



Internetsuche und Google Page-Rank -

Wie wird was durch wen gefunden?

Woche 2

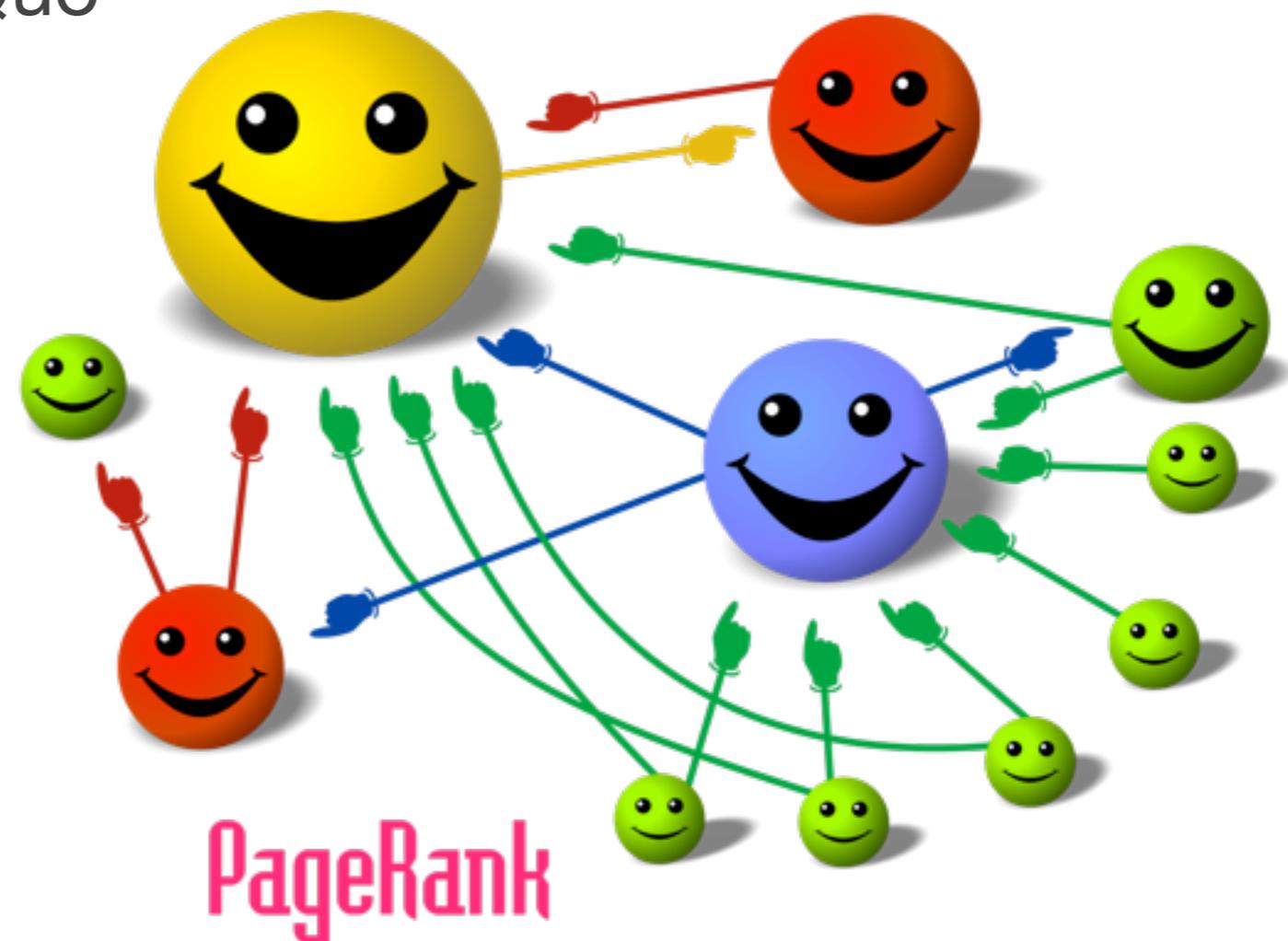
HPI

Magnus Knuth & Nadine Steinmetz
Hasso Plattner Institute for IT-Systems Engineering
University of Potsdam

Internetsuche und Google Page-Rank

1. Wiederholung

2. Term- und Dokumentenranking
3. SEO-Experiment - Status Quo



Was ist ein Index?

Suchindex - Arten

Suchindex - Arten

- 4
- **Forward-Index:** Zuordnung von Termen zu Webseiten (Dokument muss nicht komplett durchsucht werden, Terme sortiert)

Suchindex - Arten

- 4
- **Forward-Index:** Zuordnung von Termen zu Webseiten (Dokument muss nicht komplett durchsucht werden, Terme sortiert)

Dok1	Die, Katze, macht, miau
Dok2	Die, Kuh, macht, muh
Dok3	Die, Erbse, grün, ist
Dok4	Die, Kuh, lacht

Suchindex - Arten

- 4
- **Forward-Index:** Zuordnung von Termen zu Webseiten (Dokument muss nicht komplett durchsucht werden, Terme sortiert)

Dok1	Die, Katze, macht, miau
Dok2	Die, Kuh, macht, muh
Dok3	Die, Erbse, grün, ist
Dok4	Die, Kuh, lacht

- **Invertierter Index:** Zuordnung von Webseiten zu Termen (findet schnell Dokumente zu Suchterm)

Suchindex - Arten

- 4
- **Forward-Index:** Zuordnung von Termen zu Webseiten (Dokument muss nicht komplett durchsucht werden, Terme sortiert)

Dok1	Die, Katze, macht, miau
Dok2	Die, Kuh, macht, muh
Dok3	Die, Erbse, grün, ist
Dok4	Die, Kuh, lacht

- **Invertierter Index:** Zuordnung von Webseiten zu Termen (findet schnell Dokumente zu Suchterm)

Die	Dok1, Dok2, Dok3, Dok4
Kuh	Dok2, Dok4
macht	Dok1, Dok2
muh	Dok2

Suchindex - Arten

- 4
- **Forward-Index:** Zuordnung von Termen zu Webseiten (Dokument muss nicht komplett durchsucht werden, Terme sortiert)

Dok1	Die, Katze, macht, miau
Dok2	Die, Kuh, macht, muh
Dok3	Die, Erbse, grün, ist
Dok4	Die, Kuh, lacht

