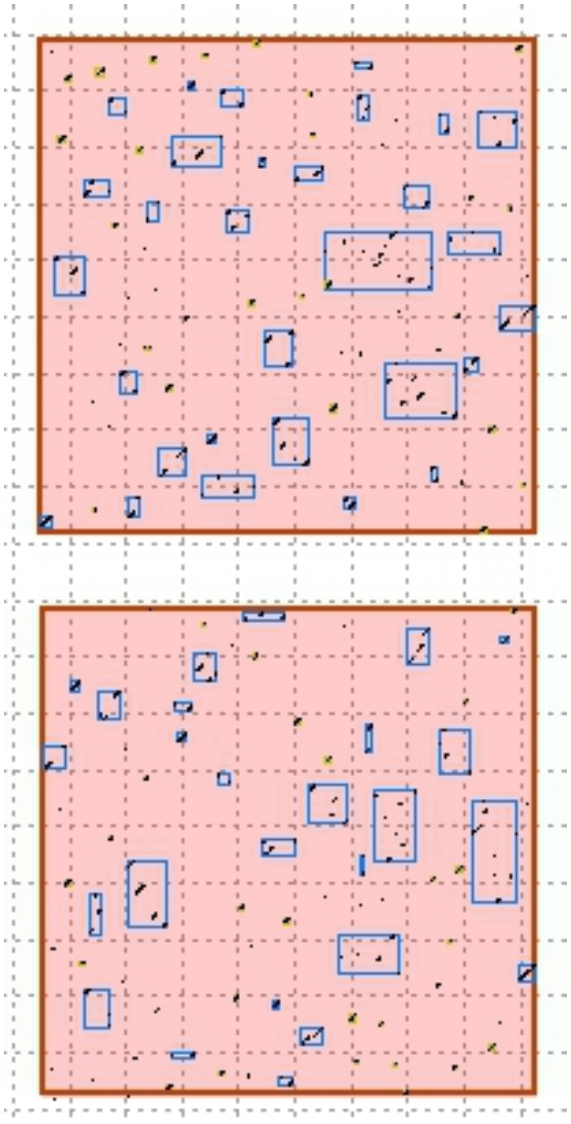
- Clusteranalyse globale Bereiche.
  - Dichtebasierte Clusteranalyse mit DBScan.
  - Analyse und Vereinigung erkannter Cluster.



Verfahren zur Text-Reuse-Extraktion

Beispiele





Überblick

Evaluierung

In dieser Arbeit

- Leistungsfähigkeit mit dem PAN-Framework untersucht.
- Großer Korpus mit Textwiederverwendung als Datengrundlage.
- Gütemaße zur Bewertung der Erkennungsleistung des Verfahrens.

PAN

- Jährlich stattfindender Wettbewerb zur Plagiaterkennung.
- Bislang 42 Detektoren in den Jahren 2009 - 2011 evaluiert.

Evaluierung

PAN-Framework

PAN-Framework

- Testumgebung zur ganzheitlichen Evaluierung von Detektoren.
- Der PAN plagiarism corpus (PAN-PC) bildet die Grundlage für die getesteten Textanalysen.
- Vordefinierte Gütemaße kombiniert im Effizienzwert Plagdet.

PAN plagiarism corpus 2010 (PAN-PC-10)

- 27073 Dokumente und 68558 künstliche Plagiatsfälle.
- 50% Quelldokumente.
- 50% verdächtige Dokumente von denen die Hälfte Plagiate enthält.

Evaluierung

PAN-Framework

PAN-Framework

- Testumgebung zur ganzheitlichen Evaluierung von Detektoren.
- Der PAN plagiarism corpus (PAN-PC) bildet die Grundlage für die getesteten Textanalysen.
- Vordefinierte Gütemaße kombiniert im Effizienzwert Plagdet.

