

Plagiatserkennung

Motivation

- ▶ Akademische Plagiate (Guttenberg, Schavan, von der Leyen) [3]
- ▶ Finden von Plagiaten schwierig und langwierig
- ▶ Aktuelle Software nur teilweise nützlich bis nutzlos
- ▶ Erleichtern der Plagiatserkennung durch Software

Zielsetzung

- ▶ Software zur effizienten Plagiatserkennung
- ▶ Methoden aus Stringology und Information Retrieval
- ▶ Einlesen von Textdokumenten und Abschlussarbeiten

Methodik

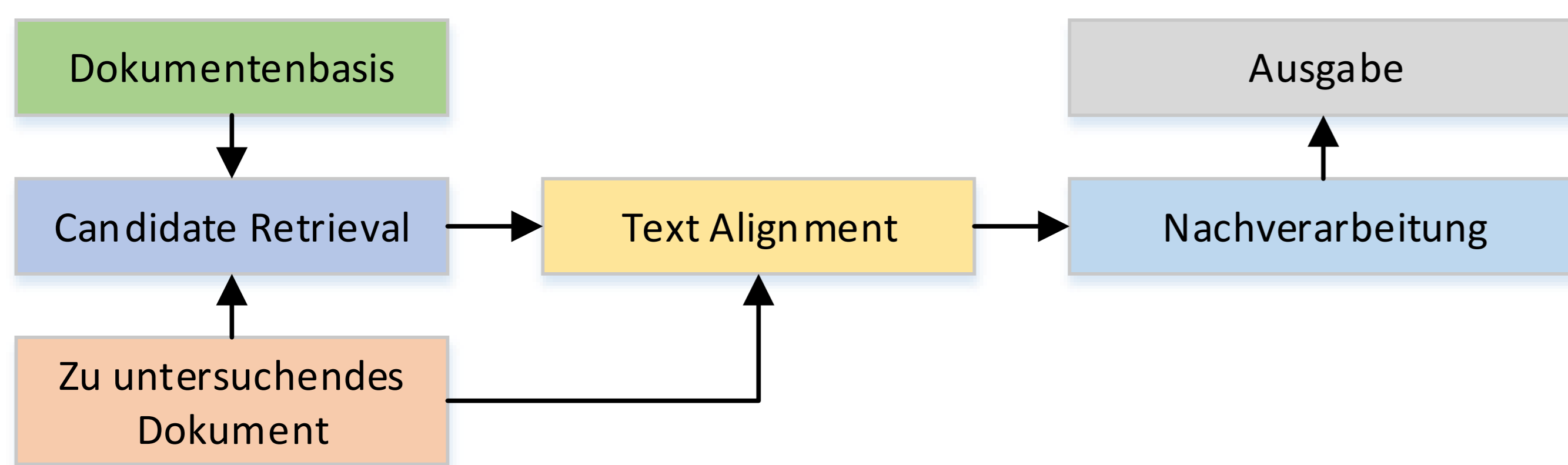


Abbildung: Vorgehensweise der Software

Schritte der Software

- ▶ Vorverarbeitung der Dokumente
- ▶ Aufbau einer Dokumentenbasis
- ▶ Kandidatendokumente mit Methoden des Information Retrieval suchen
- ▶ Ähnliche Stellen mit Methoden der Stringology finden
- ▶ Nachverarbeitung und Ausgabe möglicherweise plagiierter Stellen

Vorverarbeitung

- ▶ *Tokenization*: Aufteilen nach Satzzeichen
- ▶ *Stemming*: Reduzierung von Wörtern auf den Stamm
- ▶ *Stop Word Removal*: Wörter mit wenig Informationsgehalt entfernen
- ▶ *Normalisierung*: Entfernung von Sonderzeichen
- ▶ *Integer-Alphabet*: Wörter werden auf Zahlen abgebildet