- **Invertierter Index:** Zuordnung von Webseiten zu Termen (findet schnell Dokumente zu Suchterm)

Die	Dok1, Dok2, Dok3, Dok4
Kuh	Dok2, Dok4
macht	Dok1, Dok2
muh	Dok2

- **Dokument-Term-Matrix:** zweidimensionale Matrix (Anzahl des Auftreten in Dokument)

Suchindex - Arten

- 4 **Forward-Index:** Zuordnung von Termen zu Webseiten (Dokument muss nicht komplett durchsucht werden, Terme sortiert)

Dok1	Die, Katze, macht, miau
Dok2	Die, Kuh, macht, muh
Dok3	Die, Erbse, grün, ist
Dok4	Die, Kuh, lacht

- Invertierter Index:** Zuordnung von Webseiten zu Termen (findet schnell Dokumente zu Suchterm)

Die	Dok1, Dok2, Dok3, Dok4
Kuh	Dok2, Dok4
macht	Dok1, Dok2
muh	Dok2

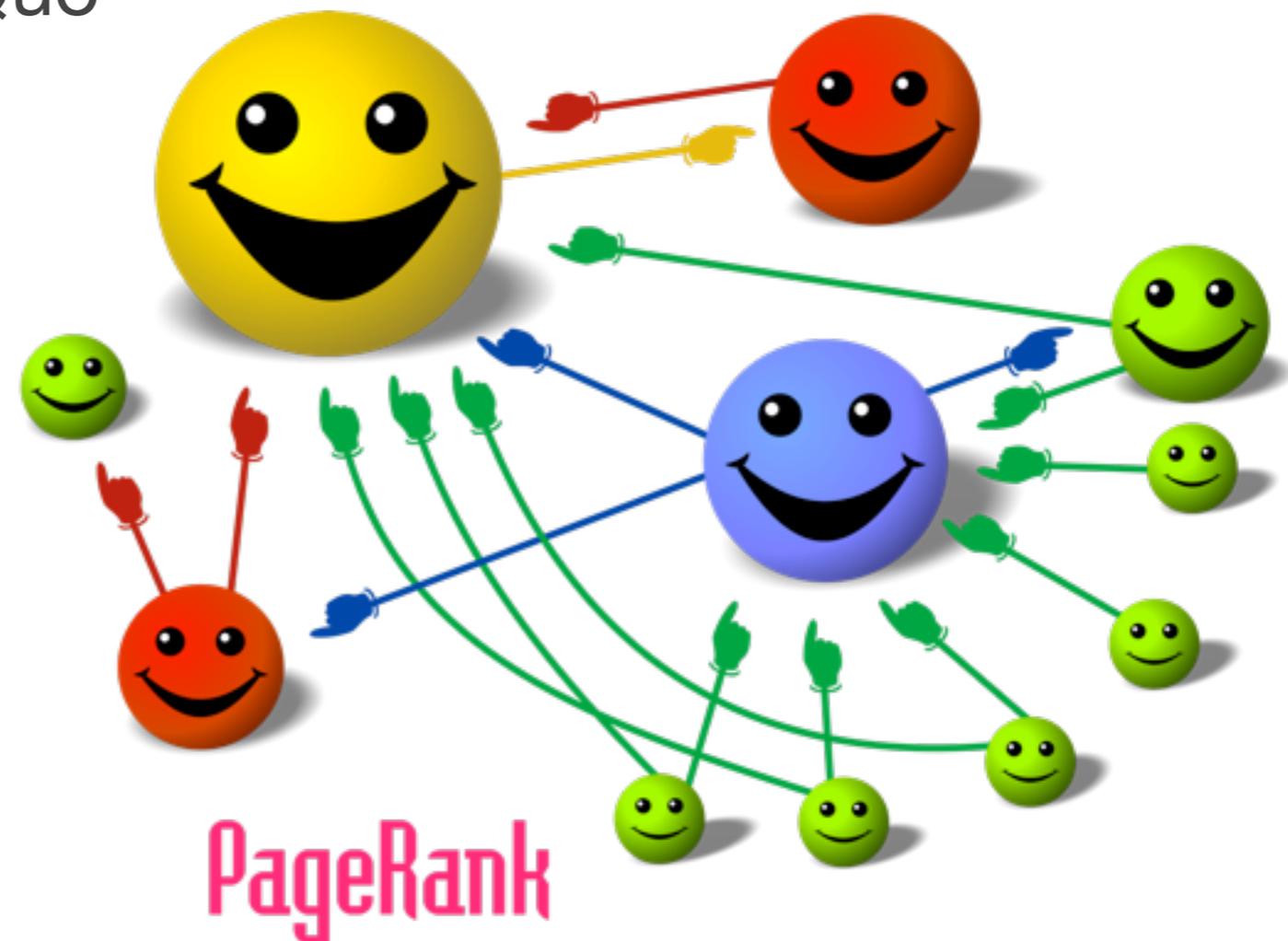
- Dokument-Term-Matrix:** zweidimensionale Matrix (Anzahl des Auftreten in Dokument)

