EINFOSE

http://einfose.ffos.hr/

PRINZIPIEN DES INFORMATION SEEKING UND DES INFORMATION RETRIEVAL

THOMAS MANDL,
UNIVERSITÄT HILDESHEIM

Erasmus+

AGENCIJA ZA MOBILNOST I PROGRAME EU

SVEUČILIŠTE J. J. STROSSMAYERA U OSIJEKU
FILOZOFSKI FAKULTET
FACULTY OF HUMANITIES AND SOCIAL SCIENCES

Kurs Beschreibung

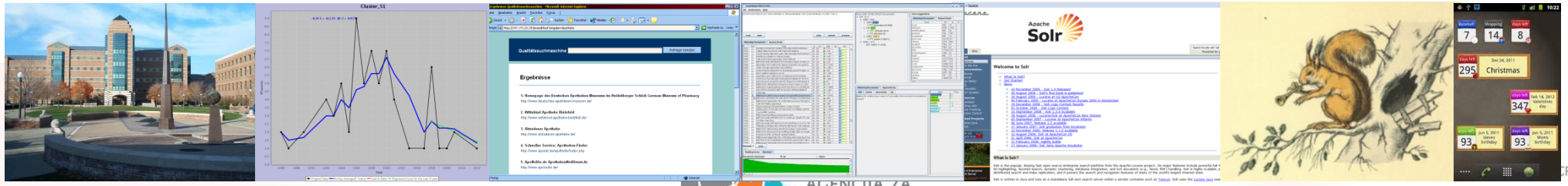
- PRINCIPLES OF INFORMATION SEEKING AND RETRIEVAL

- Kurs besteht aus fünf Teilen

- Einführung in Information Retrieval, Systeme und Ranking (Thomas Mandl)
- Search Process (Andres Sule)
- Organisation of Information in IR (Jan Pisanski)
- User Behavior (Polona Vilar)
- Collaborative Information Seeking and Retrieval (Stefanie Elbeshausen)

Definition

Information Retrieval (IR) deals with the search for information and with the representation, storage and organisation of knowledge.



Definition

„Information retrieval covers the problems relating to the effective storage, access and searching of information required by individuals“ (Ingerwersen 1992)

Definitionen

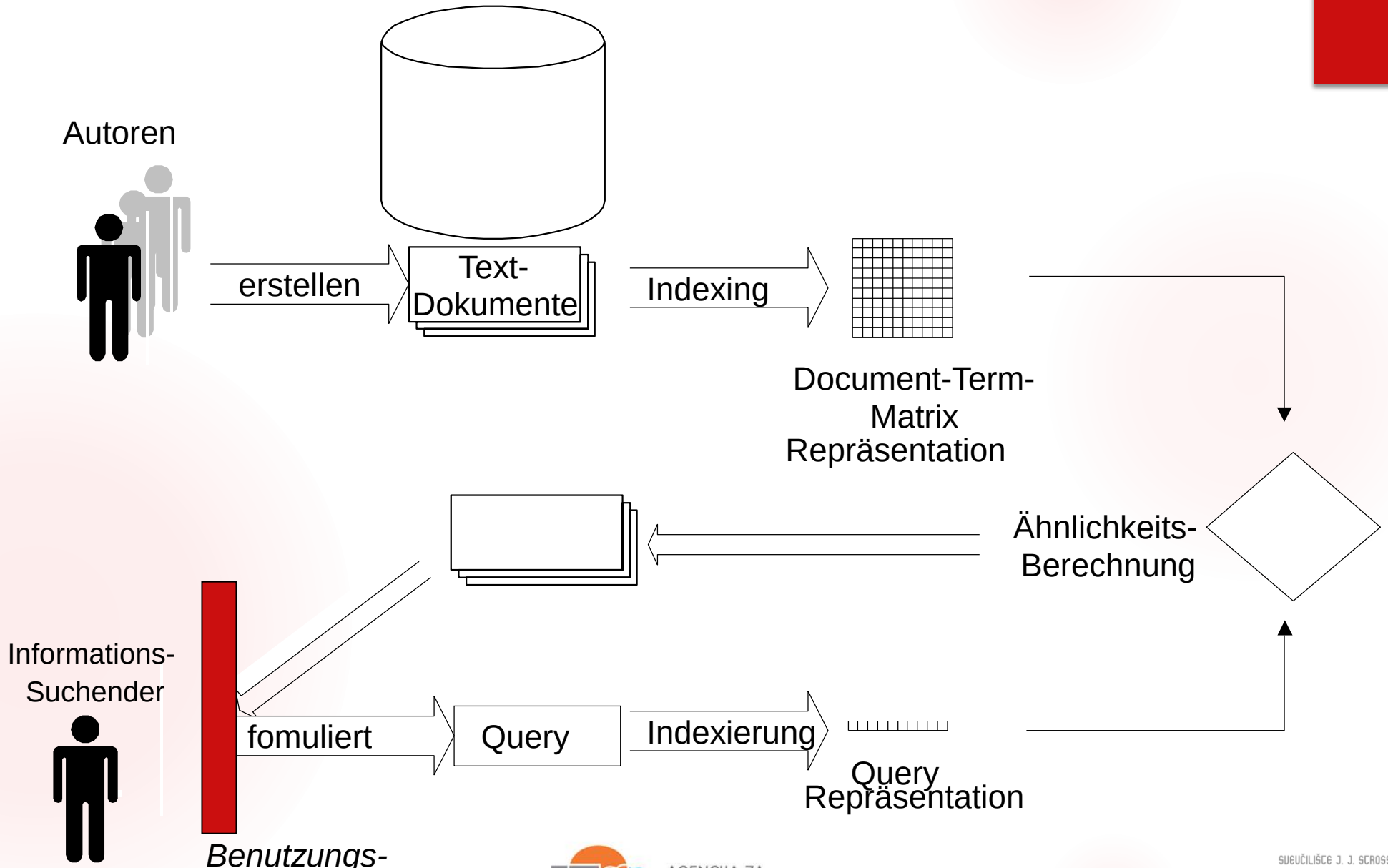
„leading the user to those documents that will be best enable him/her to satisfy his/her need for information“

(Robertson 1981, p.10)

„the goal of an information retrieval system is for the user to obtain information from the knowledge resource which helps her/him in problem management“

(Belkin 1984)