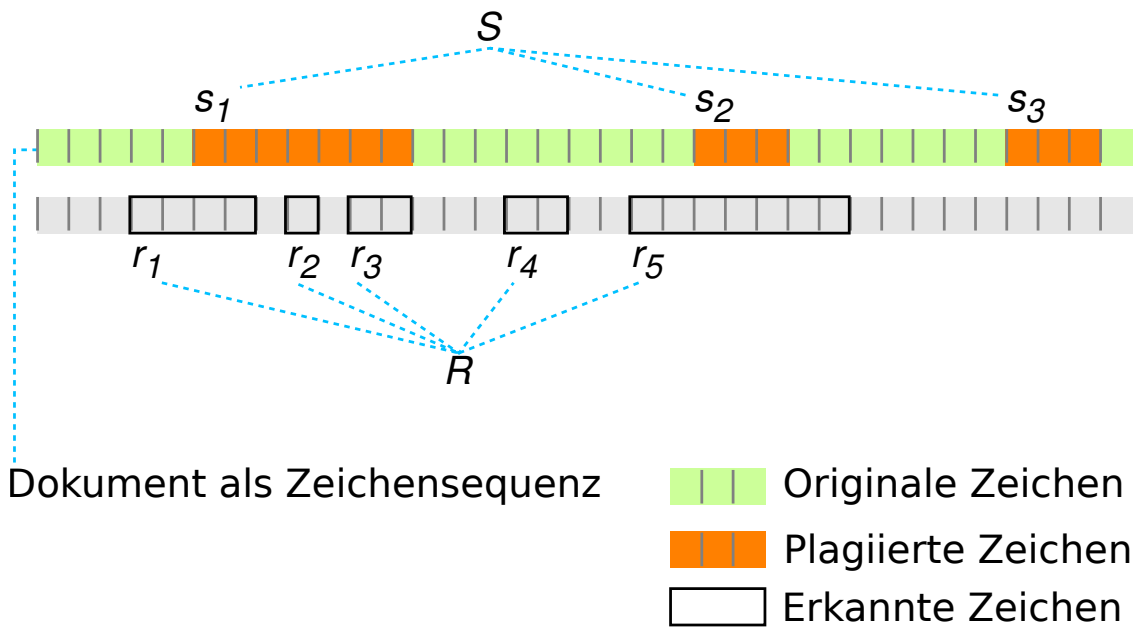
PAN plagiarism corpus 2010 (PAN-PC-10)

- 27073 Dokumente und 68558 künstliche Plagiatsfälle.
- 50% Quelldokumente.
- 50% verdächtige Dokumente von denen die Hälfte Plagiate enthält.

Evaluierung

PAN-Framework

- Recall misst die Vollständigkeit der Erkennungen.
- Precision misst die Genauigkeit der Erkennungen.
- Granularity misst die Häufigkeit der Erkennungen.



[Quelle: Potthast et al. (2009) - Overview of the 1st International Competition on Plagiarism Detection]

Evaluierung

PAN-Framework

- Plagdet verrechnet vorangegangene Maße zu einem vergleichbaren Effizienzwert eines Plagiatdetektors.

$$plagdet = \frac{F_1}{\log_2(1+granularity(S,R))}$$

F_1 Gewichtetes harmonisches Mittel aus *precision* und *recall*

Evaluierung

Analyse der Parameter

- 9 analysierte Parameter aus Vorverarbeitung und Datenanalyse.
- Subkorpus des PAN-PC-10 als Datengrundlage.
- Parameteranpassung anhand der Entwicklung des Plagdet.
- Durchgeführte Experimente:

Stoppwortentfernung	On/Off
Stemming	On/Off regel-/zeichenbasiert [2, 3 ,4, 5] Zeichen Präfixlänge
Wort-*N*-Grammlänge	[3, 4, 5, 6] Wörter
Wort-*N*-Gramm Sortierung	On/Off
DBScan Radius (lokal)	[100, 500, 1000, 1500, 2000] Zeichen
Clusterfilterung	[60, 70, 80, 90, 100, 110] Zeichen
DBScan Radius (global)	[2000, 2500, 3000, 3500, 4000] Zeichen

Evaluierung

Analyse der Parameter

- 9 analysierte Parameter aus Vorverarbeitung und Datenanalyse.
- Subkorpus des PAN-PC-10 als Datengrundlage.
- Parameteranpassung anhand der Entwicklung des Plagdet.
- Durchgeführte Experimente:

Stoppwortentfernung	On/Off
Stemming	On/Off regel-/zeichenbasiert [2, 3 ,4, 5] Zeichen Präfixlänge
Wort-*N*-Grammlänge	[3, 4, 5, 6] Wörter
Wort-*N*-Gramm Sortierung	On/Off
DBScan Radius (lokal)	[100, 500, 1000, 1500, 2000] Zeichen
Clusterfilterung	[60, 70, 80, 90, 100, 110] Zeichen
DBScan Radius (global)	[2000, 2500, 3000, 3500, 4000] Zeichen

Evaluierung

Experimente mit dem PAN-PC

PAN-PC-10 (18 Teilnehmer)

Rank	Plagdet	Recall	Precision	Granularity	Team
1	0.7971	0.6917	0.9414	1.0006	Kasprzak,Brandejs
	0.7127	0.6895	0.8800	1.1211	Kümmel, Potthast
2	0.7090	0.6299	0.9055	1.0675	Zou, Long, Ling
3	0.6948	0.7057	0.8417	1.1508	Muhr, Kern, Zechner, Granitzer
4	0.6209	0.4808	0.9085	1.0177	Grozea, Popescu

PAN-PC-11 (9 Teilnehmer)

Rank	Plagdet	Recall	Precision	Granularity	Team
1	0.5563	0.3965	0.9368	1.0022	Grman, Ravas
2	0.4153	0.3376	0.8119	1.2167	Grozea, Popescu
3	0.3468	0.2257	0.9116	1.0611	Oberreuter
	0.2893	0.2032	0.5811	1.0576	Kümmel, Potthast
4	0.2467	0.1500	0.7106	1.0058	Gillam

Überblick

Zusammenfassung

In dieser Arbeit

- Entwicklung eines Verfahrens zur Extraktion von Text-Reuse.
- Einsatz von Algorithmen des IR und verwandter Forschungsgebiete.
- Analyse des Verfahrens und der Parameter mit dem PAN-Framework.

Zukünftige Problemstellungen

- Weiterführende Ausforschung des Parameterraums.
- Adaptiver DBScan-Algorithmus abhängig von den Eingabedaten.
- Konzeptionelle Weiterentwicklung mit Hinblick auf Paraphrasen.

Zusammenfassung

In dieser Arbeit

- Entwicklung eines Verfahrens zur Extraktion von Text-Reuse.
- Einsatz von Algorithmen des IR und verwandter Forschungsgebiete.
- Analyse des Verfahrens und der Parameter mit dem PAN-Framework.

Zukünftige Problemstellungen

- Weiterführende Ausforschung des Parameterraums.
- Adaptiver DBScan-Algorithmus abhängig von den Eingabedaten.
- Konzeptionelle Weiterentwicklung mit Hinblick auf Paraphrasen.

Vielen Dank!

Literaturverzeichnis I


P. Clough.
Measuring text reuse in a journalistic domain.
In *Conference Proceedings of the 4th Annual CLUK Colloquium, University of Sheffield, Sheffield, UK*, 2001.


P. McNamee, C. K. Nicholas, and J. Mayfield.
Don't have a stemmer?: be un+concern+ed.
In *SIGIR*, pages 813–814, 2008.


Metzler, D. and Bernstein, Y. and Croft, W.B. and Moffat, A. and Zobel, J.
Similarity measures for tracking information flow.
In *Proceedings of the 14th ACM international conference on Information and knowledge management*, page 524. ACM, 2005.