	Dok1	Dok2	Dok3	Dok4
Die	1	1	1	2
Kuh		1		1
macht	1	1		
muh		1		

Internetsuche und Google Page-Rank

5

1. Wiederholung
2. Term- und Dokumentenranking
3. SEO-Experiment - Status Quo

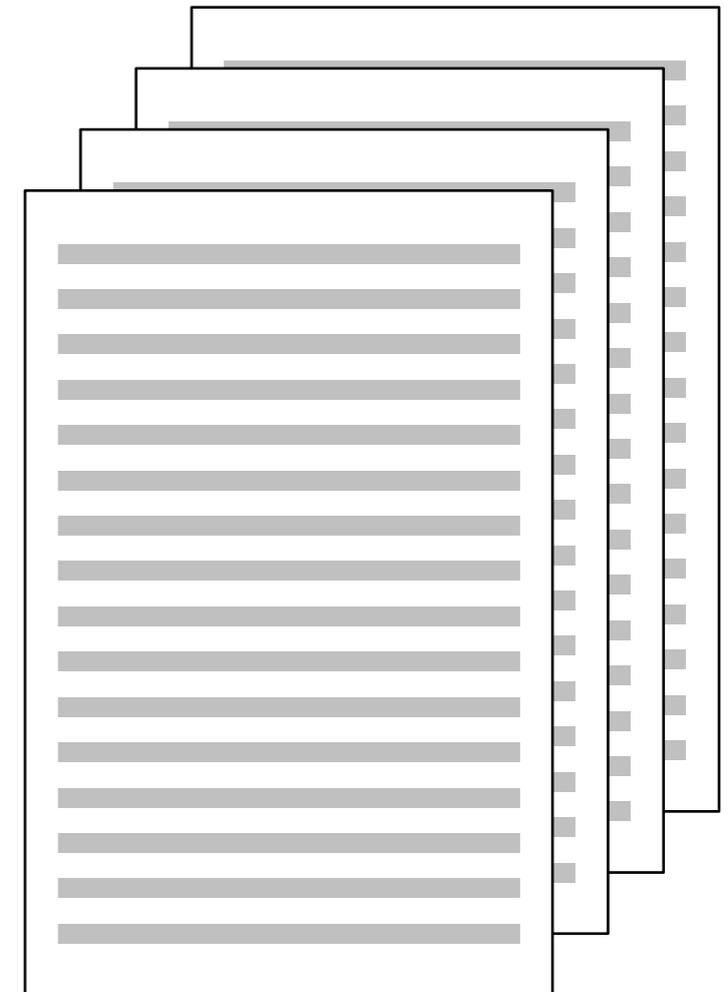


Dokumentensuche

6

Angenommen, wir haben eine Menge von Dokumenten (oder Webseiten).

Welches Ergebnis bekommen, wenn wir diese Dokumente nach dem Wort „das“ durchsuchen?



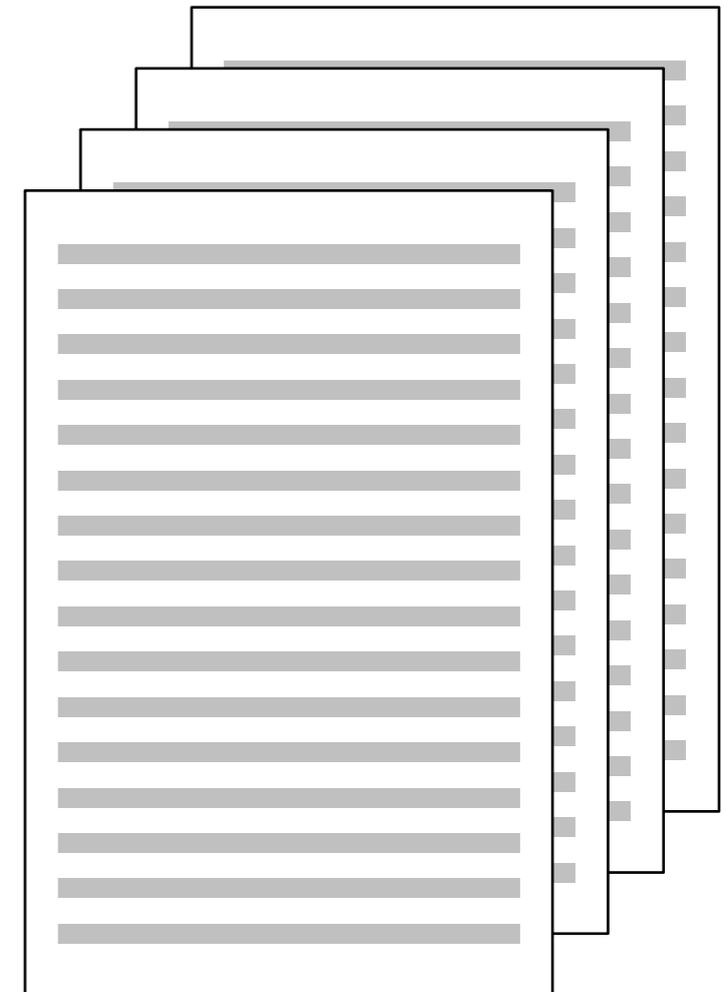
Dokumentensuche

6

Angenommen, wir haben eine Menge von Dokumenten (oder Webseiten).

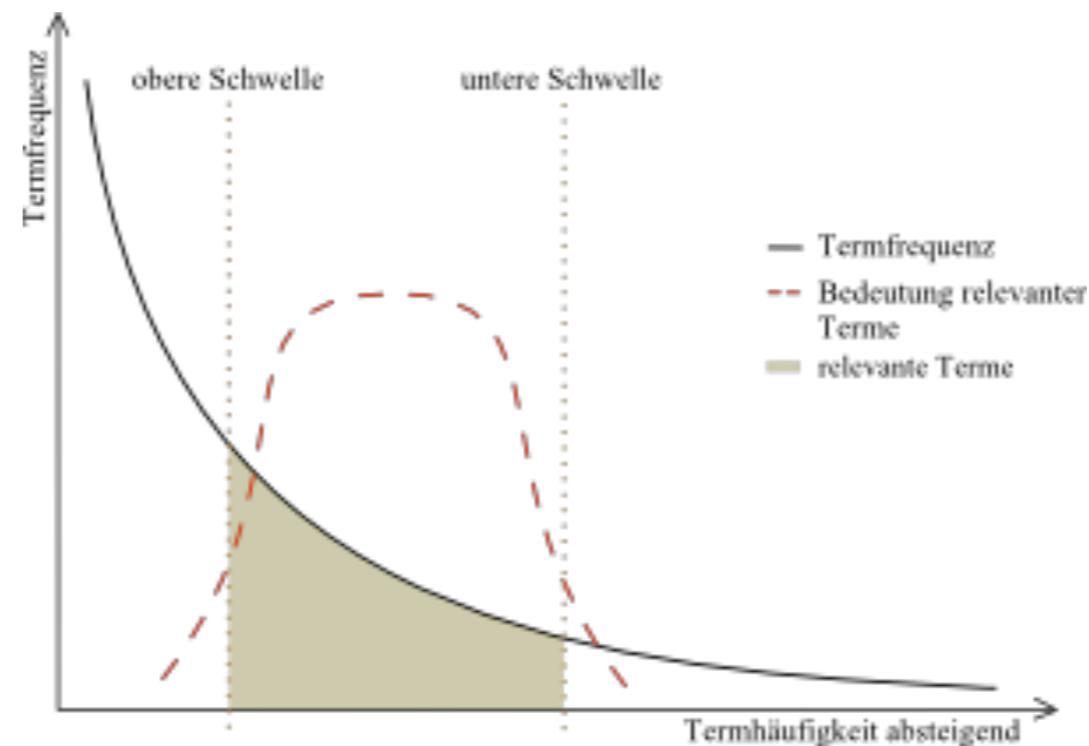
Welches Ergebnis bekommen, wenn wir diese Dokumente nach dem Wort „das“ durchsuchen?