IR Prozess

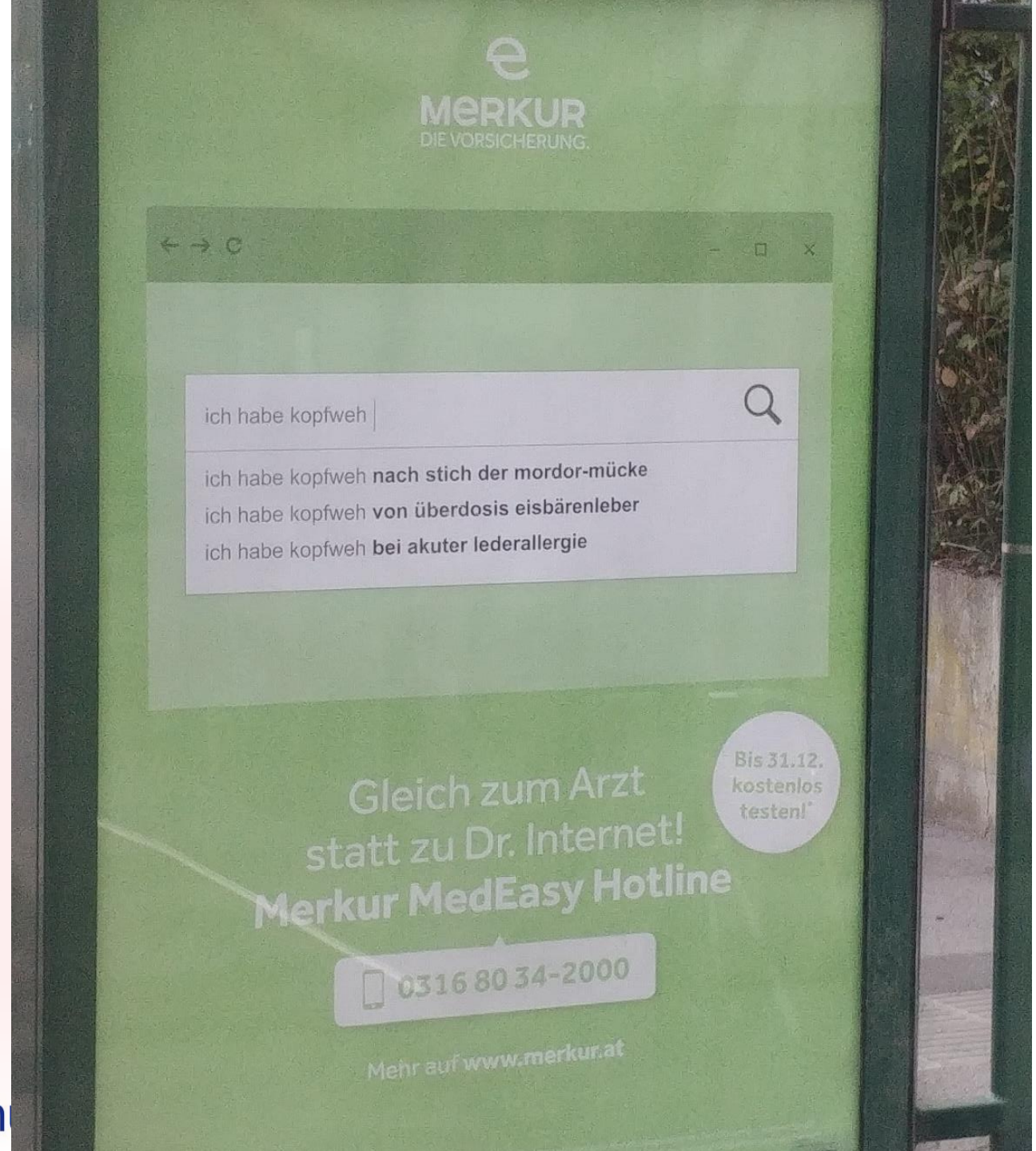


Technologie

- Müssen wir Technologie überhaupt verstehen, um sie zu nutzen?
  - Z.B. Auto, Flugzeug, Medizin?

Einfluss von Suchmaschinen

- Gesellschaftliche Relevanz
  - Bestehen Gefahren?
  - Regulierung durch Staaten und Gesetze


MERKUR
DIE VERSICHERUNG.
ich habe kopfweh
ich habe kopfweh nach stich der mordor-mücke
ich habe kopfweh von überdosis eisbärenleber
ich habe kopfweh bei akuter lederallergie
Gleich zum Arzt
statt zu Dr. Internet!
Merkur MedEasy Hotline
Bis 31.12. kostenlos testen!*
0316 80 34-2000
Mehr auf www.merkur.at

Rolle Information Professionals

- Verständnis von Such-Technologie
- Ergebnisse und Grenzen erkennen und erklären können
  - Z.B. für Benutzer und Manager
- Such-Technologie optimal nutzen
- Such-Technologie verbessern

Repräsentation

- Wie Sortiert eine Suchmaschine ihr Ergebnis?
  - Warum steht ein Dokument an Position 1 und warum andere darunter?

Repräsentation von Dokumenten

- Repräsentation der wichtigsten Merkmale und des Inhalts eines Dokuments
  - z. B. ein Textdokument durch einige Begriffe (Worte)
  - z. B. ein Geschäftsprozesses durch einige Einträge in einer Datenbank
  - z.B. ein Bild durch Tags oder seine Farbverteilung

Kernprobleme im IR

Homo-nyme

- *Bank*
  - *Finanz-Institution oder Möbelstück*

Syno-nyme

- *Geld, Kapital*

Inhaltsrepräsentation im IR

- Indexierung
  - Manuelle Indexierung
  - Automatische Indexierung
- Abstracting
- Clustering

Manuelle Indexierung

- Auch bezeichnet als: Intellektuelle Indexierung
- Informationsexperten weisen Objekten Repräsentations-Elemente zu

- Beispiele?
- Wie funktierung diese Indexierung? Wie läuft sie ab?

- Intellektuelle Indexierung ist noch weit verbreitet(auch wenn im Internet häufig automatische Indexierung eingesetzt wird):
  - Meta-Daten
  - Bibliotheken, Informationsinfrastruktur-Einrichtungen (Fachinformationszentren)
  - Nachrichten-Agenturen (Reuters oder dpa)
  - Firmen-Archive
  - Patent-Ämter
  - Tagging (in Social Media wie Facebook)

Manuelle Indexierung

- Informationsexperten weisen Objekten Repräsentations-Elemente zu
  - Selektion der Repräsentations-Elemente (Terme, Schlagwörter)
  - So spezifisch wie notwendig
  - So vollständig wie möglich
  - Kenntnis von Werkzeugen und Wissensquellen erforderlich
  - Quantitä vs. Qualität
  - Nutzungskontext muss beachtet werden

Arten von kontrolliertem Vokabular

- Schlagwortliste
  - Sammlung erlaubter Repräsentations-Begriffe
- Thesaurus
  - systematische Sammlung
  - Zusätzliche Information (Oberbegriff, Unterbegriffe, Synonyme ...)
- Klassifikation
  - Hierarchische Struktur (Yahoo, Bibliothekskatalog)
- Ontologie
  - Beschreibung der Eigenschaften des Objekts und deren erlaubter Werte
  - Datenmodellierung
  - Erlaubt logisches Schließen

Automatische Indexierung

- Software erstellt Repräsentation für ein Wissensobjekt (Dokument)
  - Auswahl der angemessenen Repräsentations-Elemente (oft Wörter)
  - Selten: Auswahl aus einem kontrolliertem Vokabular

„Bag of words“

- Document wird von wenigen Wörtern repräsentiert
  - Kontext geht verloren
  - Inhalt eines Dokuments kann nicht vollständig wiederhergestellt werden, wenn nur die Repräsentation betrachtet wird
- Weshalb?
- Was ist mit Syntax und Semantik?

Stopwort-Eliminierung: Standard

- Artikel, Pronomen, Präpositionen, …
- Diese Wörter enthalten keinen Inhalt, wenn sie alleine stehen
- Sind die häufigsten Wörter in einer Text-Kollektion
- Ihre Streichung reduziert eine Kollektion meist um 30% bis 50% (in ihrer Größe)
- Umfasst einige hundert Wörter
- je nach Sprache unterschiedlich

Linguistische Vorverarbeitung: Stammformreduktion (Stemming)

- Query:
  - lesen
- Dokument
  - lese
  - gelesen
  - las

Information Retrieval und Database Retrieval

- Wie unterscheiden sich Databank Retrieval (z.B. die Anwendung von SQL) von IR?

Paradigmen

- Exact Match
  - Genaue Übereinstimmung mit Suchbedingungen
- Partial Match
  - Bestmögliche Übereinstimmung mit Suchbedingungen

Paradigmen

- Exact Match
  - Genaue Spezifizierung der Bedingungen für die Suche
  - Eine Menge wird beschrieben
  - Alle Dokumente, welche diese Bedingungen erfüllen, sind in der Ergebnismenge
  - Reihenfolge der Dokumente wird durch Meta-Daten defniert
  - Boole‘sche Logik wird eingesetzt
  - Typische Beispiele:
    - Bibliothekskatalog
    - Datei-Suche
    - Bibliographische Suche

Beispiel

Paradigmen

- Partial Match
  - Vage Beschreibung der gewünschten Inhalte
  - Reihenfolge gibt die Wahrscheinlichkeit für die Relvanz an
  - Nutzer entscheidet, wie viele Treffer sie/er betrachtet
  - Keine Syntax muss erlernt werden
  - Typische Beispiele:
    - Web-Suchmaschinen
    - Meist unternehmensweite Suchdienste

Bai du 百度

样读

百度一下

网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»

百度为您找到相关结果约24,100,000个

搜索工具

这个烽字怎么样读_百度知道

2个回答 - 提问时间: 2015年12月21日

最佳答案: 读音【xīng】 康熙字典介绍 【已集中】 【火部】 ·康熙笔画:14·部外笔画:10 《集韵》思营切,音騂。《博雅》赤也。《类篇》作。

https://zhidao.baidu.com/quest...

