Near-Duplicates im ClueWeb12

Eine Pilotstudie zur Übertragbarkeit von Forschungsergebnissen

Jan Graßegger
Bauhaus-Universität Weimar

25.09.2014

Near-Duplicates

Freiwilligentag in Weimar: Ein Tag mit 680 Stunden

23.09.2014 - 11:00 Uhr

Wenn 170 Leute vier Stunden lang ranklotzen, schaffen sie viel mehr als ein Arbeiter in 680 Stunden. Gemeinsamkeit macht stärker, deshalb sind die Einsätze an den Weimarer Freiwilligentagen nicht mit Geld aufzuwiegen.

Bäumchen, rüttel dich: Auf der Streuobstwiese bei Gaberndorf hängt kein Apfel mehr am Baum, nachdem die Grüne Liga mit zehn Helfern alles erntete, was sich zu Saft verarbeiten lässt. Foto: Sabine Brandt

Weimar. "Wir haben alles geschafft, was wir uns vorgenommen hatten", zieht Stefanie Lachmann von der Ehrenamtsagentur zufrieden Bilanz unter Weimars größten Subbotnik, der am zurückliegenden Samstag ausgerufen worden war. Seither sind der Lebenshilfeladen um einen Bilderständer und die Grüne Liga um rund zwei Tonnen Obst aus Gaberndorf für die Saftpresse reicher und das Schlachthofviertel um einige Kilo Müll

ZUM THEMA

Rekord an guten Taten beim Freiwilligentag in Jena

Rekordbeteiligung beim 10. Jenaer Freiwilligentag: Mehr als 320 Freiwillige kümmerten sich an 34 Ein... mehr

Freiwilligentag in Weimar: Ein Tag mit 680 Stunden

Erfurter Freiwilligentag für das Gemeinwohl

Kindertag mit Spaß und Wünschen in Gera

Erster Freiwilligentag in Eisenach

Bilder des Tages im Monat September

Erster Thüringer Freiwilligentag im Landkreis Nordhausen

Thüringer Freiwilligentag im Landkreis Nordhausen

Im Nachbarschaftszentrum in Eisenach wird beraten und auch gemeinsam gekocht

Helfer für Freiwilligentag am 20. September in Gera gesucht

Freiwilligentag in Weimar: Ein Tag mit 680 Stunden

23.09.2014 - 11:00 Uhr

Wenn 170 Leute vier Stunden lang ranklotzen, schaffen sie viel mehr als ein Arbeiter in 680 Stunden. Gemeinsamkeit macht stärker, deshalb sind die Einsätze an den Weimarer Freiwilligentagen nicht mit Geld aufzuwiegen.

Bäumchen, rüttel dich: Auf der Streuobstwiese bei Gaberndorf hängt kein Apfel mehr am Baum, nachdem die Grüne Liga mit zehn Helfern alles erntete, was sich zu Saft verarbeiten lässt. Foto: Sabine Brandt

Weimar. "Wir haben alles geschafft, was wir uns vorgenommen hatten", zieht Stefanie Lachmann von der Ehrenamtsagentur zufrieden Bilanz unter Weimars größten Subbotnik, der am zurückliegenden Samstag ausgerufen worden war. Seither sind der

MEISTGELESEN

1 Wenn Paare sehr unterschiedlich aussehen

2 Blutspuren an der Windschutzscheibe: Polizei-Großeinsatz nach Unfallflucht in Heiligenstadt

3 Spurensuche in Jena: Wie ein Sportstudent zum Islamisten wurde

4 Entsetzen bei Mitschülern der in Jena getöteten Leila

5 Lieberknecht entschuldigt sich bei SPD für Wahlkampf-Polemik

MEISTKOMMENTIERT

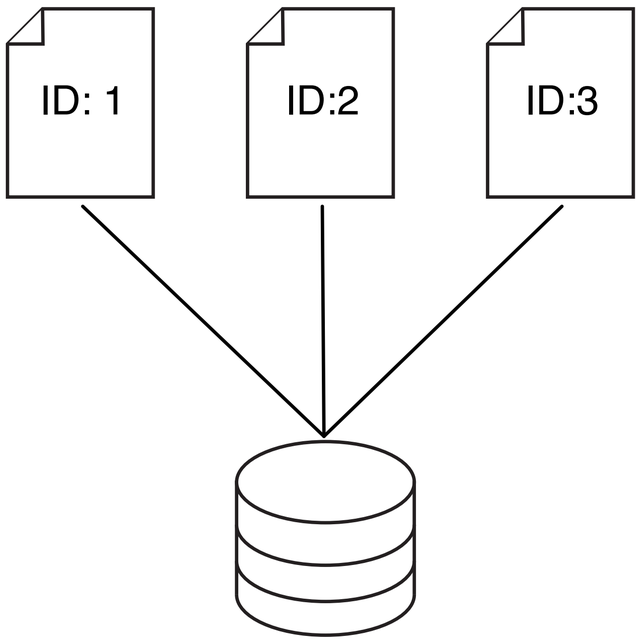
1 AfD will Lieberknecht nicht unterstützen: Unterstützung für Ramelow möglich

2 Skepsis in Erfurt: Ein Kilometer

Near-Duplicates

Die meisten Änderungen auf Webseiten sind nur minimal und betreffen nicht den Inhalt [OP08].

Web-Korpora zur Evaluierung



- Sammlungen von Web-Dokumenten
- Dokumente besitzen eindeutige ID
- verwendete Korpora
  - ClueWeb09 (504 Mio. Dokumente)
  - ClueWeb12 (733 Mio. Dokumente)
- Problem: IDs zwischen Korpora nicht gleich

Relevanzurteile

- Beschreiben Relevanz eines Dokuments für Fragestellung
- Werden manuell erstellt
- 540.551 Relevanzurteile
- 155.419 Dokumente im ClueWeb09
- 2,5 Mannjahre Aufwand (bei 30 Sek. pro Urteil)

Sind Relevanzurteile zwischen Korpora übertragbar?

Vergleichsansätze

- Gemeinsame URLs
- Near-Duplicates

Gemeinsame URLs

Relevanzurteile ClueWeb09

- 34.219 Dokumente mit URL im ClueWeb12

- 23% sind Near-Duplicates

Korpora

- 64,5 Mio. gemeinsame URLs.