...und was passiert, wenn wir nach „venezianischer Zwergsprinter“ suchen?



Termfrequenz - Zipf'sches Gesetz

7 ... gibt die Wertigkeit von Wörtern anhand der Häufigkeit ihres Vorkommens an (Wert umgekehrt proportional zum Rang).

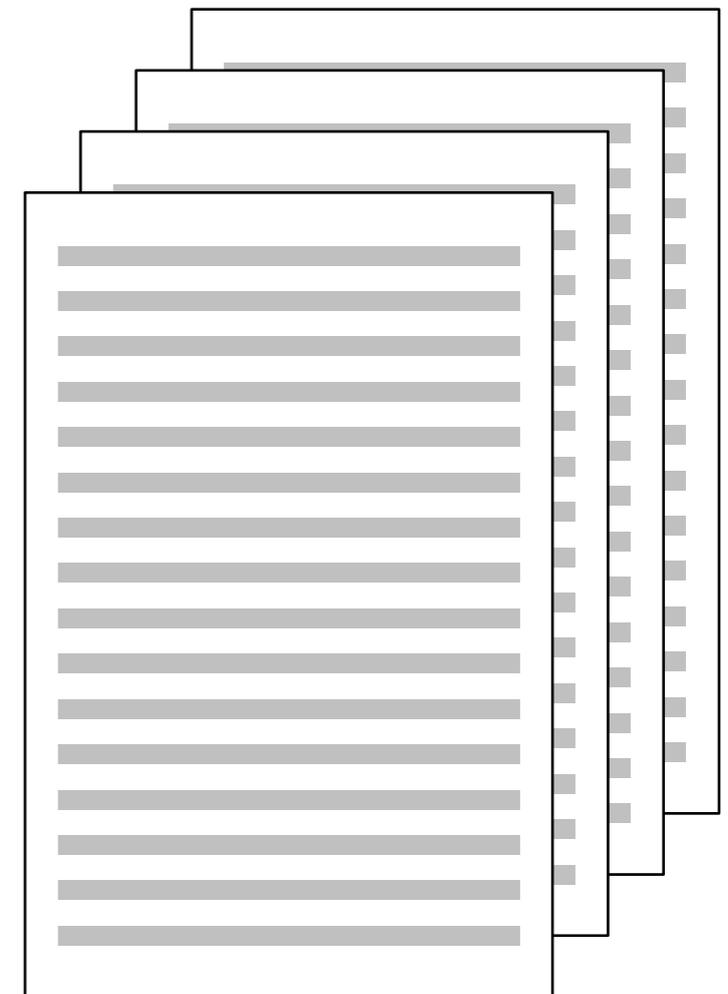


Dokumentenranking

8

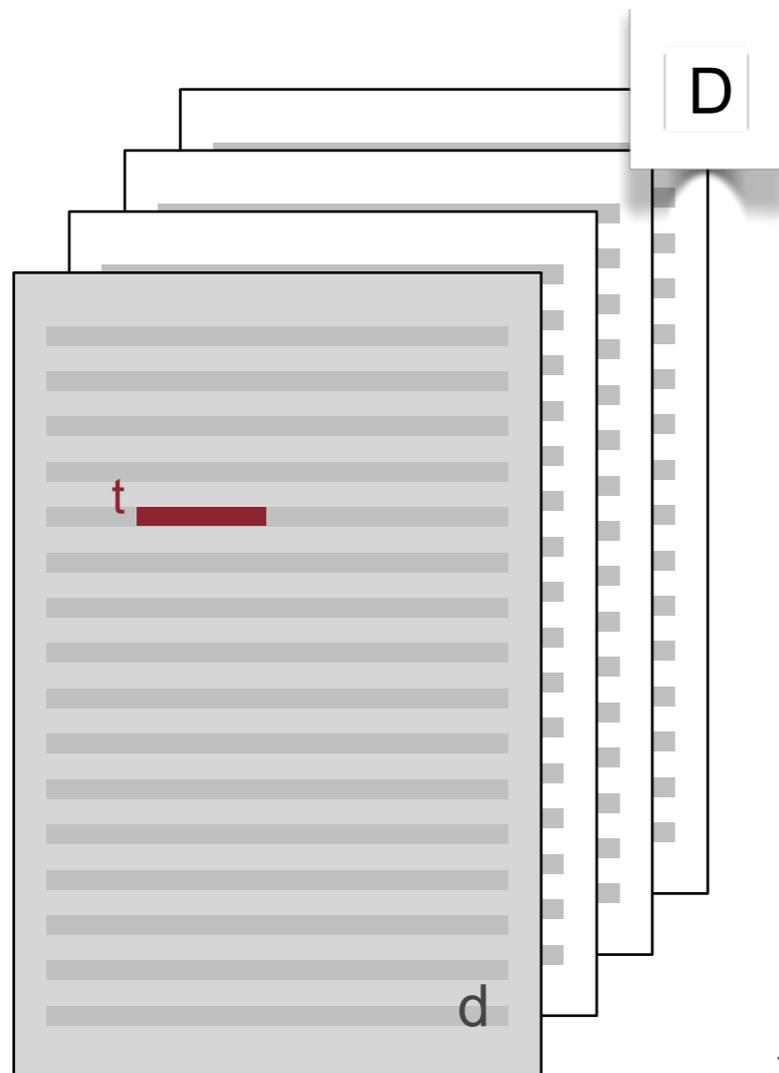
Angenommen, wir haben mehrere Dokumente gefunden, in denen das gesuchte Wort vorkommt.

Welches Dokument geben wir als erstes in einer Liste der Suchergebnisse aus?



Dokumentenranking - TF-IDF

9



$$tf(t, d) = \frac{|\{t \in d\}|}{|\{w \in d\}|}$$

$$idf(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D)$$

Dokumentenranking - TF-IDF

10 Beispiel:

Dokument A:

In Bayern leben viele Bayern. Das kann man sich auch in einem Statistik-Amt bestätigen lassen.