60700800该样读	1个回答	2016-09-03
2400002603怎么样读	2个回答	2016-09-29
我不知道怎么读怎么样读	2个回答	2018-02-18

更多知道相关问题>>

如何像读书一样读人_百度百科

2017年1月11日 - 《如何像读书一样读人-微动作读心术》是2013年新世界出版社出版的图书,作者是亨利·卡莱罗 , 杰勒德·尼伦伯格 , 加布里埃尔·格雷森。...

https://baike.baidu.com/item/如... V3 - 百度快照

怎么读书读书读什么?怎么样读才叫读懂。怎么才能看出... 爱问知识人

2018年1月25日 - 读书读什么?怎么样读才叫读懂。怎么才能看出书中的精华 w627287572 | 举报 报告...创新读书三部曲 拿到一本书,认真读三遍,每一遍采用不同的方法和眼光...

https://iask.sina.com.cn/b/159... - 百度快照

其他人还搜

出师表

三国

相关汉字

忿

怒

相关书籍

Google Play

Graz

Apps

Suchergebnisse Android-Apps Alle Preise Alle Bewertungen

Meine Apps

Einkaufen

Spiele

Familie

Empfehlungen

Konto

Meine Abos

Code einlösen

Apps

Graz Offline Stad
Topobyte.de

Graz App
CITYAPP.WIEN

qando Graz
Fluidtime

BusBahnBim
Verkehrsauskunft Ös

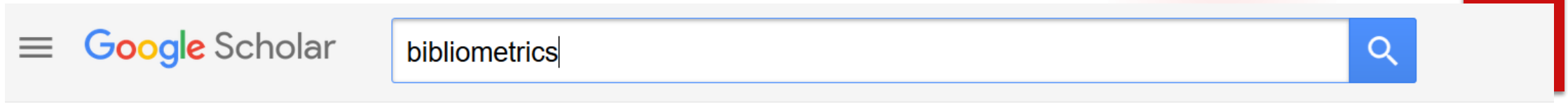
Graz Reiseführer
tripwolf

Graz CityGuide

Erasmus+

AGENCIJA ZA MOBILNOST I PROGRAME EU

SVEUČILIŠTE J. J. STROSSMAYERA U OSIJEKU
FILOZOFSKI FAKULTET
FACULTY OF HUMANITIES AND SOCIAL SCIENCES



Artikel Ungefähr 549 Ergebnisse (**0,06** Sek.) Mein Profil

Beliebige Zeit
Seit 2018
Seit 2017
Seit 2014
Zeitraum wählen...

Nach Relevanz sortieren
Nach Datum sortieren

Beliebige Sprache
Seiten auf Deutsch

☑ Patente einschließen
☑ Zitate einschließen

Tipp: Suchen Sie nur nach Ergebnissen auf **Deutsch**. Sie können Ihre Sprache in den Scholar-Einstellungen. festlegen.

[ZITATION] Eine bibliometrische Analyse eines Dokumentlieferdienstes am Beispiel von Subito: Zusammenhang von Zeitschriftennachfrage und-zitationshäufigkeiten
J Gorraiz, C **Schlögl** - Zeitschrift für Bibliothekswesen und …, 2003 - Klostermann
☆ Zitiert von: 10 Ähnliche Artikel

A **bibliometric** analysis of pharmacology and pharmacy journals: Scopus versus Web of Science **[PDF]** rclis.org
J Gorraiz, C **Schloegl** - Journal of Information Science, 2008 - journals.sagepub.com
Our study examines the suitability of Scopus for bibliometric analyses in comparison with the Web of Science (WOS). In particular we want to explore if the outcome of bibliometric analyses differs between Scopus and WOS and, if yes, in which aspects. Since journal …
☆ Zitiert von: 74 Ähnliche Artikel Alle 14 Versionen

Information and knowledge management: dimensions and approaches. **[PDF]** ed.gov
C **Schlögl** - Information Research: An International Electronic …, 2005 - ERIC
Introduction. Though literature on information and knowledge management is vast, there is much confusion concerning the meaning of these terms. Hence, this article should give some orientation and work out the main aspects of information and knowledge …
☆ Zitiert von: 91 Ähnliche Artikel Alle 2 Versionen »

IR Prozess

- Sammeln der Dokumente
  - Im Web: Crawling
- Analyse der Dokumente
  - Stammformreduktion
  - Indexing -> Repräsentation
- Verarbeitung von Queries
  - Analyse der Queries
  - Queries mit den Dokument-Repräsentationen vergleichen

Stemming

IR Modelle

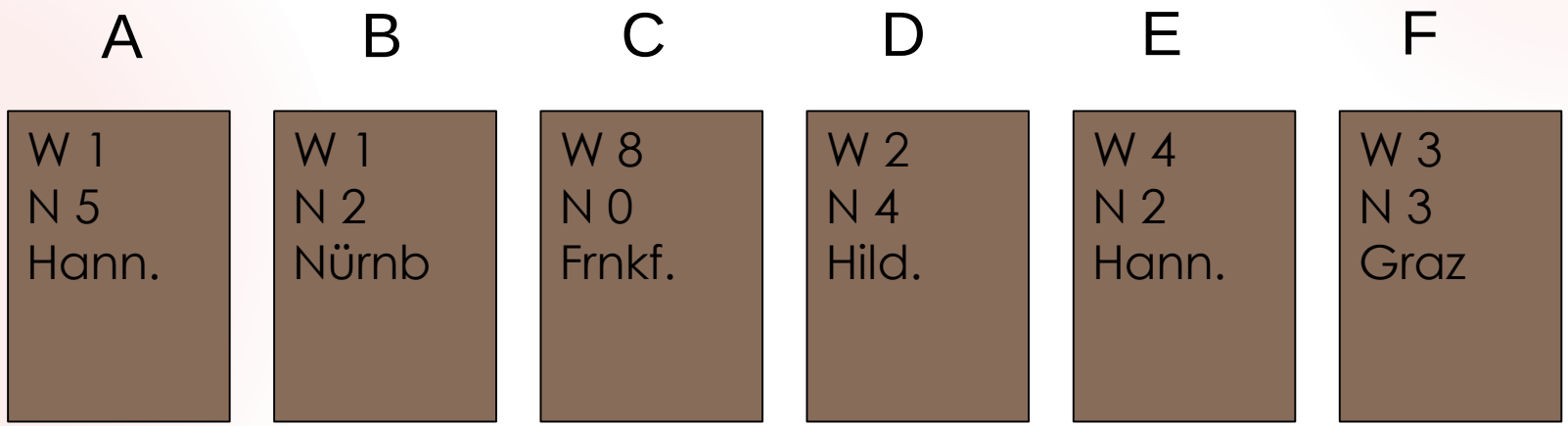
Erasmus+
AGENCIJA ZA MOBILNOST I PROGRAME EU
SVEUČILIŠTE J. J. STROSSMAYERA U OSIJEKU
FILOZOFSKI FAKULTET
FACULTY OF HUMANITIES AND SOCIAL SCIENCES

Was muss über die Wörter in einem Dokument bekannt sein?