- Häufigste Domain ist *en.wikipedia.org*

Near-Duplicates

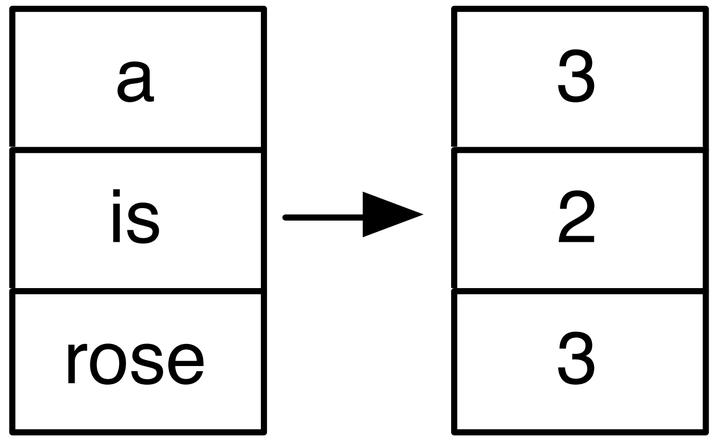
Near-Duplicate Hashingverfahren

- Reduzieren Dokumente auf Hashwerte
- Komplexität und Speicheraufwand wird reduziert
- Zwei Ansätze: Projektion und Einbettung [PS08]
- Untersuchte Verfahren:
  - MinHash (Projektion) [BGMZ97]
  - SimHash (Einbettung) [Cha02]

SimHash [Cha02]

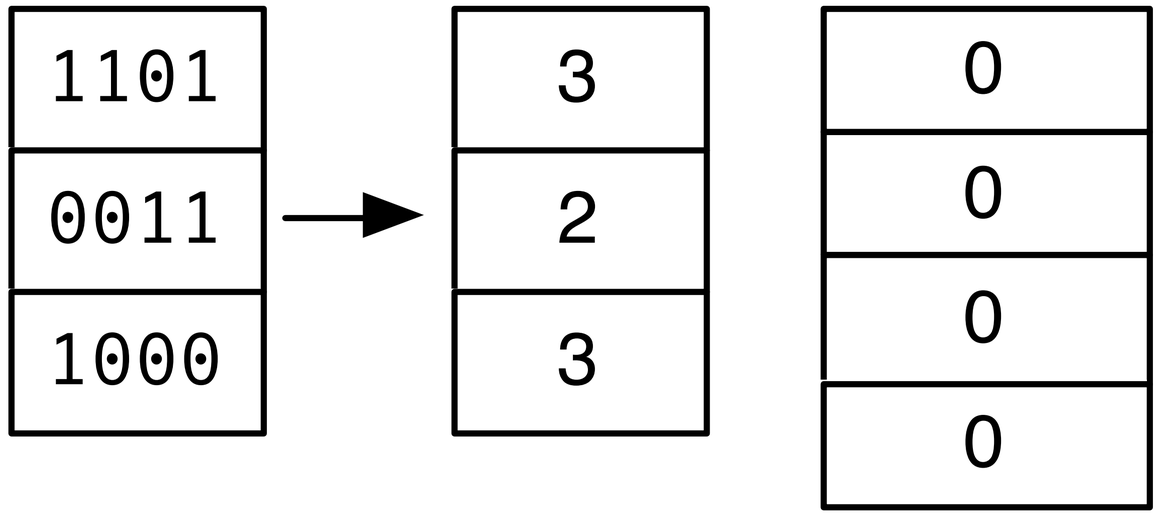
- Erzeugt Locality Sensitive Hashes
- Bildet das ganze Dokument ab (Einbettung)
- Funktioniert mit kurzen Hashlängen (64-bit)
- Vergleich über Hamming-Distanz
- Korreliert mit Kosinus-Ähnlichkeit