Dokument B:

Heute Abend ist das erste Spiel des Halbfinals der Champions League. Es spielt Bayern München gegen Barcelona. Morgen Abend spielt der BVB.

Dokument C:

Uli Hoeneß ist der Präsident des FC Bayern München. Vielleicht ist er das aber bald nicht mehr.

Wie groß ist die Relevanz des Wortes „Bayern“ für Dokument A in dieser Sammlung?

Dokumentenranking - TF-IDF

10 Beispiel:

Dokument A:

In Bayern leben viele Bayern. Das kann man sich auch in einem Statistik-Amt bestätigen lassen.

Dokument B:

Heute Abend ist das erste Spiel des Halbfinals der Champions League. Es spielt Bayern München gegen Barcelona. Morgen Abend spielt der BVB.

Dokument C:

Uli Hoeneß ist der Präsident des FC Bayern München. Vielleicht ist er das aber bald nicht mehr.

Wie groß ist die Relevanz des Wortes „Bayern“ für Dokument A in dieser Sammlung?

$$tf(t, d) = \frac{|\{t \in d\}|}{|\{w \in d\}|}$$

$$idf(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D)$$

Dokumentenranking - TF-IDF

10 Beispiel:

Dokument A:

In Bayern leben viele Bayern. Das kann man sich auch in einem Statistik-Amt bestätigen lassen.

Dokument B:

Heute Abend ist das erste Spiel des Halbfinals der Champions League. Es spielt Bayern München gegen Barcelona. Morgen Abend spielt der BVB.

Dokument C:

Uli Hoeneß ist der Präsident des FC Bayern München. Vielleicht ist er das aber bald nicht mehr.

Wie groß ist die Relevanz des Wortes „BVB“ für Dokument B in dieser Sammlung?

$$tf(t, d) = \frac{|\{t \in d\}|}{|\{w \in d\}|}$$

$$idf(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D)$$

Dokumentenranking - TF-IDF

10 Beispiel:

Dokument A:

In Bayern leben viele Bayern. Das kann man sich auch in einem Statistik-Amt bestätigen lassen.

Dokument B:

Heute Abend ist das erste Spiel des Halbfinals der Champions League. Es spielt Bayern München gegen Barcelona. Morgen Abend spielt der BVB.

Dokument C:

Uli Hoeneß ist der Präsident des FC Bayern München. Vielleicht ist er das aber bald nicht mehr.

Wie groß ist die Relevanz des Wortes **„München“** für Dokument C in dieser Sammlung?

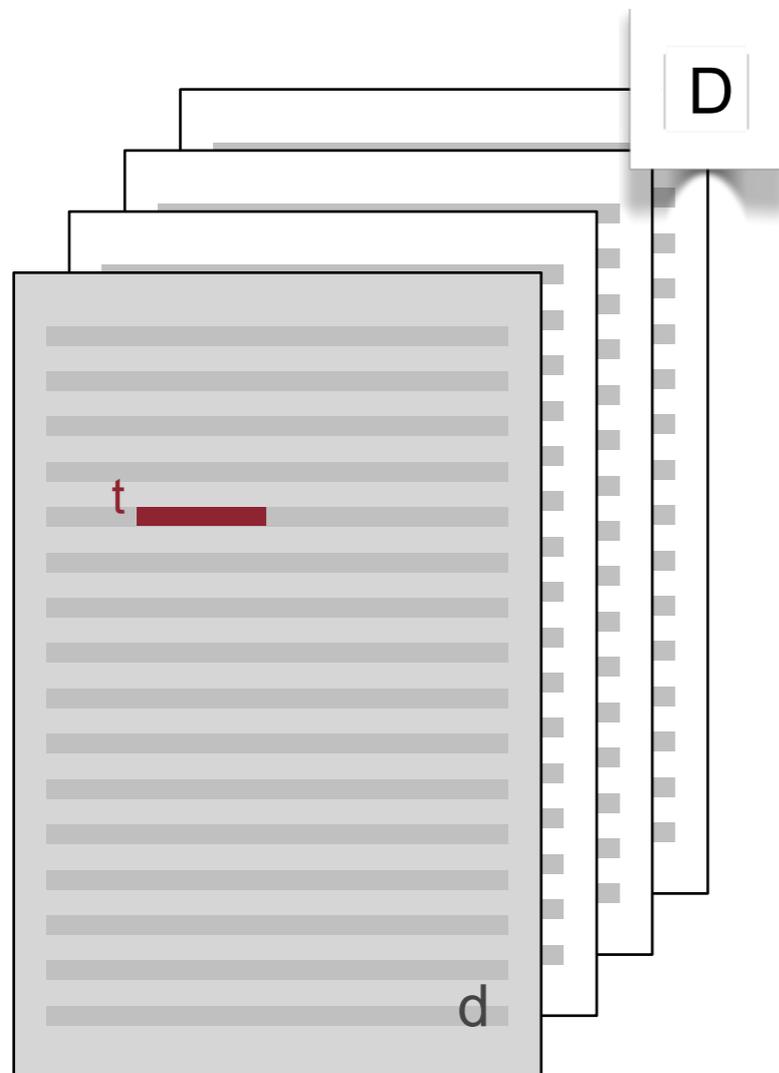
$$tf(t, d) = \frac{|\{t \in d\}|}{|\{w \in d\}|}$$

$$idf(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D)$$

Dokumentenranking - WDF-IDF

11



$$\text{wdf}(t, d) = \frac{\log(f(t, d))}{\log(|\{w \in d\}|)}$$

$$\text{idf}(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

$$\text{wdfidf}(t, d, D) = \text{wdf}(t, d) \times \text{idf}(t, D)$$

Dokumentenranking - Stoppwörter

- ¹² ...Wörter, die bei einer Volltextindexierung nicht beachtet werden, da sie sehr häufig auftreten und gewöhnlich keine Relevanz für die Erfassung des Dokumentinhalts besitzen.

Dokumentenranking - Stoppwörter

¹² ...Wörter, die bei einer Volltextindexierung nicht beachtet werden, da sie sehr häufig auftreten und gewöhnlich keine Relevanz für die Erfassung des Dokumentinhalts besitzen.