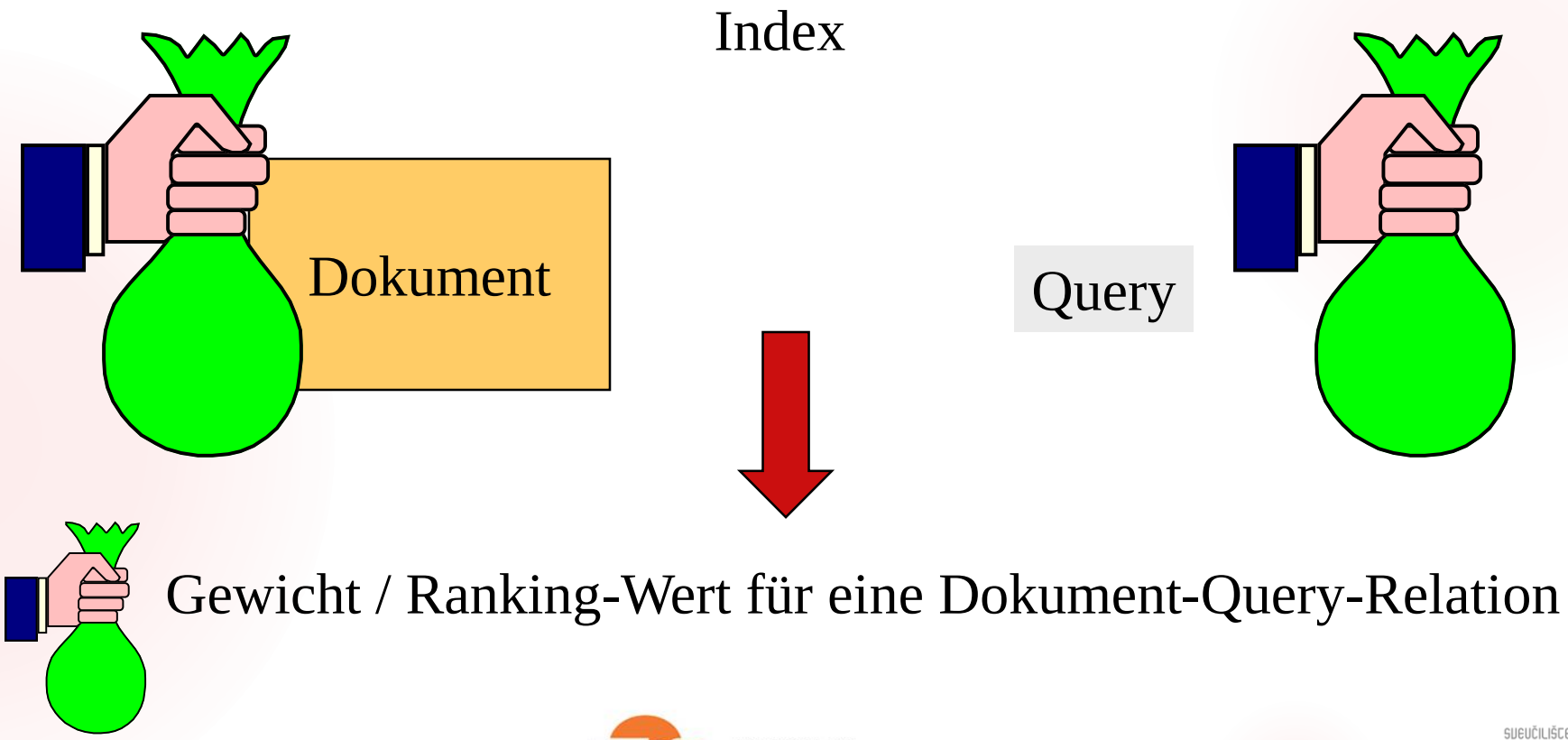
- Welche Information könnte für das Ranking sinnvoll sein?

Beispiel

- Query:
  - Wolf Niedersachsen



Grundidee: Gewichtung



34

Gewichtung

- Wie gut repräsentiert ein Term ein Dokument?

Grundidee

- „It is here proposed that the frequency of word occurrence in an article furnishes a useful measurement of word significance“ (Luhn nach Rijsbergen 1999)

IR und Grundkonzepte

- IR zählt Wörter!
  - In Dokumenten
  - In Kollektionen

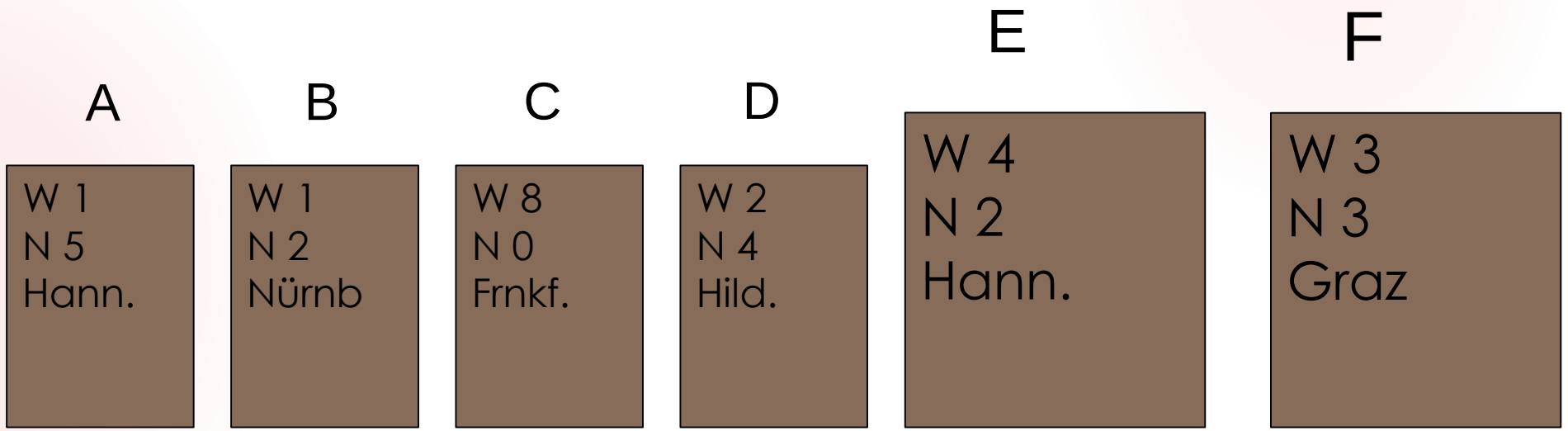
Term-Frequenz

- Formel: TF (t, d) = N

  N wie oft tritt Term t
  in Dokument d auf?