1. TF-Vektor erzeugen

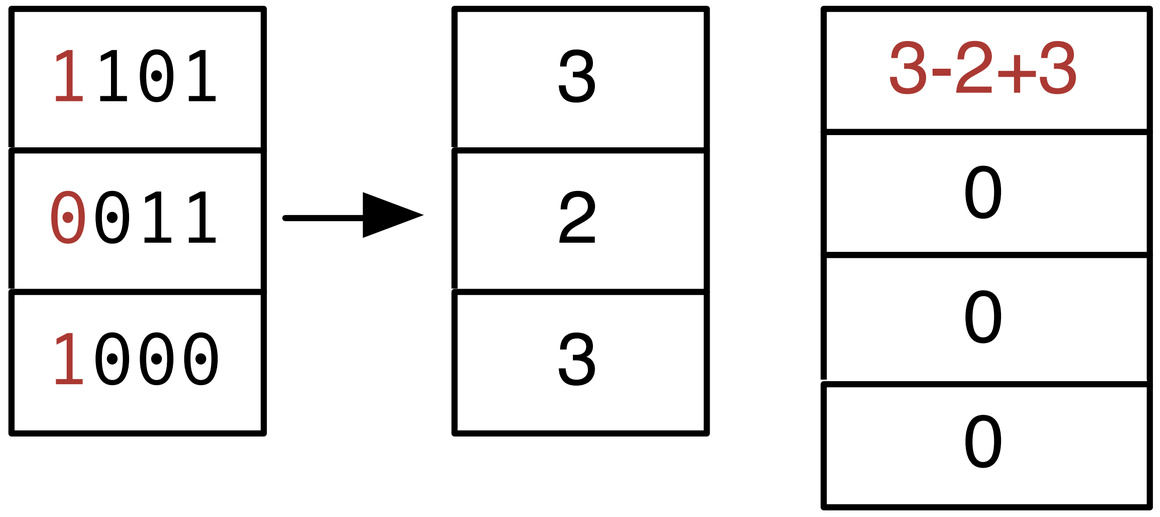


Ausgangstext:
a rose is a rose is a rose

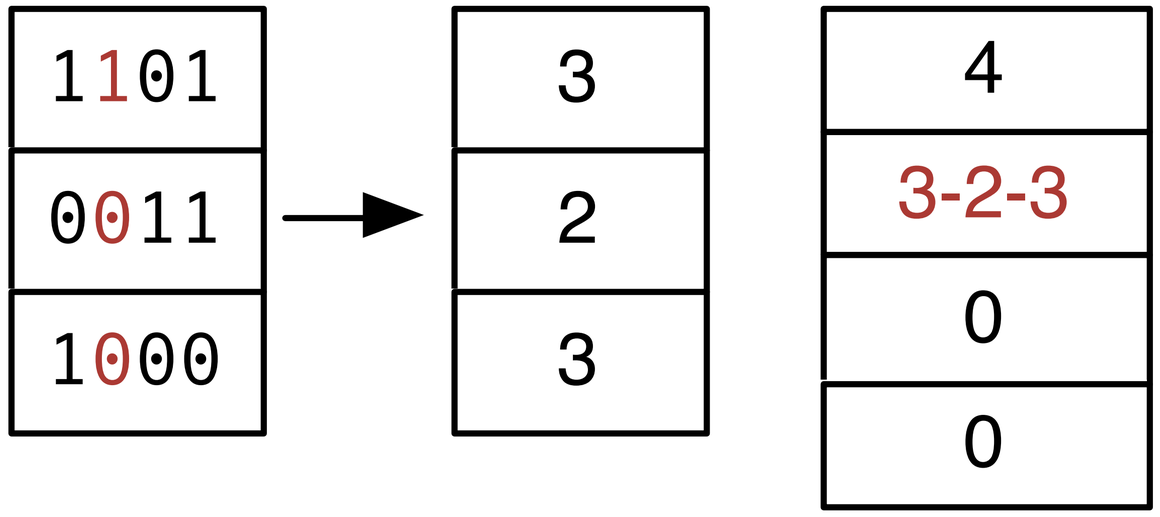
2. Hashes bilden und SimHash-Vektor initialisieren