Literaturverzeichnis II

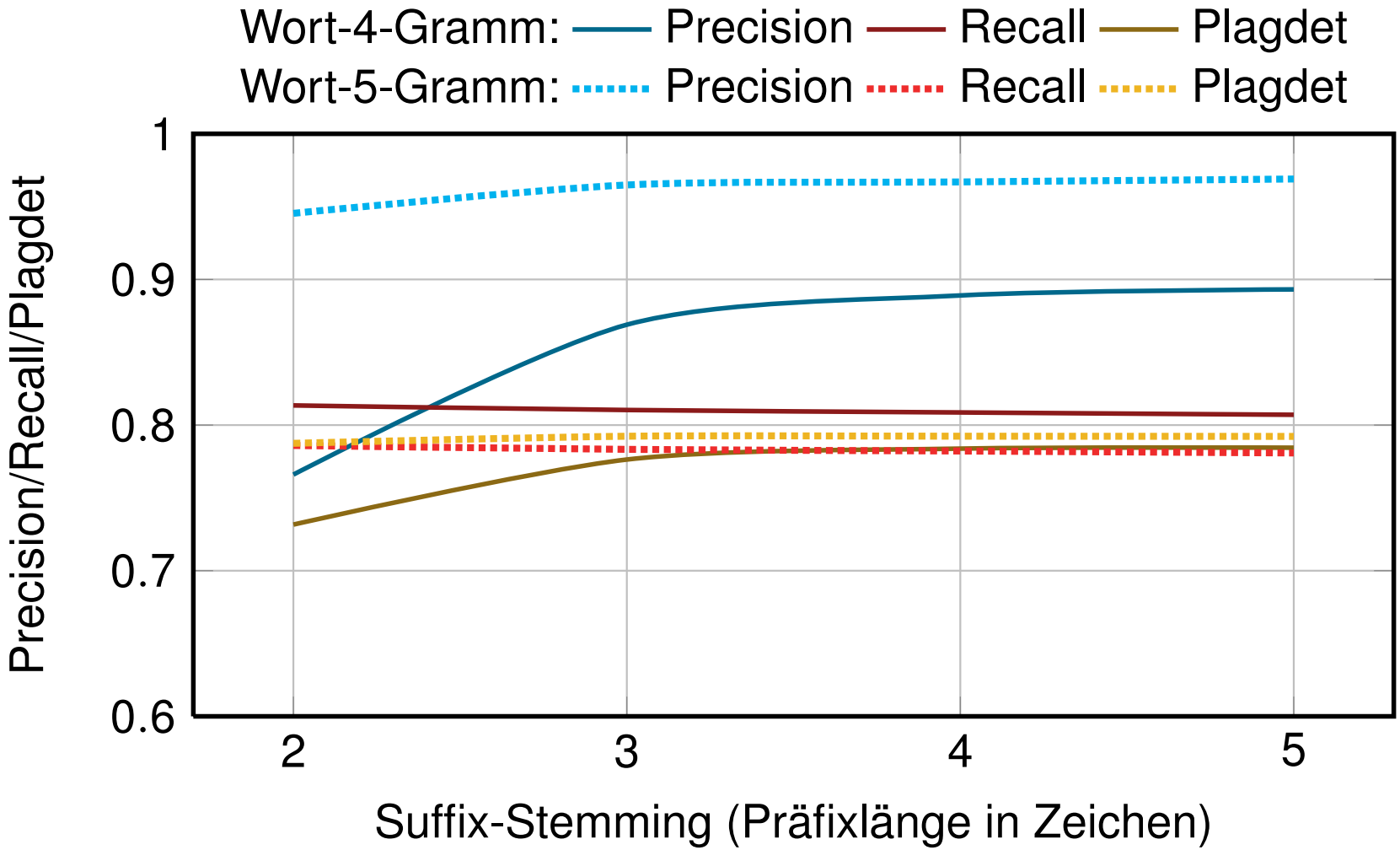

N.N.
Appendix: Common Evaluation Measures.
In Ellen M. Voorhees and Lori P. Buckland, editor, *19th International Text Retrieval Conference (TREC 10)*. National Institute of Standards and Technology (NIST), 2010.


M. Potthast, B. Stein, A. Eiselt, A. Barrón-Cedeño, and P. Rosso.
Overview of the 1st International Competition on Plagiarism Detection.
In B. Stein, P. Rosso, E. Stamatatos, M. Koppel, and E. Agirre, editors, *SEPLN 09 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)*, pages 1–9. CEUR-WS.org, Sept. 2009.