- bestimmte Artikel
- unbestimmte Artikel
- Konjunktionen
- häufig gebrauchte Präpositionen
- Hilfsverben

Dokumentenranking - Stemming

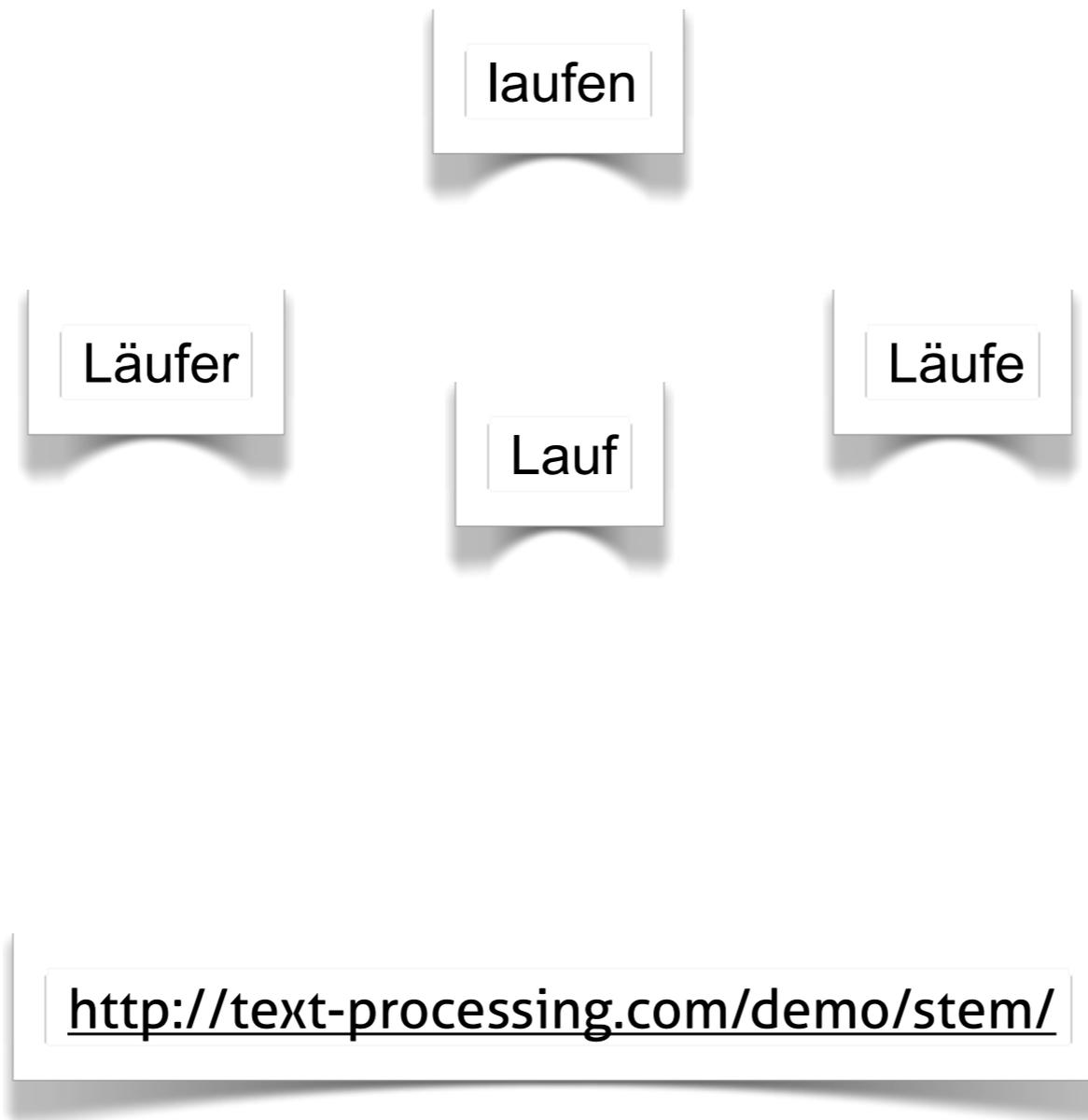
13

laufen

<http://text-processing.com/demo/stem/>

Dokumentenranking - Stemming

13



Dokumentenranking - Synonyme

¹⁴ ... mehrere unterschiedliche Wörter, die die gleiche Bedeutung haben oder mit denen das gleiche gemeint sein kann.

Beispiele:

Dokumentenranking - Synonyme

¹⁴ ... mehrere unterschiedliche Wörter, die die gleiche Bedeutung haben oder mit denen das gleiche gemeint sein kann.

Beispiele:



Pferd:

Dokumentenranking - Synonyme

¹⁴ ... mehrere unterschiedliche Wörter, die die gleiche Bedeutung haben oder mit denen das gleiche gemeint sein kann.

Beispiele:

Pferd:

Gaul, Klepper, Schimmel, Ross, Rappen ...

Dokumentenranking - Synonyme

¹⁴ ... mehrere unterschiedliche Wörter, die die gleiche Bedeutung haben oder mit denen das gleiche gemeint sein kann.

Beispiele:

Pferd:

Gaul, Klepper, Schimmel, Ross, Rappen ...

Ballspielverein Borussia 09 e. V. Dortmund:

Dokumentenranking - Synonyme

¹⁴ ... mehrere unterschiedliche Wörter, die die gleiche Bedeutung haben oder mit denen das gleiche gemeint sein kann.

Beispiele:

Pferd:

Gaul, Klepper, Schimmel, Ross, Rappen ...

Ballspielverein Borussia 09 e. V. Dortmund:

BVB, die Schwarz-Gelben, die Borussen, Dortmund ...

Dokumentenranking - Synonyme

¹⁴ ... mehrere unterschiedliche Wörter, die die gleiche Bedeutung haben oder mit denen das gleiche gemeint sein kann.

Beispiele:

Pferd:

Gaul, Klepper, Schimmel, Ross, Rappen ...

Ballspielverein Borussia 09 e. V. Dortmund:

BVB, die Schwarz-Gelben, die Borussen, Dortmund ...

Angela Merkel

Dokumentenranking - Synonyme

¹⁴ ... mehrere unterschiedliche Wörter, die die gleiche Bedeutung haben oder mit denen das gleiche gemeint sein kann.

Beispiele:

Pferd:

Gaul, Klepper, Schimmel, Ross, Rappen ...

Ballspielverein Borussia 09 e. V. Dortmund:

BVB, die Schwarz-Gelben, die Borussen, Dortmund ...