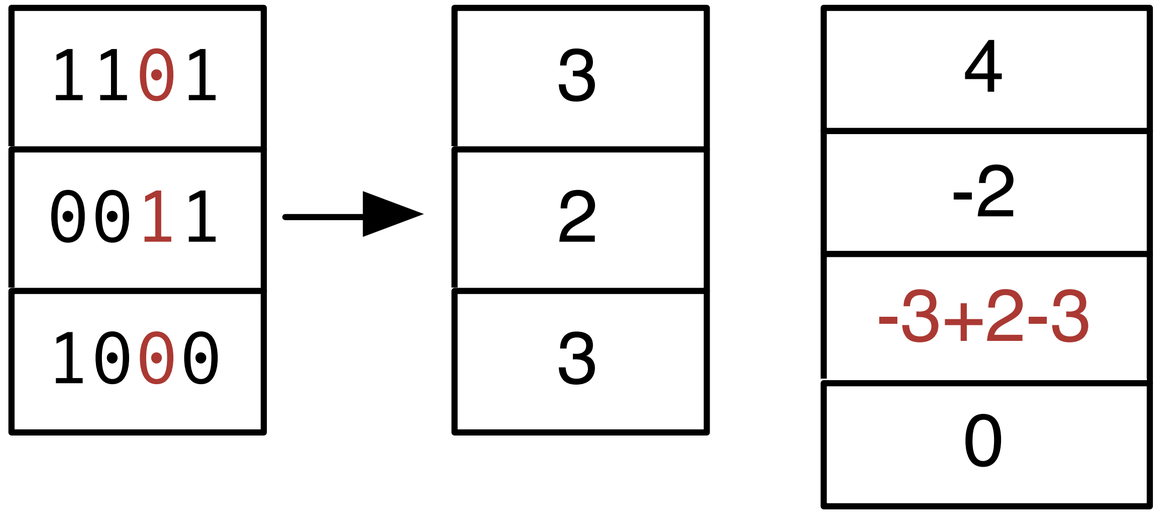
3. SimHash berechnen



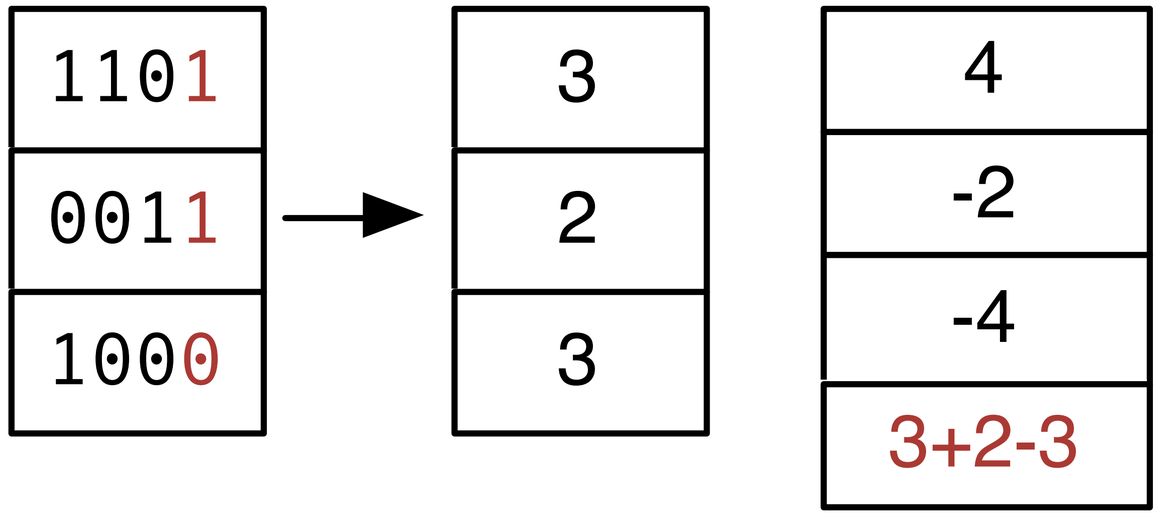
3. SimHash berechnen



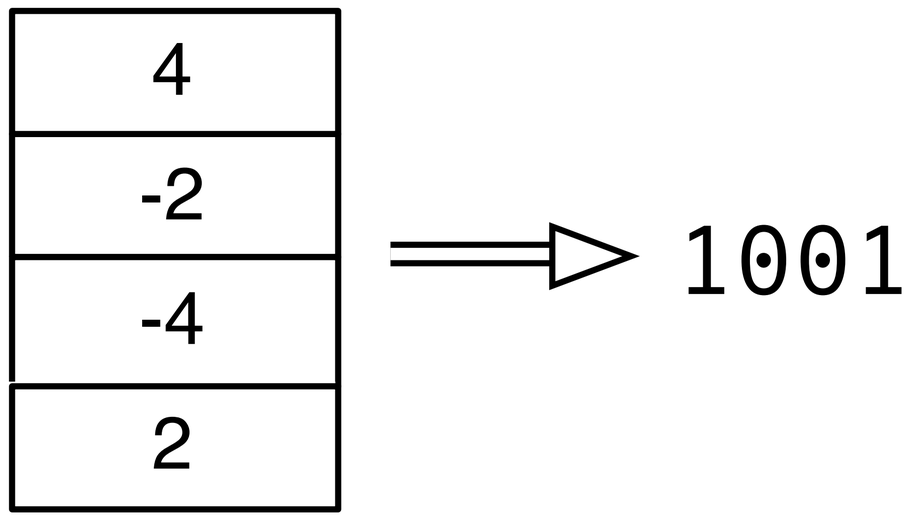
3. SimHash berechnen



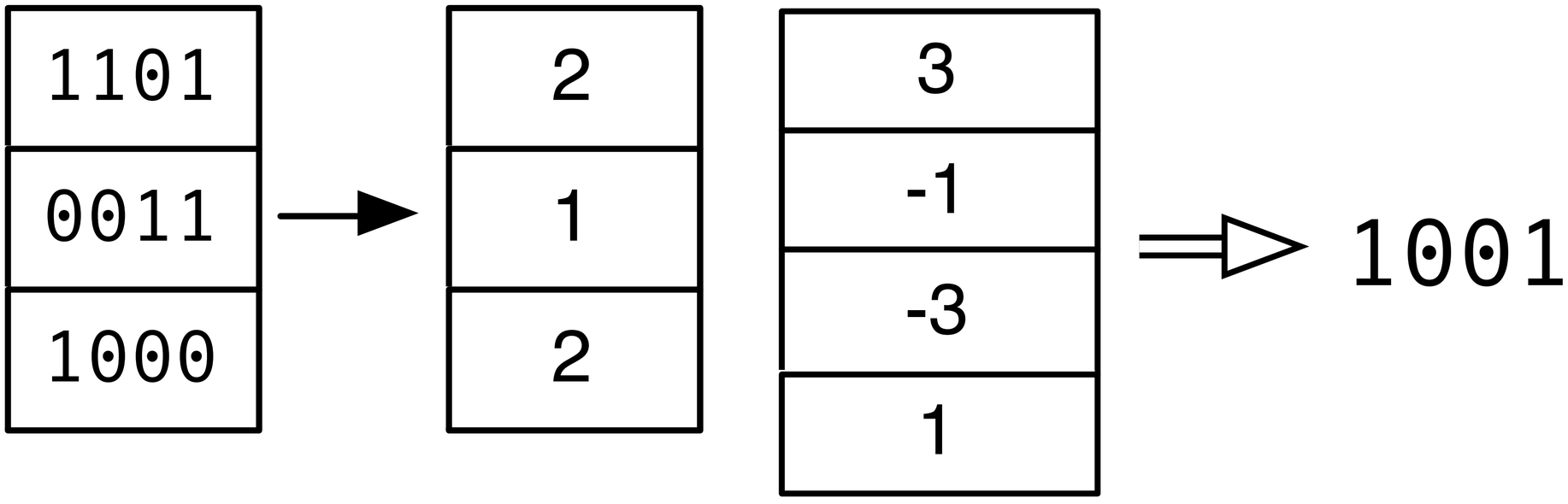
3. SimHash berechnen



4. SimHash erzeugen

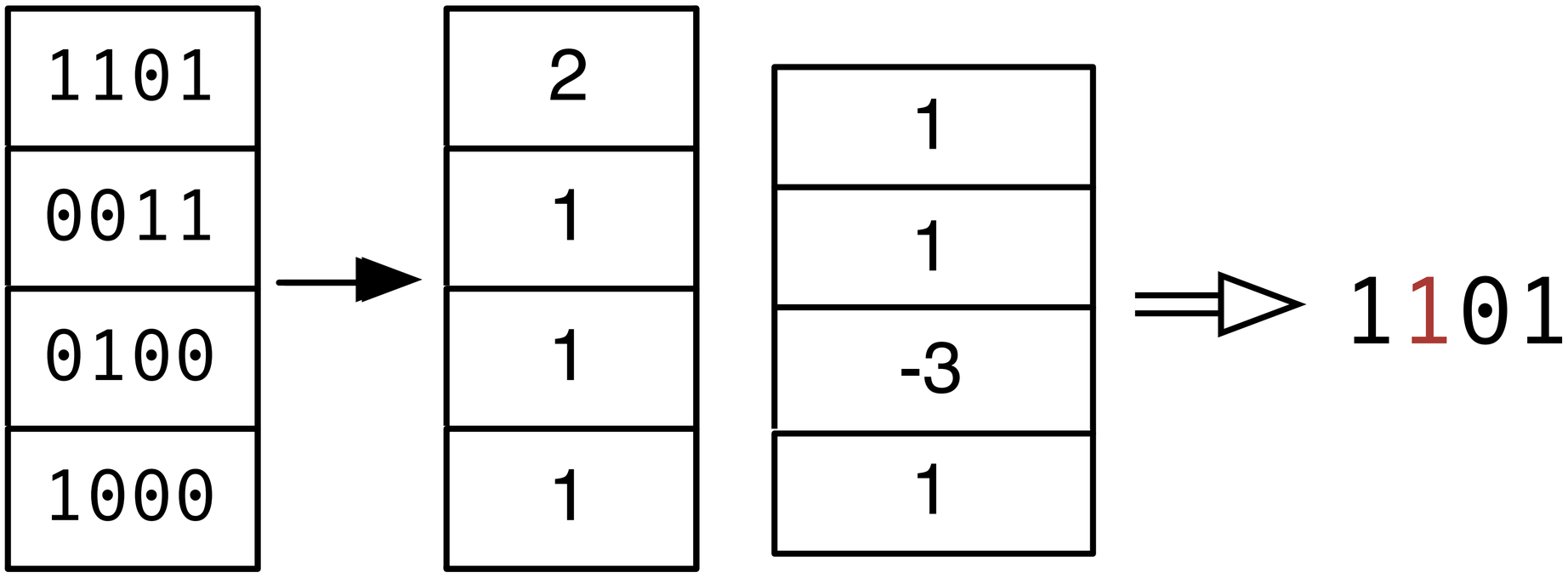


Beispiel 2



Ausgangstext:
is a rose a rose?

Beispiel 3



Ausgangstext:
is a rose a lily?