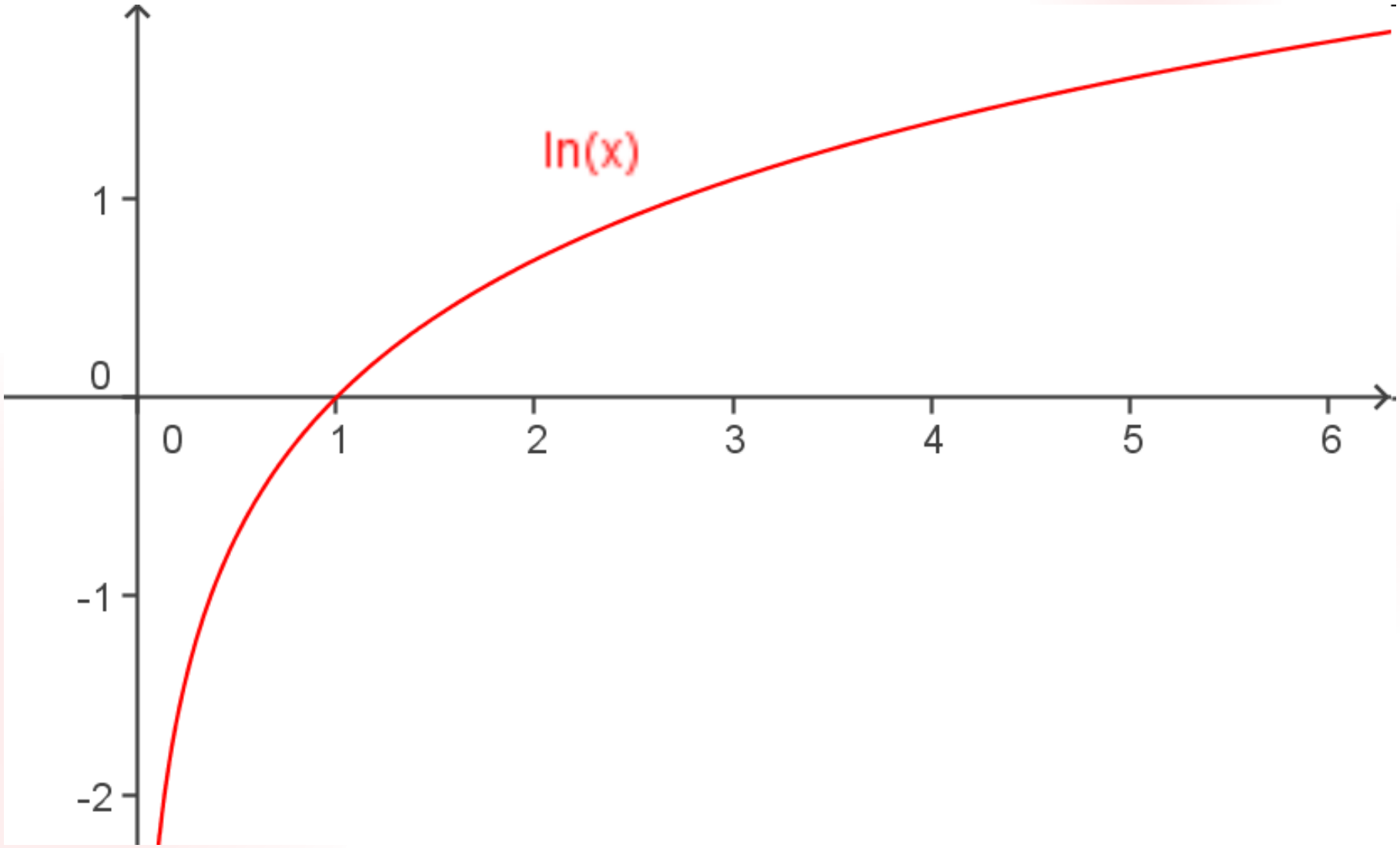
Beispiel

- Query:
  - Wolf Niedersachsen



Term-Frequenz (TF)

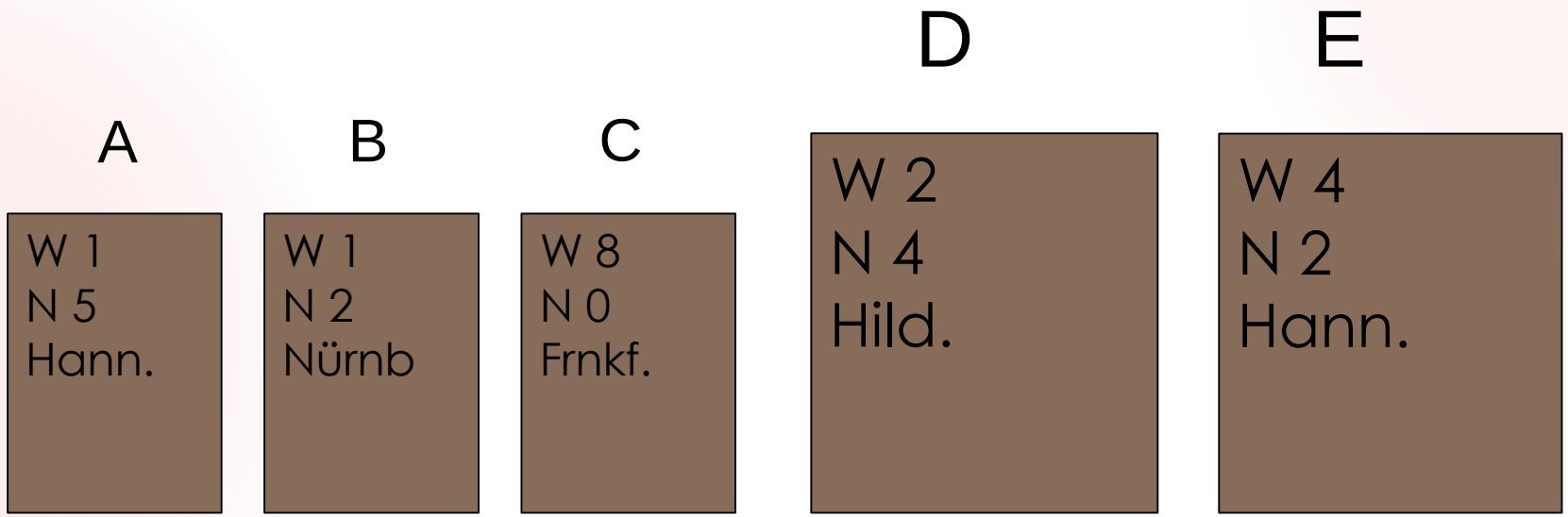
– Soll jedes Auftreten den gleichen Beitrag zur Gewichtung liefern?



http://exbook.de/wp-content/uploads/2008/03/ln_funktion.png

Beispiel

- Query:
  - Wolf Niedersachsen

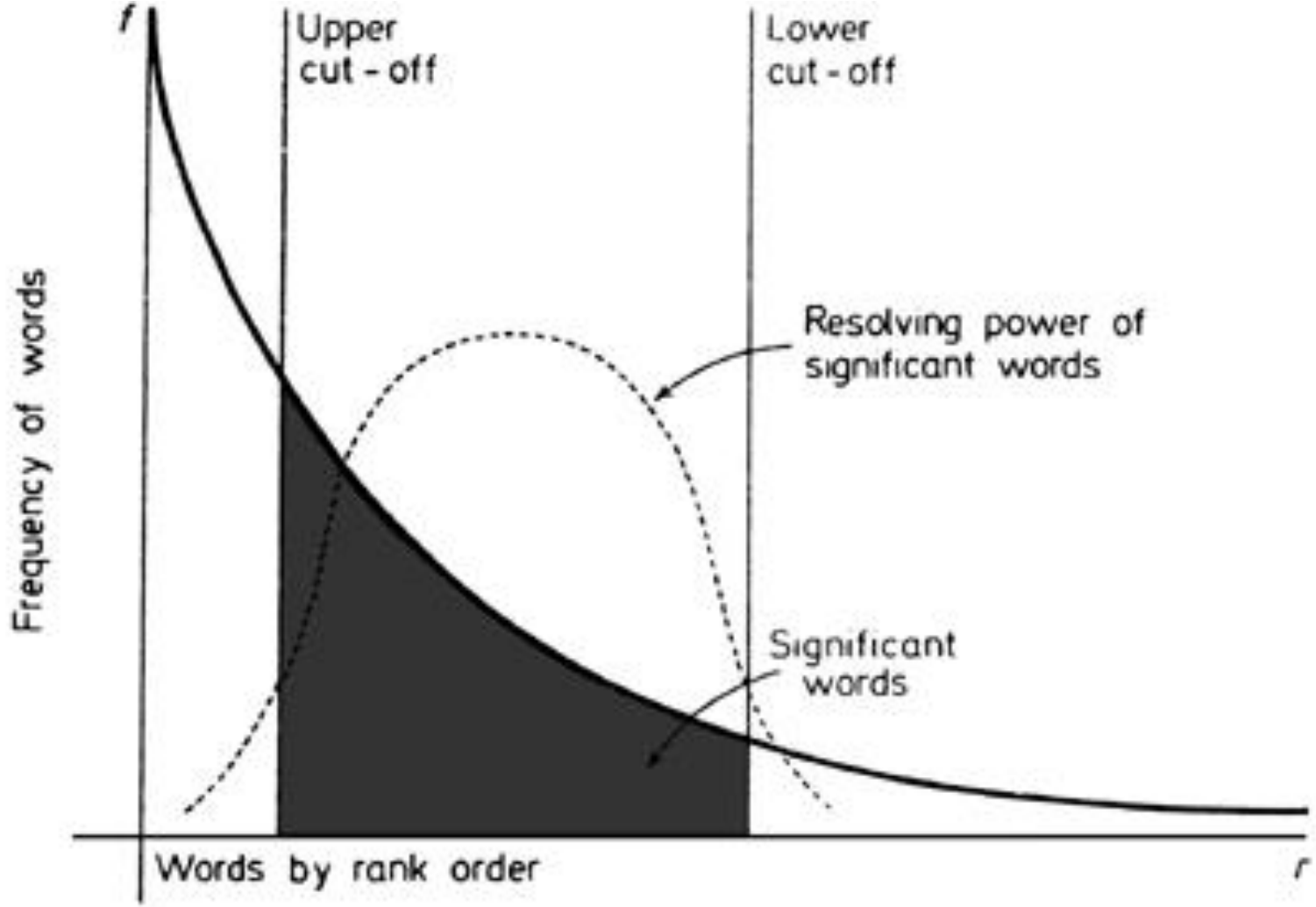


- Unterschiede zwischen Wörtern?
- Größe des Vokabulars
- Verteilung der Wörter in einer Sprache?