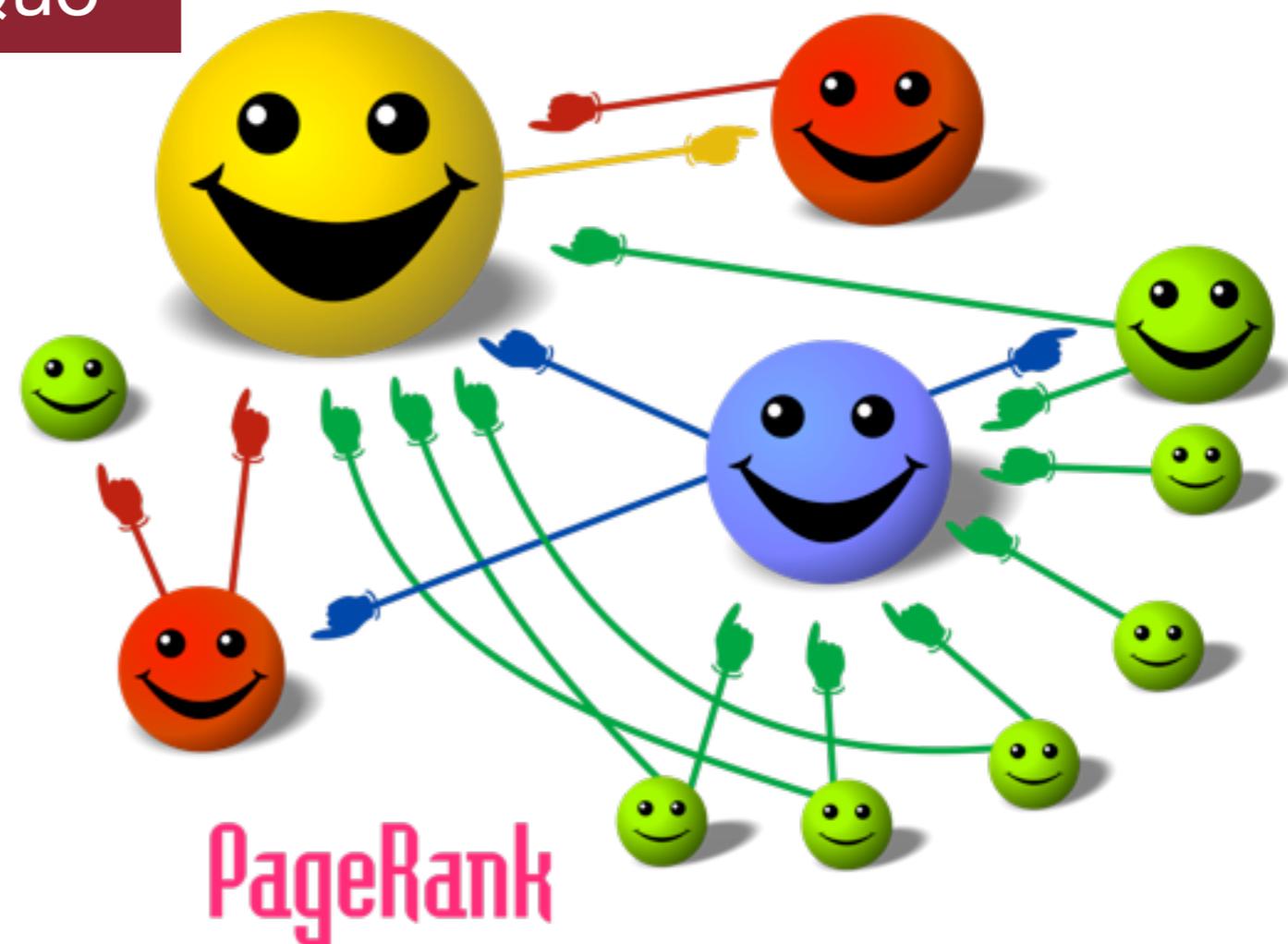
Angela Merkel

Bundeskanzlerin, CDU-Vorsitzende, Angela Kasner ...