Skalierung

- Vergleich großer Kollektionen nicht trivial
- Ohne Optimierung 155.000 Hashvergleiche pro Dokument
- Verfahren von Manko et al. ermöglicht effektives Preprocessing über Lookup-Tabellen [MJS07]
- Mit Tabellen 4 Lookups und ca. 16 Vergleiche pro Dokument

Near-Duplicates im ClueWeb12

Suche nach Relevanzurteilen (1)

- SimHash wurde mit einer Länge von 64-bit erzeugt
- Maximale Hamming-Distanz von 3 für Near-Duplicates
- Verschiedene N-Grammlängen wurden bei der Erzeugung genutzt

N-Gramme

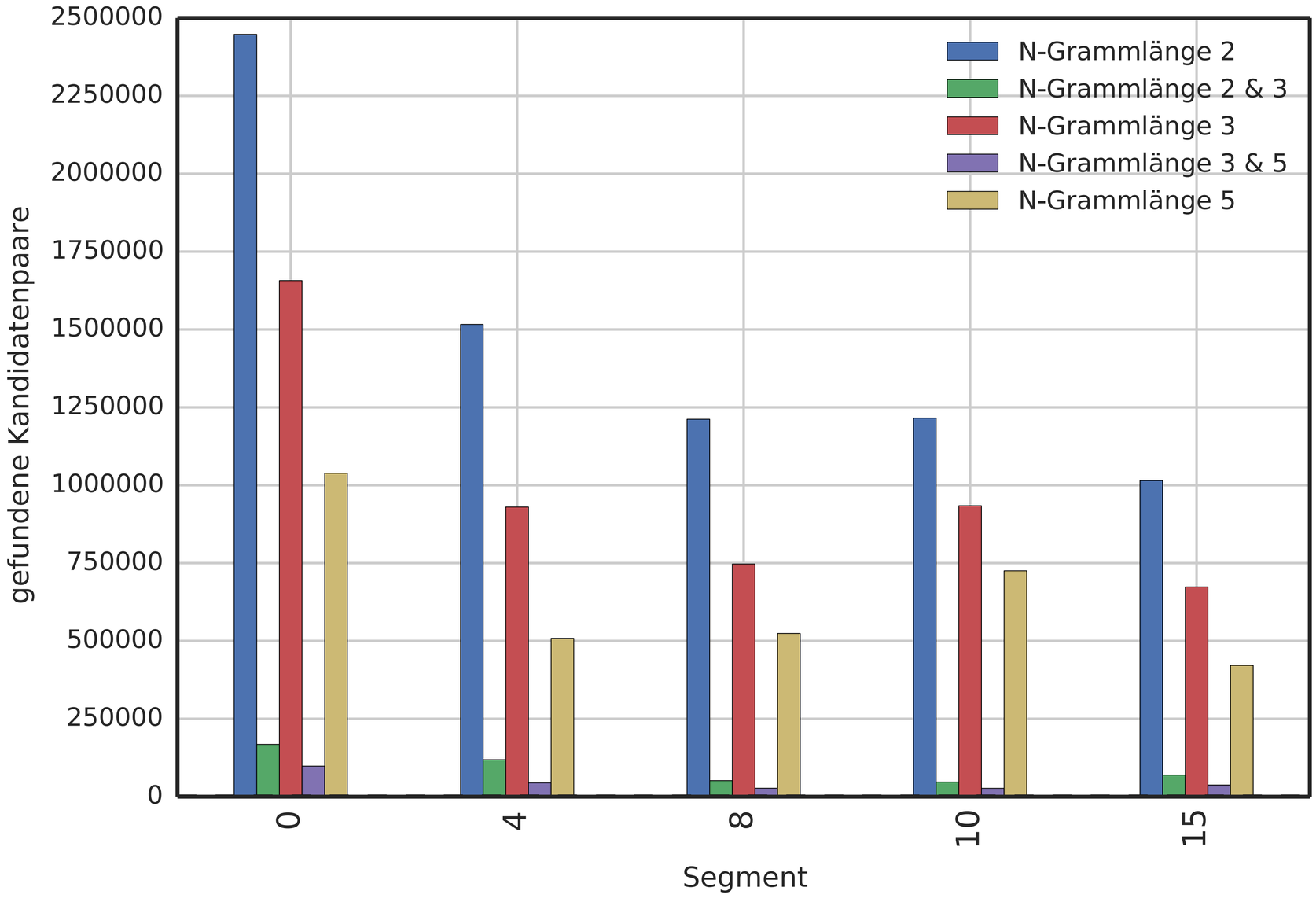
1-Gramme:
{a, rose, is}

2-Gramme:
{a rose, rose is, is a}

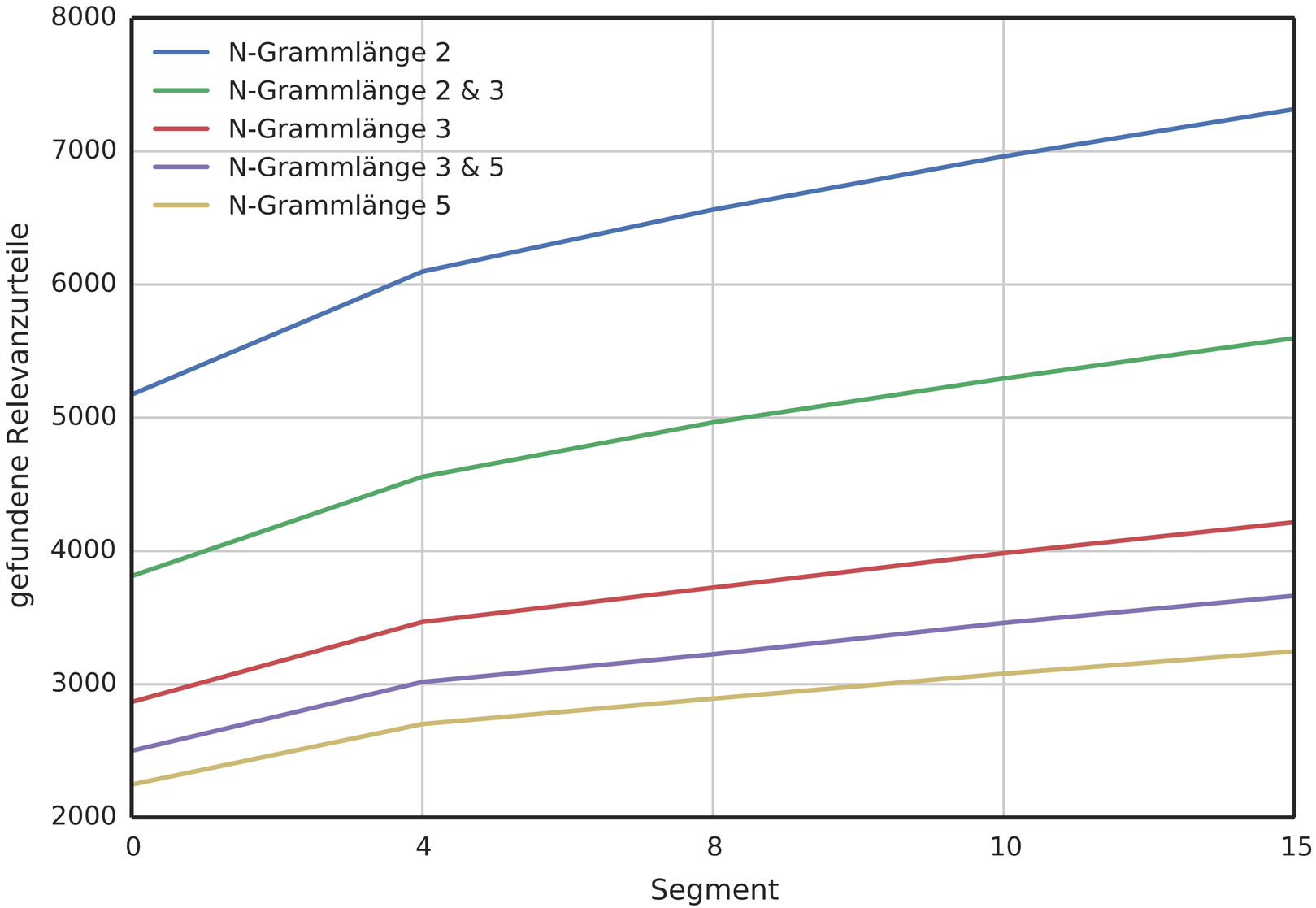
3-Gramme:
{a rose is, rose is a, is a rose}