Candidate Retrieval

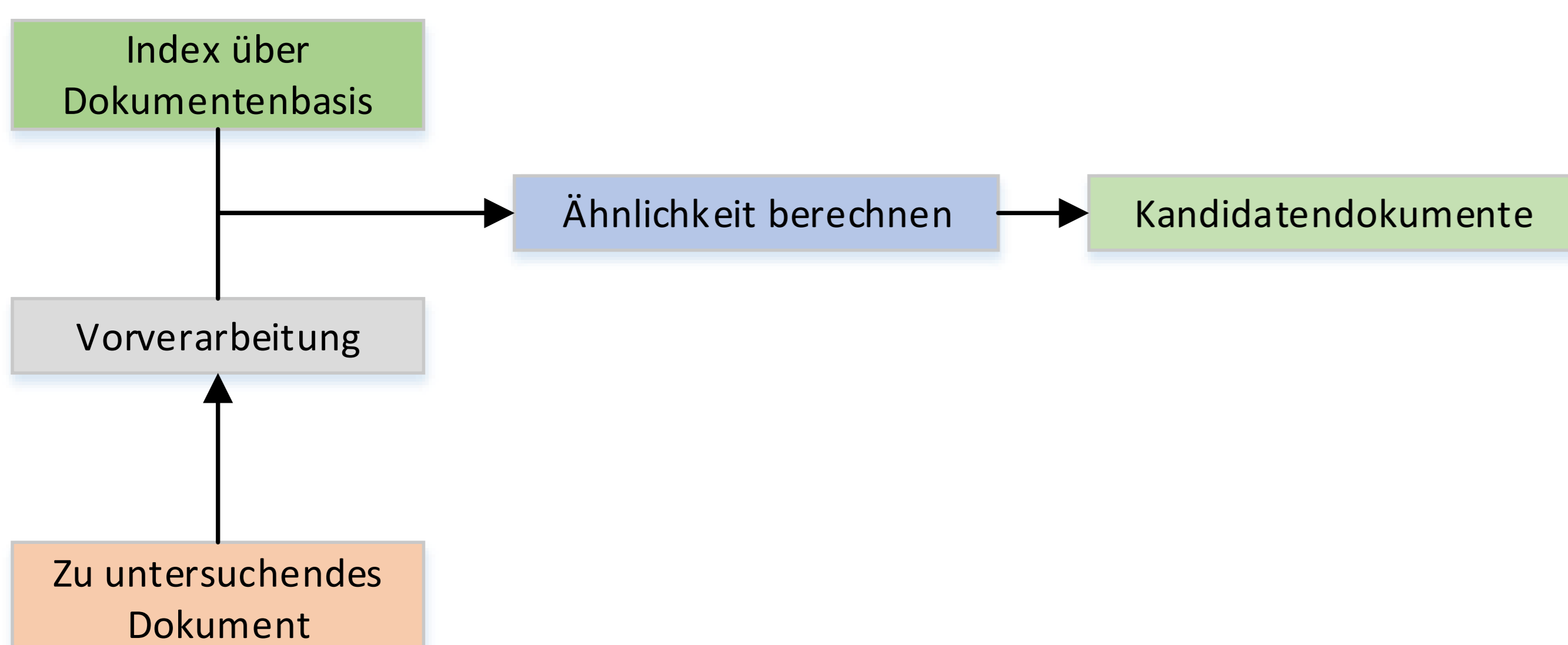


Abbildung: Vorgehensweise bei Candidate Retrieval

- ▶ Eingabe ist eine Dokumentenbasis und ein zu untersuchendes Dokument (Anfrage)
- ▶ Vollständige Suche in der Dokumentenbasis nach Dokumenten, die ähnlich zur Anfrage sind
- ▶ Kandidatendokumente sind Teilmenge der Dokumentenbasis

Ähnlichkeit von Dokumenten

- ▶ Repräsentation der Dokumente als *tf-idf-Vektor* [2]
 - ▶ *tf*: Term Frequency - Häufigkeit des Wortes im Dokument
 - ▶ *idf*: Inverse Document Frequency - Größe der Dokumentenbasis geteilt durch Anzahl der Dokumente, die dieses Wort enthalten
 - ▶ *tf-idf-Vektor*: Vektor über allen Wörtern, die im Dokument vorkommen. Die *i*-te Zeile repräsentiert den $tf \cdot idf$ Wert des Wortes *i* im Dokument *d*.
- ▶ Ähnlichkeit zweier Dokument-Vektoren *d* und *d'* mit Kosinus-Ähnlichkeit berechnen: $\text{sim}(d, d') := \cos(\angle(d, d'))$
- ▶ Ausgabe von Dokumenten, deren Ähnlichkeit einen Schwellwert überschreitet

Text Alignment mit Suffixarrays

$k = 1\ 6\ 7\ 9\ 7\ 4\ 3\ 5\ 4\ 9\ 3$
 $t_1 = 3\ 4\ 5\ 2\ 7\ 1\ 6\ 7\ 9\ 1\ 6$
 $t_2 = 9\ 3\ 5\ 1\ 2\ 3\ 5\ 4\ 7\ 8\ 1$
 $T = 1\ 6\ 7\ 9\ 7\ 4\ 3\ 5\ 4\ 9\ 3\ 3\ 4\ 5\ 2\ 7\ 1\ 6\ 7\ 9\ 1\ 6\ 9\ 3\ 5\ 1\ 2\ 3\ 5\ 4\ 7\ 8\ 1$

i	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	...
SA	33	26	17	01	21	27	15	11	12	24	28	07	06	13	30	09	25	
LCP	-	1	1	4	2	0	1	0	1	1	2	3	0	1	1	1	0	

Abbildung: Beispiel eines Suffixarrays (SA)