Berechnung der Gewichte

- Grundidee: Terme und ihr Auftreten (Frequenz und Verteilung)
- Welche Terme sind „gute“ Indikatoren für die Modellierung von Dokument-Inhalten?
  - Hochfrequente Terme?
  - Terme mit niedriger Frequenz ?
  - Terme mit mittlerer Frequenz ?

Zipf‘sche Verteilung



in: Rijsbergen 1999[2]

Inverse Dokument-Frequenz

- Inverse Dokument-Frequenz (IDF)
- Formel: IDF (Term t) = N/n

  wobei N Anzahl der Dokumente in der Kollektion

  n Anzahl der Dokumente die t enthalten

- Was ist der Effekt dieser Parameter?
- Ist das sinnvoll?

TF IDF Gewichtung

- Term Frequenz * Inverse Dokument Frequenz

- Oft genutzt
  - Nutzt Logarithmus
  - OK als erster Versuch einer Gewichtung
  - Nicht aktueller State-of-the-art
  - Es gibt nicht die eine IDF
  - Viele verschiedene Definitionen existieren

TF.IDF

$$\text{Gewicht} = \log(tf) * \log N/n$$

Vereinfachung

- Anstatt des Logarithmus kann die Harmonische Summe als Näherung genutzt werden

Harmonische Summe

- Harmonische Summe
- Ähnlich zum Logarithmus log (n)

$$H_n = \sum_{k=1}^{n} \frac{1}{k} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n}$$

Harmonische Summe

$$H_1 = 1$$

$$H_2 = \frac{3}{2} = 1{,}5$$

$$H_3 = \frac{11}{6} = 1{,}8\overline{3}$$

$$H_4 = \frac{25}{12} = 2{,}08\overline{3}$$

$$H_5 = \frac{137}{60} = 2{,}28\overline{3}$$

$$H_6 = \frac{49}{20} = 2{,}45$$

$$H_7 = \frac{363}{140} = 2{,}59\overline{285714}$$

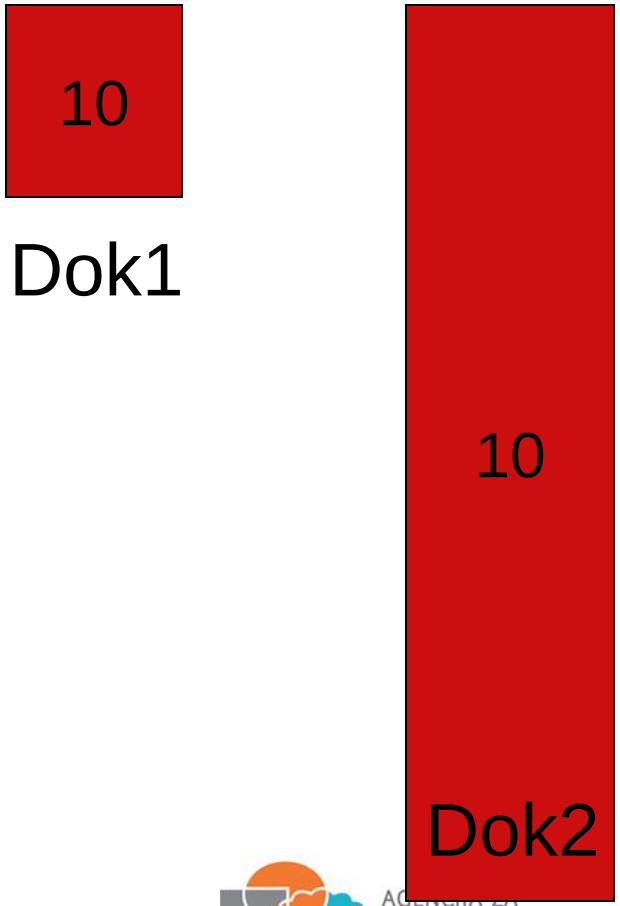
$$H_8 = \frac{761}{280} = 2{,}717\overline{857142}$$

$$H_9 = \frac{7129}{2520} = 2{,}828\overline{968253}$$

$$H_{10} = \frac{7381}{2520} = 2{,}928\overline{968253}$$

Problem verschiedener Längen

Term-Frequenz



Gleiches Term-Gewicht?

Dokument-Länge

kurzes Dokument

langes Dokument

Dokument-Term-Matrix	Haus	Bank	Geld	Park
Dok A	1	2	2	0
Dok B	9	6	3	0
Dok C	0	8	0	6
Dok D	0	6	0	7

Lösung: Längen-Normalisierung

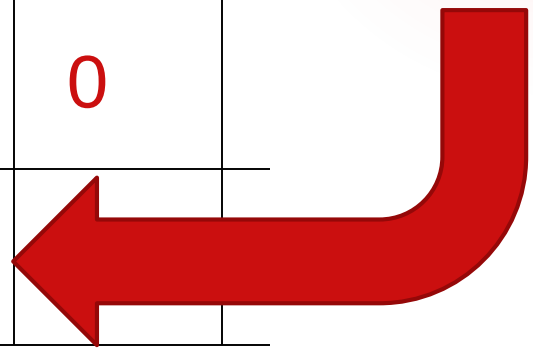
Ziel: alle Dokumente haben die gleiche Länge (und damit den gleichen Einfluss und die gleiche Chance, gefunden zu werden)

Ansatz: Division aller Term-Gewichte durch die Länge des of Dokuments

$$\text{normWeight} = w / \sqrt{\sum_{\text{Vector } i} (w_i)^2}$$

Dokument-Länge: Normalisierung

Dokument-Term-Matrix	Haus	Bank	Geld	Park	Länge: Summe der Terme
Doc A	1	2	2	0	5
Doc B	8	5	2	0	15
Doc $A_{norm.}$	1/5	2/5	2/5	0	
Doc $B_{norm.}$	8/15	5/15	2/15		



Länge für die Normalisierung
-> neue berechnete Länge = 1

Längen-Normalisierung

- Jetzt haben alle Dokumente die gleichen Chance, gefunden zu werden

Wirklich?