Ausgangstext:
a rose is a rose is a rose

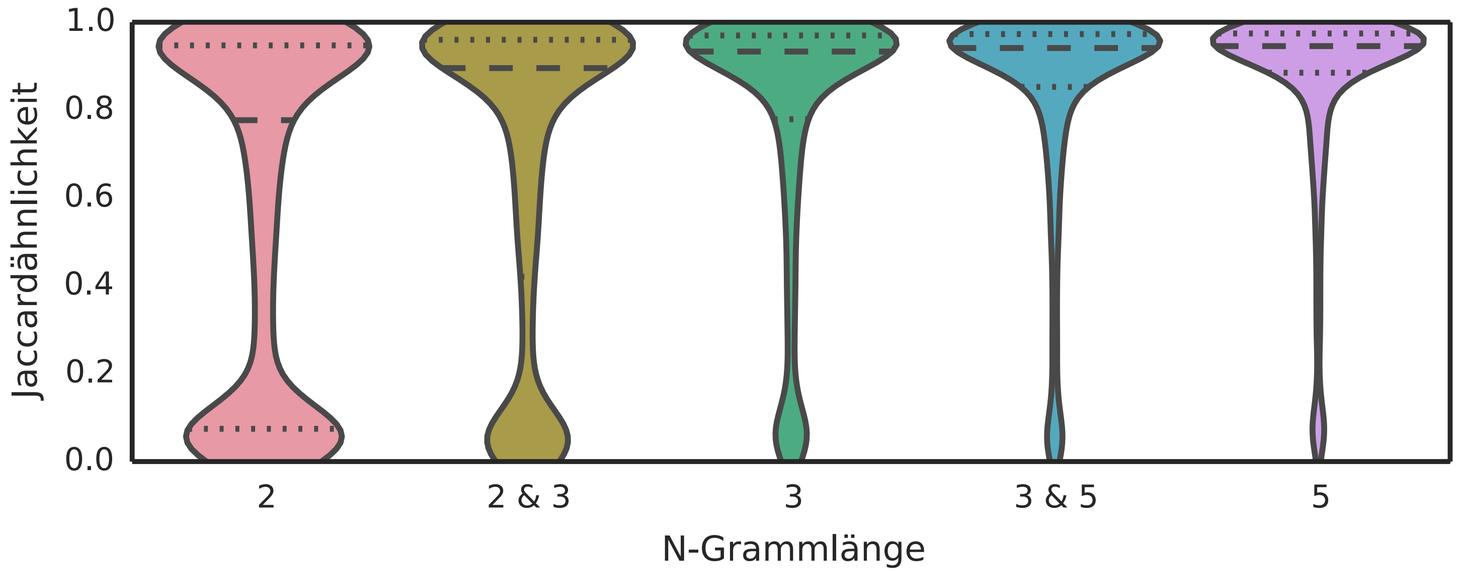
Suche nach Relevanzurteilen (2)



Suche nach Relevanzurteilen (3)



Suche nach Relevanzurteilen (4)



- Qualität steigt mit zunehmender N-Grammlänge
- Absolute Menge der gefundenen Near-Duplicates sinkt aber mit zunehmender N-Grammlänge

Fazit

- Verschiedene SimHash Konfigurationen
- Auswirkungen von N-Grammlängen
- SimHash-Verfahren mit Lookup-Tabellen geeignet
- Übertragbarkeit von Relevanzurteilen unwahrscheinlich

Ausblick

- Ergebnisse für gesamtes ClueWeb12
- Ergebnisse für gesamte Wikipedia
- Service für ClueWeb12
- Vollständiger Vergleich zwischen ClueWeb09 und ClueWeb12

Danke!

Literaturverzeichnis

[BGMZ97] Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse und Geoffrey Zweig. Syntactic clustering of the web. Computer Net- works, 29(8-13):1157–1166, 1997.

[Cha02] Moses Charikar. Similarity estimation techniques from rounding algorithms. In Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montr´eal, Quebec, Canada, Seiten 380–388, 2002.

[MJS07] Gurmeet Singh Manku, Arvind Jain und Anish Das Sarma. Detecting near-duplicates for web crawling. In Proceedings of the 16th International Conference on World Wide Web, WWW'07, Banff, Alberta, Canada, May 8-12, 2007, Seiten 141–150, 2007.