B. Stein.
Lectures on cluster analysis.
Slides at Bauhaus-Universität Weimar, 28 Mar. 2011.

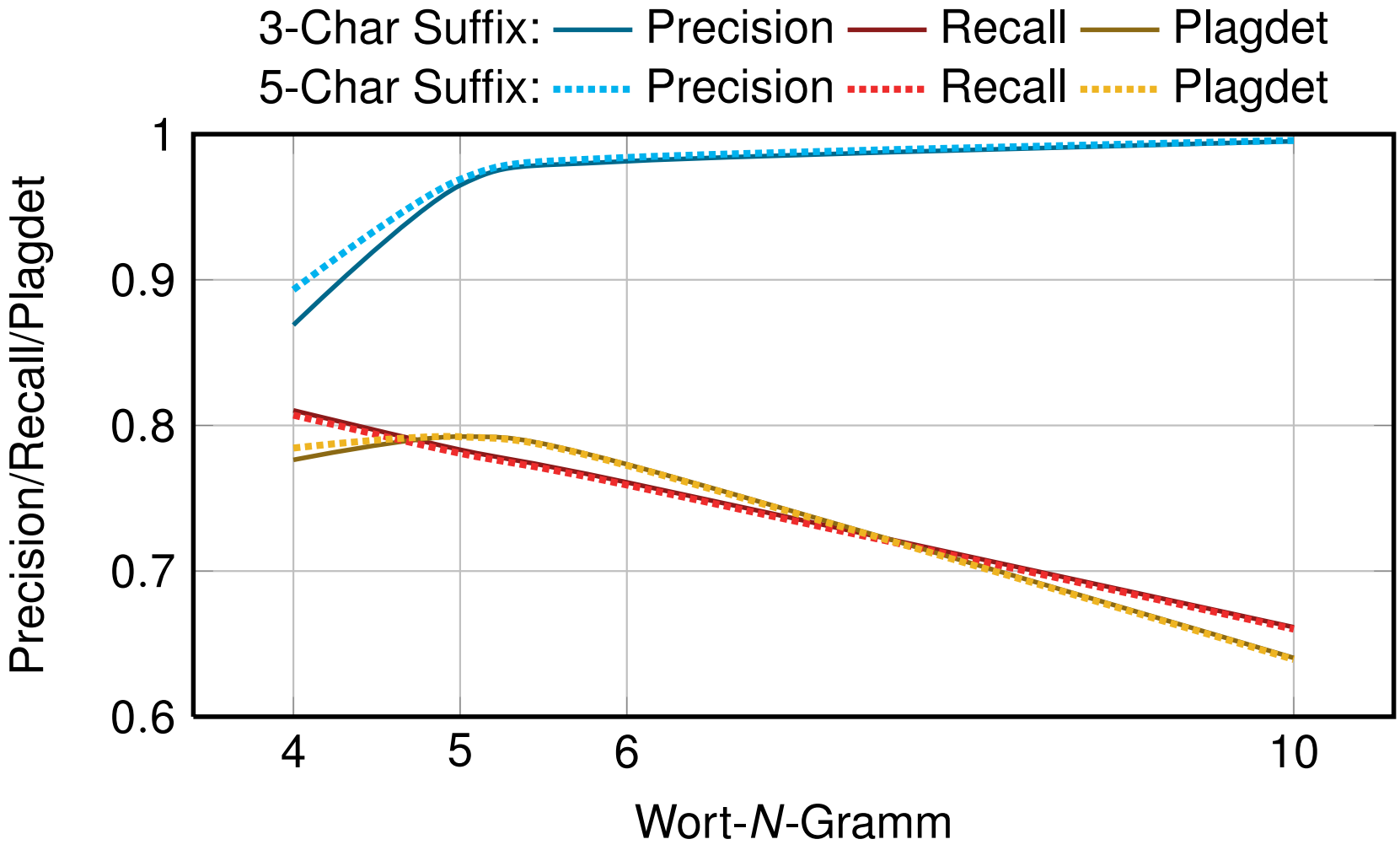
Experimente