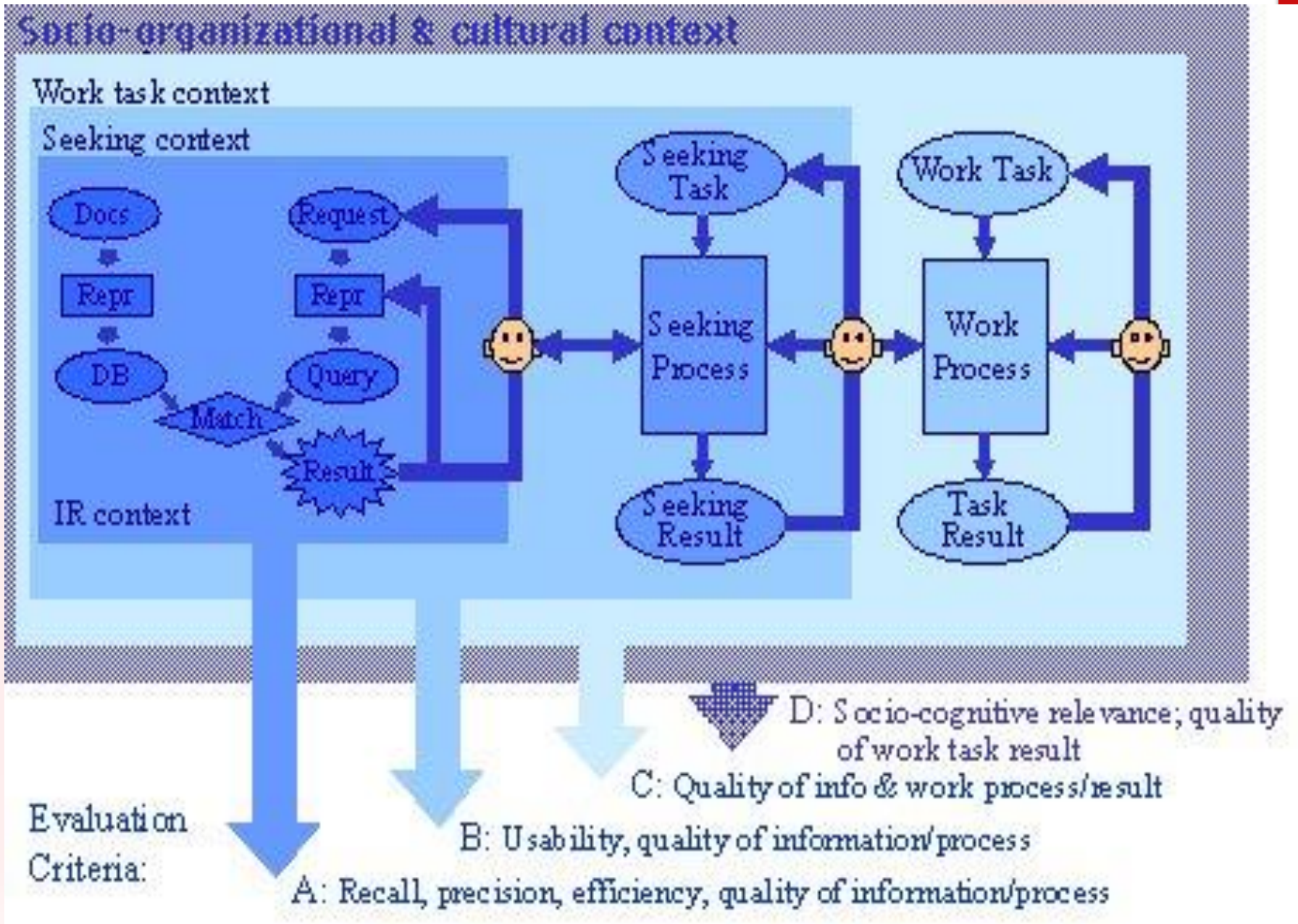
Dokument-Länge: Normalisierung

Dokument-Term-Matrix	Haus	Bank	Geld	Park	Länge: Summe der Terme
Doc A	1	2	2	0	50
Doc B	80	50	2	0	10000
Doc $A_{norm.}$	1/50	2/50	2/50	0	
Doc $B_{norm.}$	0,008	0,005	0,0002	0	

Längen-Normalisierung

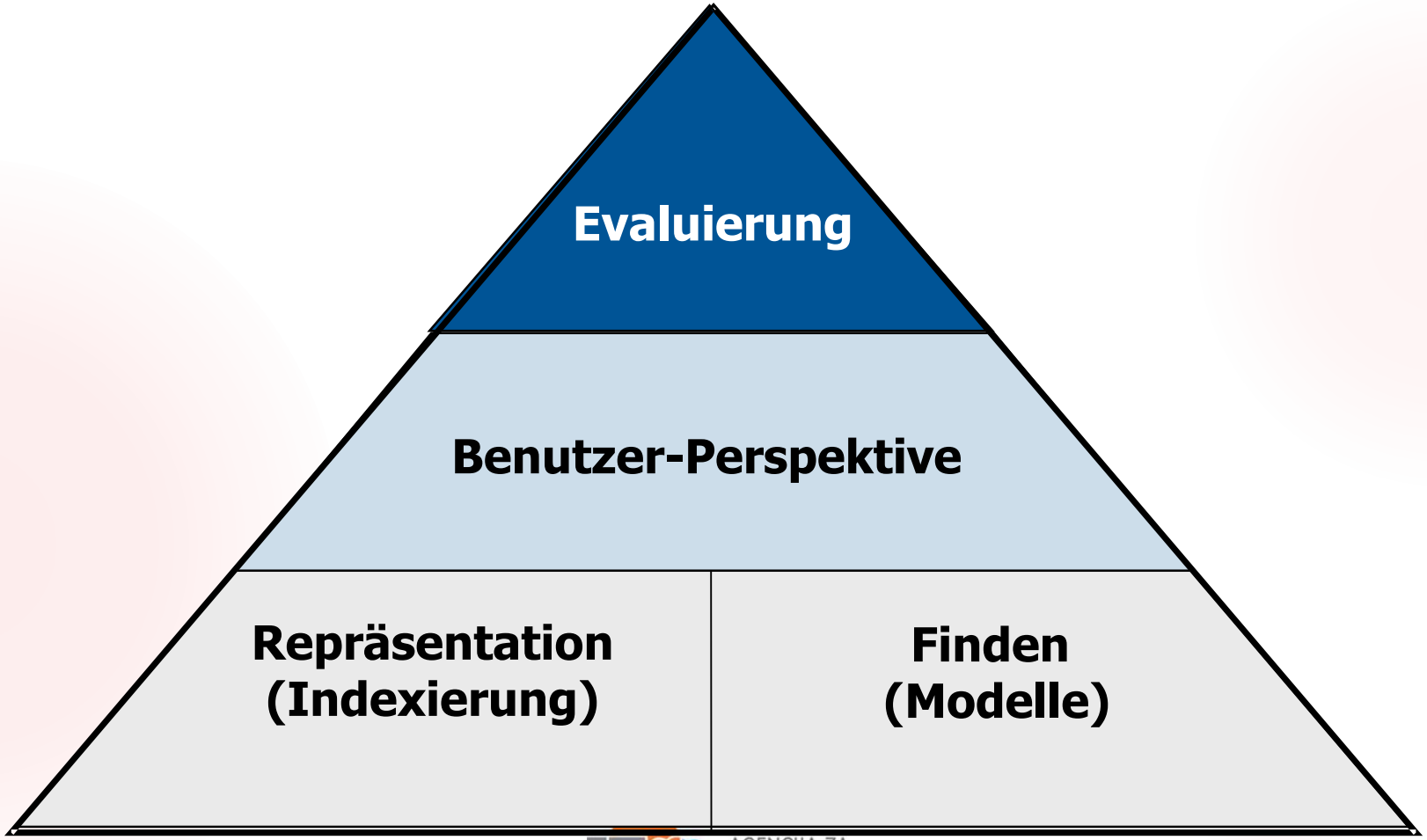
- Aber: Lange Dokumente enthalten mehr Information
- Lange Dokumente verlieren mehr Gewicht für jeden ihrer Terme
- Sie werden also seltener als Treffer auftreten als kurze Dokumente

- Vielleicht zu viel?
- Die optimale Balance zwischen keiner und zu viel Längen-Normalisierung ist schwer zu finden



Järvelin & Ingwersen 2004 http://www.informationr.net/ir/10-1/paper212.html

Information Retrieval

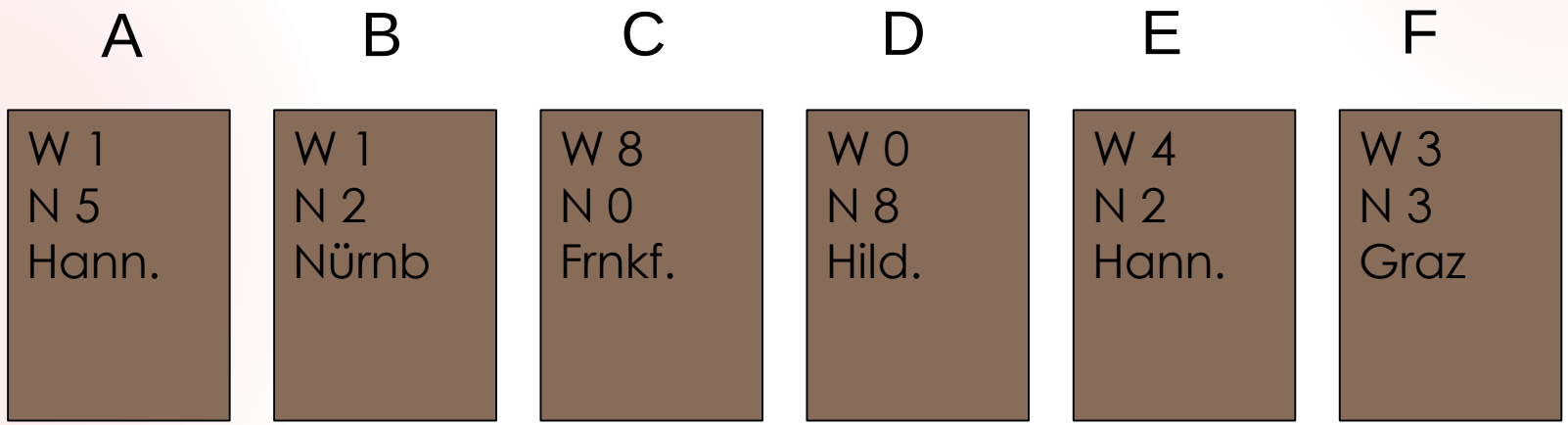


Weitere Faktoren für das Ranking

- Ort der Anfrage
- Link-Muster
- Popularität
- Design (wird mobile Anzeige unterstützt?)
- Wirtschaftliche Interessen
- Diversität der Resultate
- ….