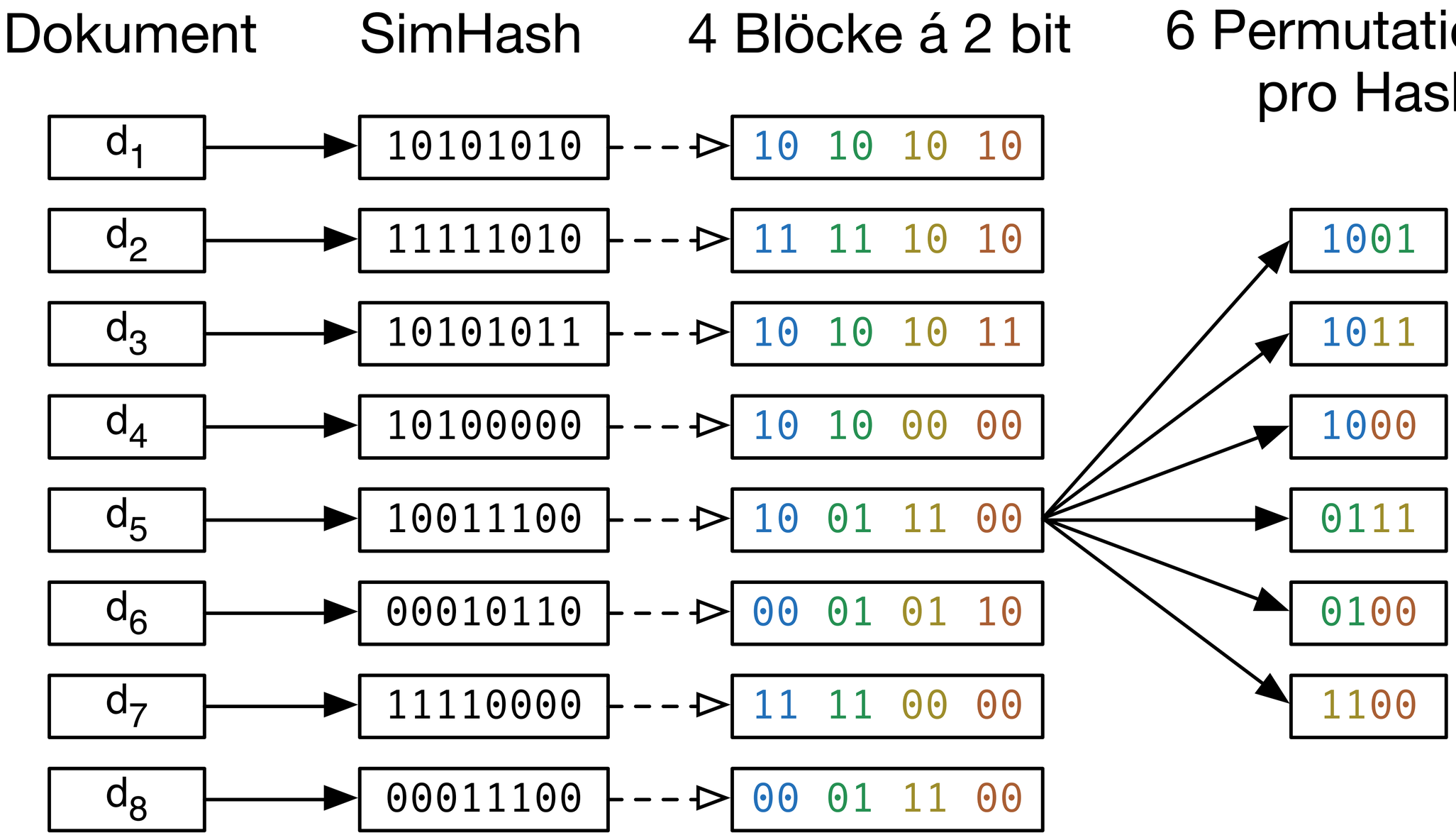
[OP08] Christopher Olston und Sandeep Pandey. Recrawl scheduling based on information longevity. In Proceedings of the 17th International Conference on World Wide Web, WWW'08, Beijing, China, April 21-25, 2008, Seiten 437–446, 2008.

[PS08] Martin Potthast und Benno Stein. New issues in near-duplicate detection. In Data Analysis, Machine Learning and Applications, Seiten 601–609. Springer, 2008.

Lookup Tabellen

Einstellungen

- Gegeben: 8 Dokumente
- Informationsgehalt: 3
- Länge SimHash: 8 bit
- Maximale Hammingdistanz: 2
- Anzahl Blöcke: 4