- Suffix-Stemming



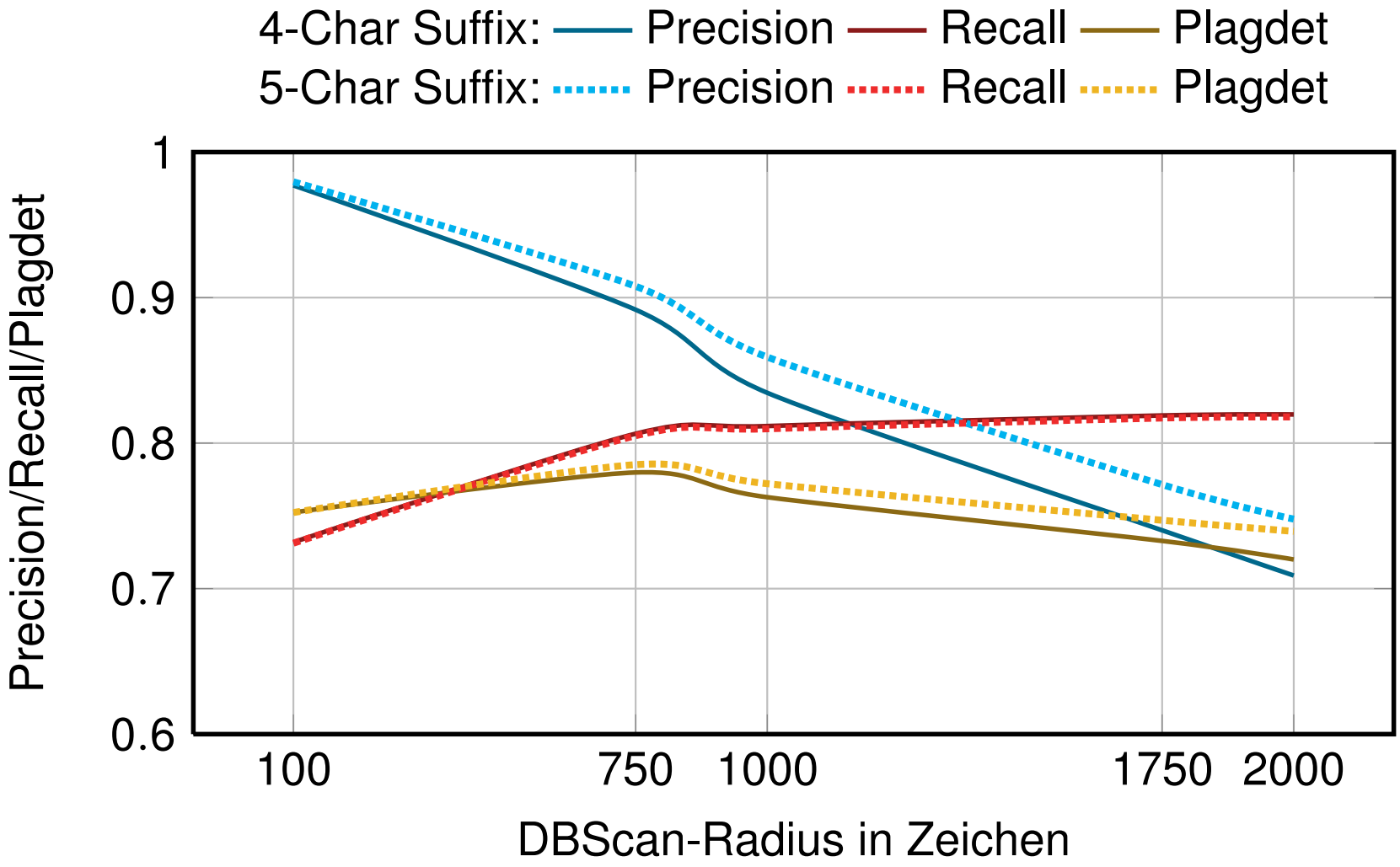
Experimente

- Chunking-Länge



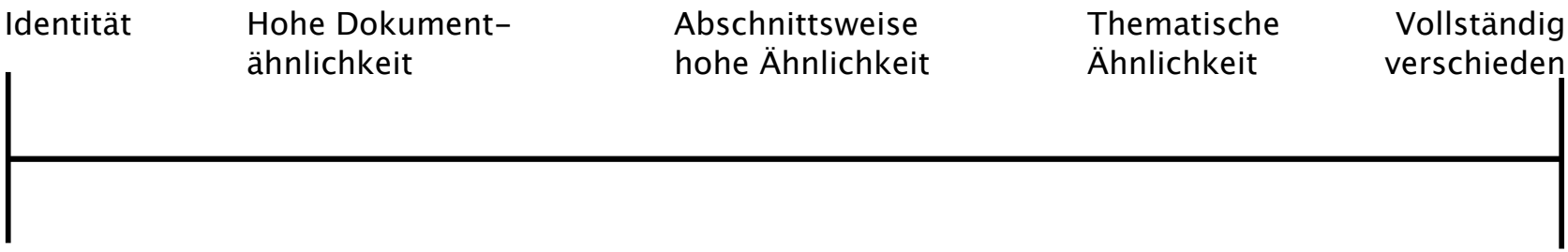
Experimente