Beispiel

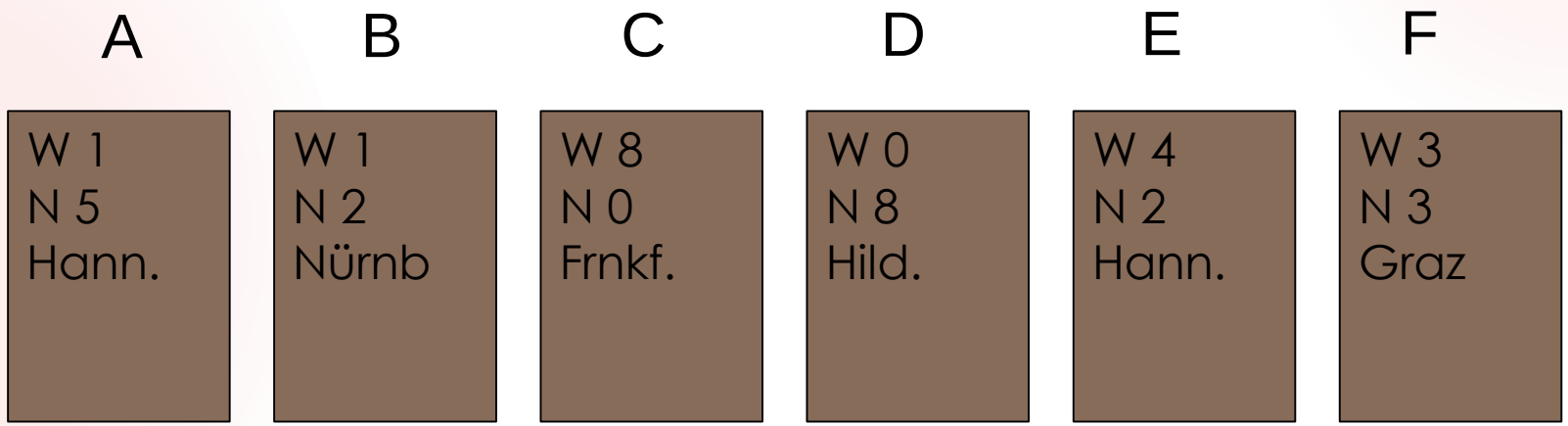
- Annahme: N (Größe der Kollektion) = 10 Milliarden
  - Wolf Idf (W) = 10000 Mio / 510 Mio.
  - Niedersachsen Idf (N) = 10000 Mio / 95 Mio.



Beispiel

- Für Dokument A (vereinfacht ohne Logarithmus)

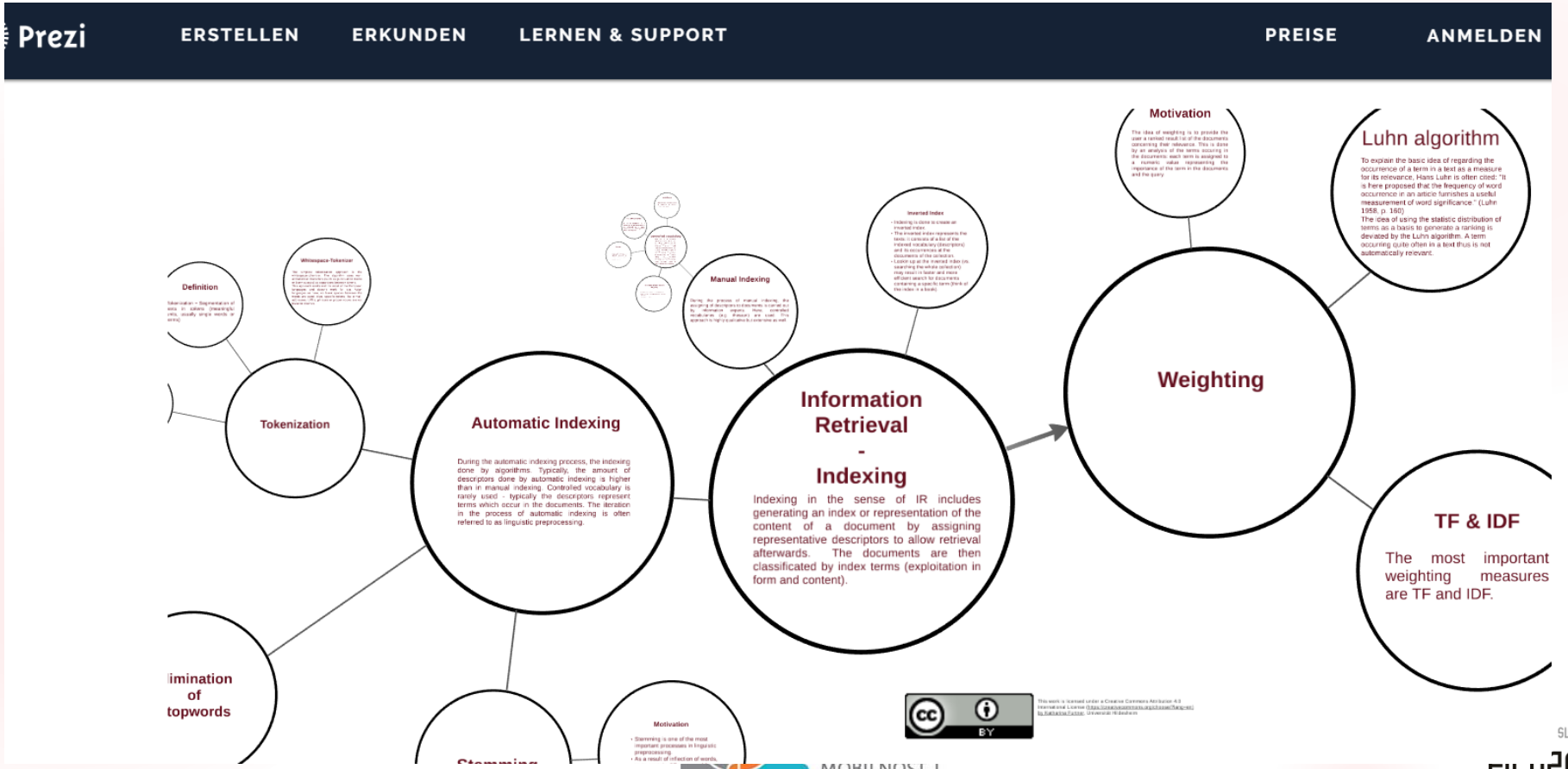
  = (200 * 1) + (100 *5) = ??



Weitere Lernmöglichkeiten

Grundkonzepte des IR (Prezi Resource)

https://prezi.com/o6xpbbuzdvfo/representation-indexing



Kontakt Information für den Lehrenden

mandl@uni-Hildesheim.de

+49 5121 883 30306

http://www.uni-hildesheim.de/~mandl/

Übungen

- Exercise 1: Suchen Sie weitere Ranking Faktoren
- Exercise 2: Berechnen Sie 2 Rankings auf Papier (Extra-Blatt)
- Exercise 3: Suchen Sie Beispiele für partial and exact match Systeme
  - Gehen Sie z.B. die Apps auf Ihrem Mobil-Telefon durch

- 15 Minuten Arbeitszeit