Internetsuche und Google Page-Rank

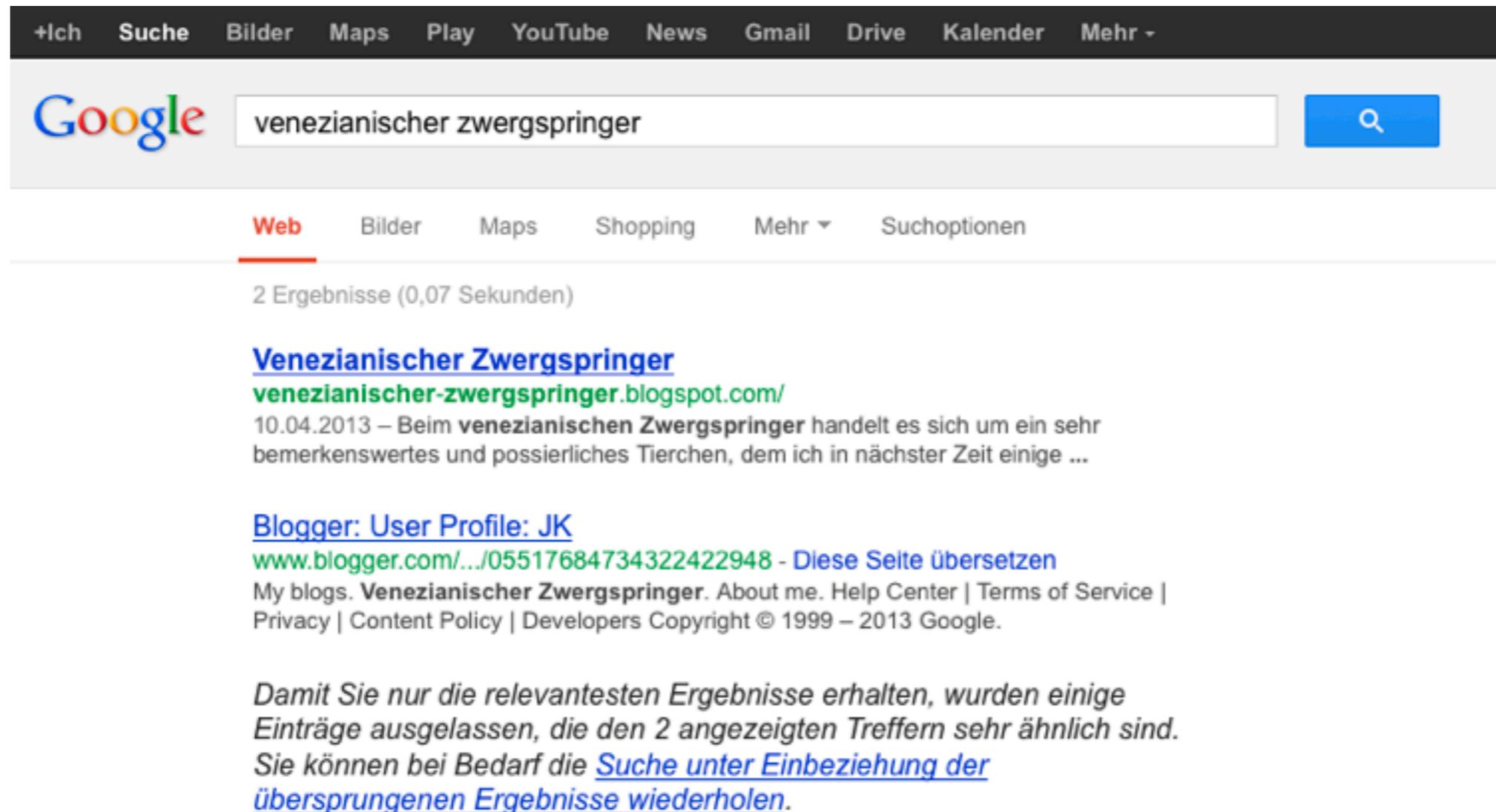
15

1. Wiederholung
2. Term- und Dokumentenranking
3. SEO-Experiment - Status Quo



Projekt: Suchmaschinenoptimierung

16



The screenshot shows a Google search interface. At the top, there is a navigation bar with links for '+Ich', 'Suche', 'Bilder', 'Maps', 'Play', 'YouTube', 'News', 'Gmail', 'Drive', 'Kalender', and 'Mehr -'. Below this is the Google logo and a search bar containing the text 'venezianischer zwergspringer'. To the right of the search bar is a blue search button with a magnifying glass icon. Below the search bar, there are tabs for 'Web', 'Bilder', 'Maps', 'Shopping', 'Mehr ▾', and 'Suchoptionen'. The 'Web' tab is selected and underlined. Below the tabs, it says '2 Ergebnisse (0,07 Sekunden)'. The first result is titled 'Venezianischer Zwergspringer' and is a link to 'venezianischer-zwergspringer.blogspot.com/'. The snippet below the link reads: '10.04.2013 – Beim venezianischen Zwergspringer handelt es sich um ein sehr bemerkenswertes und possierliches Tierchen, dem ich in nächster Zeit einige ...'. The second result is titled 'Blogger: User Profile: JK' and is a link to 'www.blogger.com/.../05517684734322422948 - Diese Seite übersetzen'. The snippet below the link reads: 'My blogs. Venezianischer Zwergspringer. About me. Help Center | Terms of Service | Privacy | Content Policy | Developers Copyright © 1999 – 2013 Google.' At the bottom of the search results, there is a note: 'Damit Sie nur die relevantesten Ergebnisse erhalten, wurden einige Einträge ausgelassen, die den 2 angezeigten Treffern sehr ähnlich sind. Sie können bei Bedarf die [Suche unter Einbeziehung der übersprungenen Ergebnisse wiederholen.](#)'

Venezianischer Zwergspringer

- Webseiten



- » <http://venezianischer-zwergspringer.blogspot.com/> - JK

- noch nicht bei Google indiziert

- » <http://venezianischer-zwergspringer.de.tl/> - Conrad

- » <http://venezianischerzwergspringer.webs.com/> - Swantje



Kontakt:

Nadine Steinmetz & Magnus Knuth

Hasso-Plattner-Institut für Softwaresystemtechnik

Universität Potsdam

Prof.-Dr.-Helmert-Str. 2-3

D-14482 Potsdam

E-Mail: vorname.nachname@hpi.uni-potsdam.de

*Danke für eure
Aufmerksamkeit.*

Suchaufgaben II

Suchaufgaben II

19

- Kultur

Rembrandt malte ein Bild von einem Philosophen, der eine Büste eines griechischen Poeten anschaut. Ein Gold-Medaillon auf der Büste zeigt einen anderen berühmten Griechen. *Wen?*

Suchaufgaben II

19

- Kultur

Rembrandt malte ein Bild von einem Philosophen, der eine Büste eines griechischen Poeten anschaut. Ein Gold-Medaillon auf der Büste zeigt einen anderen berühmten Griechen. *Wen?*

Suchaufgaben II

19

- Kultur

Rembrandt malte ein Bild von einem Philosophen, der eine Büste eines griechischen Poeten anschaut. Ein Gold-Medaillon auf der Büste zeigt einen anderen berühmten Griechen. *Wen?*

- Geschichte

Im April 1896 habe ich einen Mann erschossen. Acht Monate davor hatte dieser einen anderen Mann erschossen, der wiederum 17 Jahre davor 42 Männer erschossen haben soll. *Wen habe ich erschossen?*

Suchaufgaben II

19

- Kultur

Rembrandt malte ein Bild von einem Philosophen, der eine Büste eines griechischen Poeten anschaut. Ein Gold-Medaillon auf der Büste zeigt einen anderen berühmten Griechen. *Wen?*

- Geschichte

Im April 1896 habe ich einen Mann erschossen. Acht Monate davor hatte dieser einen anderen Mann erschossen, der wiederum 17 Jahre davor 42 Männer erschossen haben soll. *Wen habe ich erschossen?*

Suchaufgaben II

19

- Kultur

Rembrandt malte ein Bild von einem Philosophen, der eine Büste eines griechischen Poeten anschaut. Ein Gold-Medaillon auf der Büste zeigt einen anderen berühmten Griechen. *Wen?*

- Geschichte

Im April 1896 habe ich einen Mann erschossen. Acht Monate davor hatte dieser einen anderen Mann erschossen, der wiederum 17 Jahre davor 42 Männer erschossen haben soll. *Wen habe ich erschossen?*

- Geographie

Die zwei Länder, auf deren Territorium früher die Hamangia lebten, sind durch einen Grenzfluss voneinander getrennt. *Wie lang ist dieser Fluss insgesamt?*