- DBscan-Radius



Extra

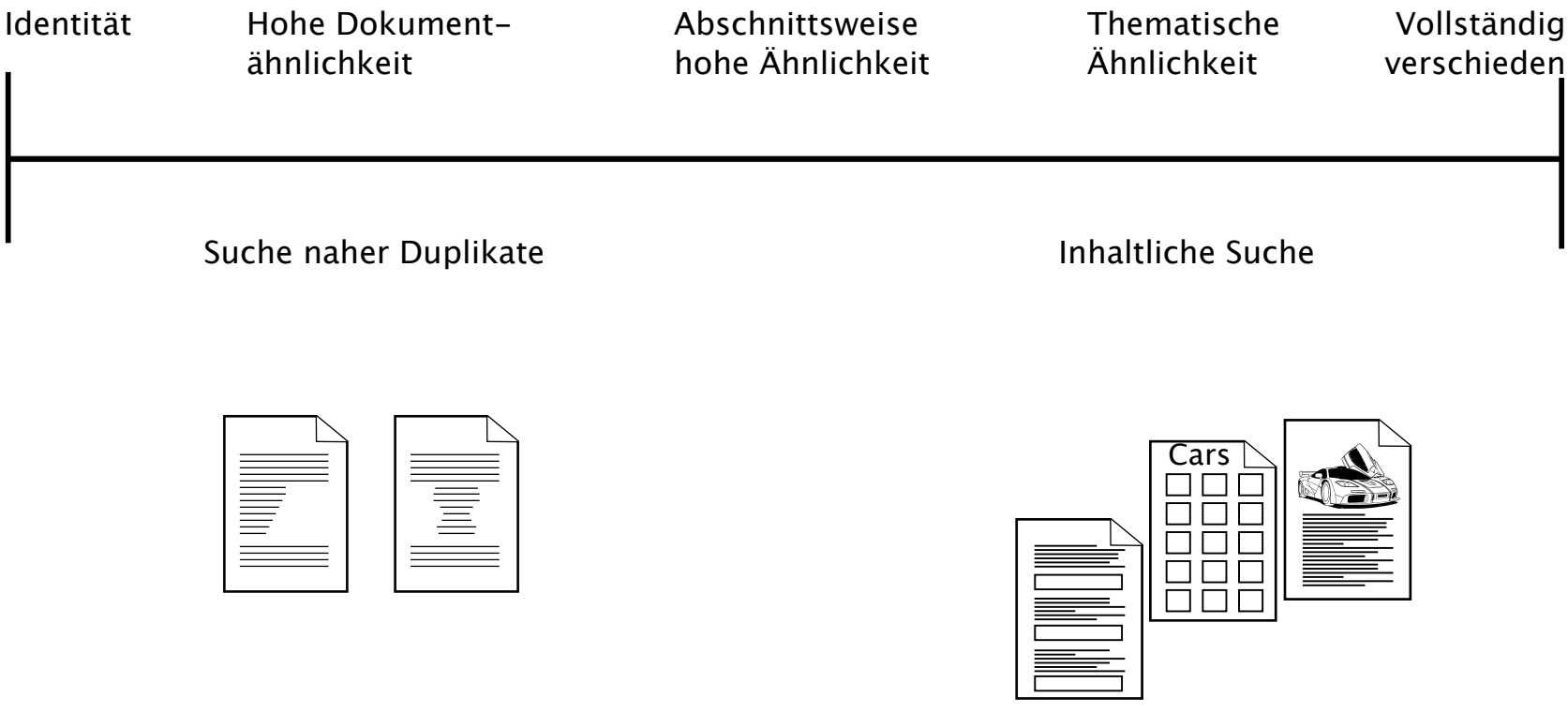
Das Ähnlichkeitsspektrum



based on [Clough et al. (2001), Metzler et al. (2005)]