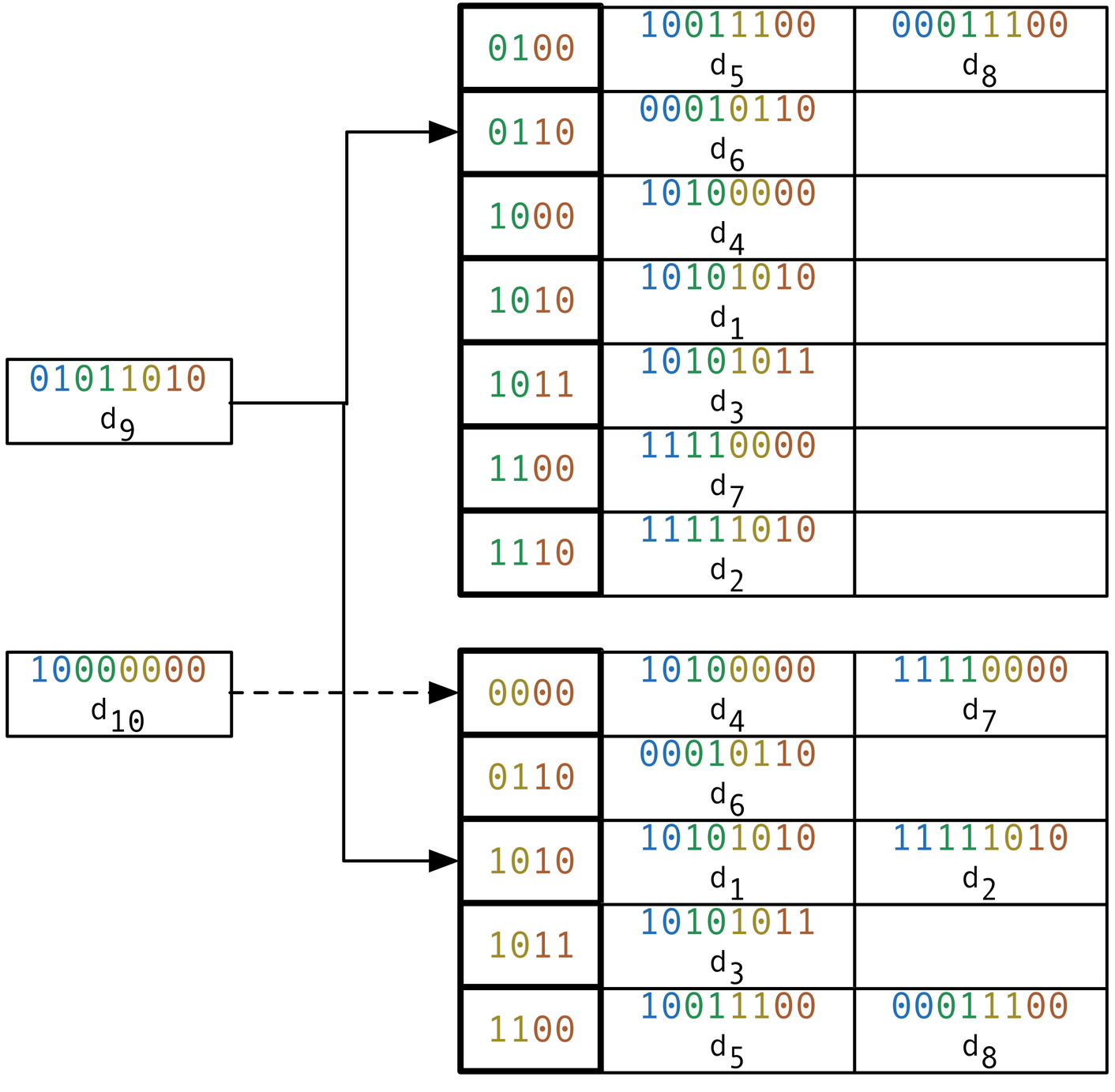

Dokument
SimHash
4 Blöcke á 2 bit
6 Permutationen pro Hash
d1
10101010
10 10 10 10
d2
11111010
11 11 10 10
d3
10101011
10 10 10 11
d4
10100000
10 10 00 00
d5
10011100
10 01 11 00
d6
00010110
00 01 01 10
d7
11110000
11 11 00 00
d8
00011100
00 01 11 00
1001
1011
1000
0111
0100
1100

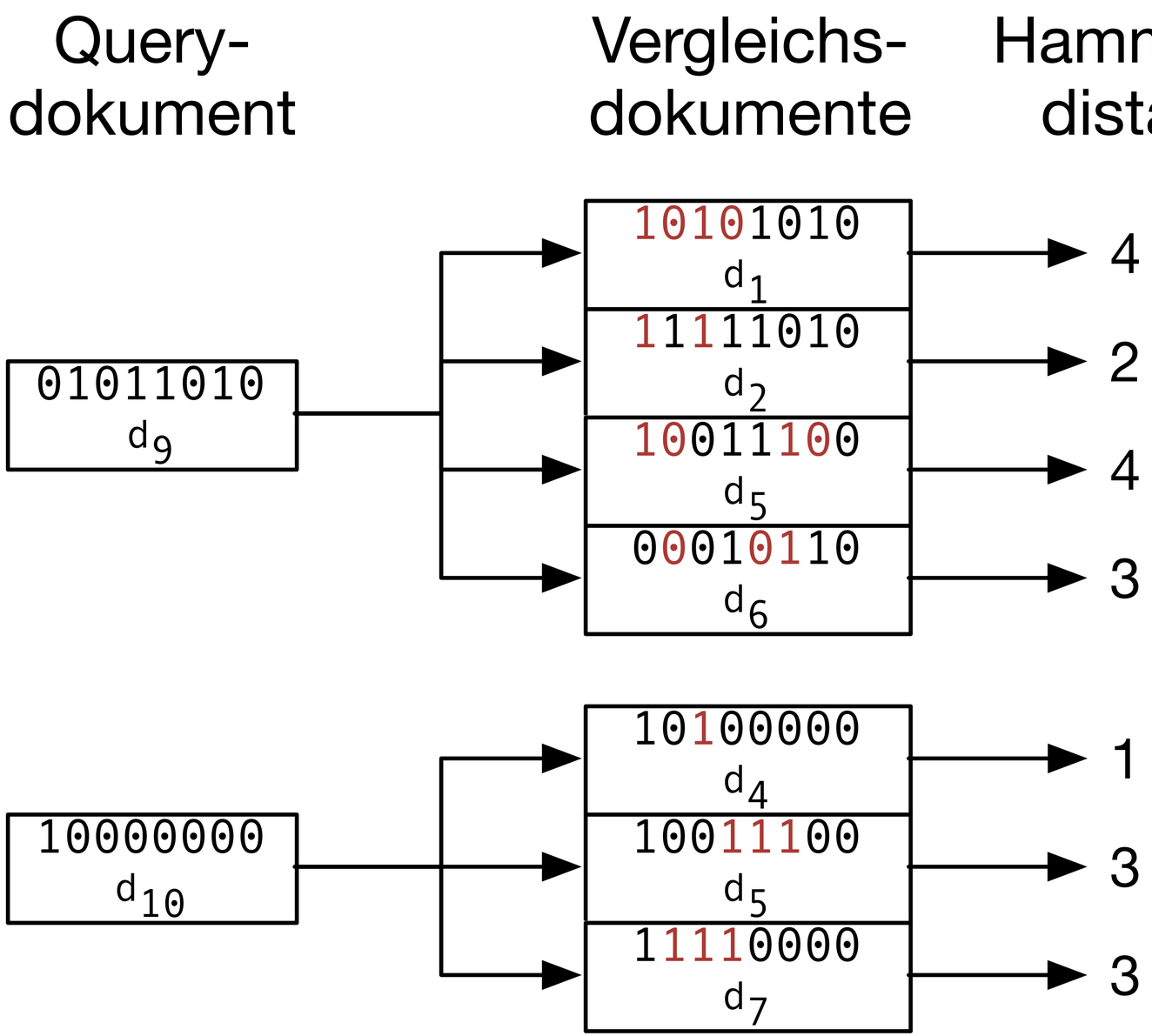
t_5

0100	10011100 d_5	00011100 d_8
0110	00010110 d_6	
1000	10100000 d_4	
1010	10101010 d_1	
1011	10101011 d_3	
1100	11110000 d_7	
1110	11111010 d_2	

t_6

0000	10100000 d_4	11110000 d_7
0110	00010110 d_6	
1010	10101010 d_1	11111010 d_2
1011	10101011 d_3	
1100	10011100 d_5	00011100 d_8


01011010
d9
10000000
d10
0100
10011100
d5
00011100
d8
0110
00010110
d6
1000
10100000
d4
1010
10101010
d1
1011
10101011
d3
1100
11110000
d7
1110
11111010
d2
0000
10100000
d4
11110000
d7
0110
00010110
d6
1010
10101010
d1
11111010
d2
1011
10101011
d3
1100
10011100
d5
00011100
d8


Query-
dokument
Vergleichs-
dokumente
Hamming-
distanz
01011010
d9
10101010
d1
4
11111010
d2
2
10011100
d5
4
00010110
d6
3
10000000
d10
10100000
d4
1
10011100
d5
3
11110000
d7
3

Shingles

Text

a rose is a rose is a rose

4-Shingles

{(a, rose, is, a), (rose, is, a, rose), (is, a, rose, is)}