- ▶ *k*: Text des verdächtigen Dokuments
- ▶ t_1, t_2 : Text der Kandidatendokumente
- ▶ Grau unterlegt: Enhanced Suffix Array [1]
 - ▶ SA: Suffixarray: Lexikographisch sortierte Wortketten
 - ▶ LCP: Longest Common Prefix Array
- ▶ Rot markiert: Fundstelle

Nachbearbeitung

- ▶ *Merge*: Vereinigung sich überlappender Fundstellen
- ▶ *Delete*: Häufig vorkommende Formulierungen entfernen
- ▶ *Dominate*: Gleiche Fundstellen in verschiedenen Kandidatendokumenten beseitigen

Ausgabe

Compare Splag Results SELECT NEW

<p>Select candidate: 2189</p> <p>Title: C-reaktives Protein als diagnostischer Parameter zur Erfassung eines Amnio</p> <p>Author: Dr. Ursula Gertrud von der Leyen</p> <p>Während eines bakteriellen Infektstimulus werden die zirkulierenden Monozyten oder das sesshafte Makrophagensystem durch mikrobielle Toxine zur Freisetzung einer Vielfalt von Mediatoren - Interleukin-1, Tumor Necrosis Factor, Interleukin-6 etc. - gebracht. Diese wirken unabhängig voneinander einerseits als Stimulatoren des zerebralen Hypothalamischen Temperaturzentrums, und andererseits als Akut-Phase-Protein-Induktoren in den Hepatozyten. Fieber und CRP-Erhöhung sind also Folge einer parallel ablaufenden Reaktion (EDGARD [sic] et al. 1989, CAMBAU 1989). Die schon länger bekannte (McCARTHY et al. 1978, PELTOLA 1982, WHICHER et al. 1985) und erneut bestätigte (KERTULLA [sic] et al. 1987) Tatsache, daß ein Virusinfekt klassischerweise zwar mit Fieber, aber ohne oder nur mit geringem CRP-Anstieg einhergeht, findet heute ebenfalls ihre pathophysiologische Erklärung: das von "reaktiven" Lymphozyten auf viralen Stimulus freigesetzte Interferon stimuliert unabhängig und ohne Induktion der Akut-Phase-Mediatoren direkt das hypothalamische Temperaturzentrum. Somit liegt einem Status febrilis ohne oder mit nur geringer CRP-Erhöhung (4-5 mg/l) mit großer Wahrscheinlichkeit eine virale Genese zugrunde. (FEHR 1988).</p>	<p>Title: Die klinische Bedeutung des CRP-(C-reaktiven Protein-)Monitoring</p> <p>Author: Jörg Fehr</p> <p>Anlässlich eines bakteriellen Infektstimulus werden die zirkulierenden Monozyten oder das sesshafte Makrophagensystem zum Beispiel durch mikrobielle (Endo-)Toxine zur Freisetzung der je unabhängig einerseits als Stimulatoren des zerebralen Hypothalamischen Temperaturzentrums, und andererseits als Akutphasenproteininduktoren in den Hepatozyten wirkenden zwei Monokine IL-1 [7, 8] und Tumor Necrosis Factor (TNF [9, 10]) gebracht. Fieber und Anstieg des CRP sind also Folge einer parallel ablaufenden direkten Zielorgan(ZNS + Hepatozyt)-Stimulation durch diese zwei Monozyten/Makrophagen-Produkte. Amplifiziert werden kann diese Reaktion dadurch, dass [...] Die schon länger bekannte [11-13] und erneut bestätigte [14] Tatsache, dass ein Virusinfekt klassischerweise zwar mit Fieber, aber ohne oder nur mit geringem CRP-Anstieg einhergeht, findet heute ebenfalls ihre pathophysiologische Erklärung, indem das von "reaktiven" Lymphozyten auf viralen Stimulus freigesetzte Interferon (v. a. Interferon-α) experimentell belegbar [15] unabhängig und ohne Induktion von IL-1 direkt das hypothalamische Temperaturzentrum stimuliert. Somit liegt einem Status febrilis ohne oder mit nur geringer CRP-Erhöhung (4 bis 5 mg/dl) mit grosser Wahrscheinlichkeit eine virale Genese zugrunde.</p>
---	---

Abbildung: Beispielausgabe des Programms

Referenzen

- [1] MANBER, Udi ; MYERS, Gene:
Suffix arrays: a new method for on-line string searches.
In: *SIAM Journal on Computing* 22 (1993), Nr. 5, S. 935-948
- [2] SPARCK JONES, Karen:
A statistical interpretation of term specificity and its application in retrieval.
In: *Journal of documentation* 28 (1972), Nr. 1, S. 11-21
- [3] WEBER-WULFF, Debora:
False feathers: a perspective on academic plagiarism.
Springer Science & Business, 2014