Extra

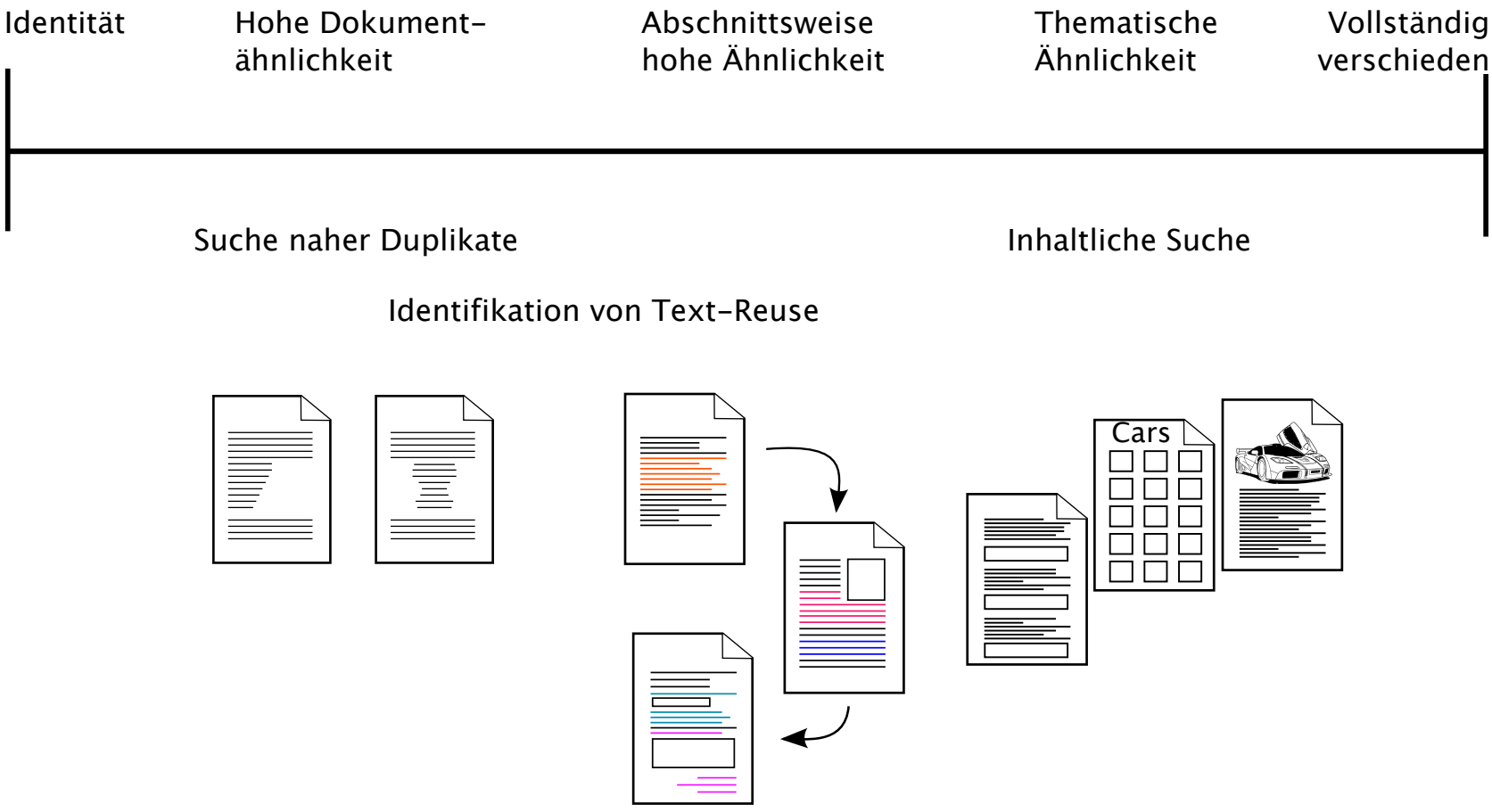
Das Ähnlichkeitsspektrum



based on [Clough et al. (2001), Metzler et al. (2005)]

Extra

Das Ähnlichkeitsspektrum



based on [Clough et al. (2001), Metzler et al. (2005)]