

# Einführung in die Computer- linguistik und Sprachtechnologie

WiSe 2018/2019  
(B-GSW-12)

Udo Hahn



FRIEDRICH-SCHILLER-UNIVERSITÄT  
JENA



Jena University Language and Information Engineering (JULIE) Lab, Germany

<http://www.julielab.de>

# Grundlagen des Information Retrieval

## Sammeln von Dokumentkollektionen vs. Erschließung von Dokumentinhalten

The collage illustrates the process of information retrieval through various media and tools. It includes a 'stern' magazine cover, a 'Disease Management' document, a 'Volks' magazine cover, a document from the 'SRU' (Schweizerischer Rat für Umweltfragen), a screenshot of a 'Bluefish 0.9.5 HTML editor' showing HTML code, a screenshot of the 'Amaya' web browser interface, and a page from the 'Süddeutsche Zeitung' newspaper.

**SRU** Sachverständigenrat für Umweltfragen

Der Vorsitzende  
Prof. Dr.-Ing. Martin Faustlich

Technische Universität München  
Lehrstuhl für Rohstoff- und  
Energietechnologie  
Petersgasse 15  
94155 Straubing  
Tel. 09421 / 187 100  
Fax 09421 / 187 111  
martin.faustlich@tum.de  
www.umwelt.at.de

13. Februar 2009

Kommentare des SRU zum Entwurf des Nationalen Biomasseaktionsplans

Sehr geehrte Frau Dr. Freier, sehr geehrter Herr Dr. Ohlhoff,

die Chancen und Risiken der Nutzung nachwachsender Rohstoffe als erneuerbare Energie-  
träger sind in den letzten Jahren intensiv erforscht und diskutiert worden. Angesichts der  
Komplexität des Themas und der Vielfalt der Nutzungsoptionen ist eine integrierte strategische  
Betrachtung des Sektors dringend geboten. Vor diesem Hintergrund ist es begrüßenswert,  
dass die Bundesregierung mit einem nationalen Biomasseaktionsplan den Vorgaben der EU-  
Biomassestrategie nachkommt. Angesichts der Tatsache, dass sowohl auf der europäischen  
als auch auf der nationalen Ebene die wesentlichen Ziele und Instrumente bereits rechtlich  
fixiert sind, kann ein solcher Plan allerdings nur im Detail nachjustieren.

Begrüßenswert ist, dass der Aktionsplan die verstärkte Erzeugung von Wärme, die Erschlie-  
ßung neuer Biomassepotenziale insbesondere aus Reststoffen und Abfällen, die Sicherung  
der nachhaltigen Erzeugung, die verstärkte Nutzung von Verwertungsoptionen mit beson-  
ders hohem Treibhausgas-Minderungspotenzial, den Vorrang der stofflichen Verwertung,  
sowie eine verstärkte dezentrale Nutzung im Sinne der Entwicklung der ländlichen Räume als  
strategische Ziele fest schreibt. Bedauerlich ist aber, dass er an vielen Stellen entgegen dem  
wissenschaftlichen Erkenntnisstand an der grundsätzlichen Gleichwertigkeit der Biomasse-

**Amaya**

Welcome to Amaya

Address: /usr/bin/\_lib/Amaya/amaya/AmayaPage.html

Title: Welcome to Amaya

Welcome to Amaya

lease 2.4  
t has been  
a fashML,

les and the  
ments  
cuments on  
d edit Web  
simple  
k, you  
ier

ra. By this  
tion, you  
so possible  
s

**Süddeutsche Zeitung**

„Kauft eure Drogen woanders“ – Krawall in Kopenhagen – Panorama

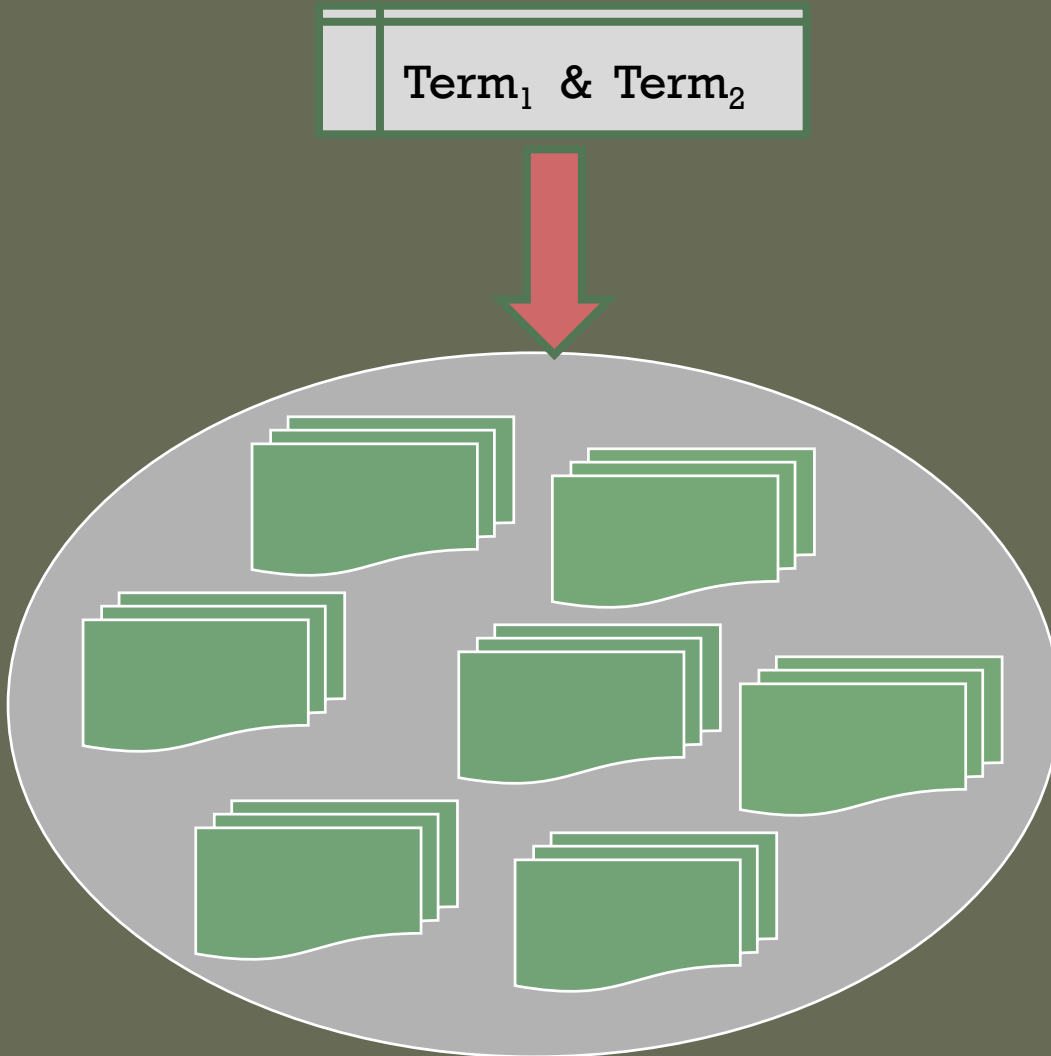
SPD stärkste Partei, AfD liegt vor CDU

Landtagswahl in Brandenburg gewonnen

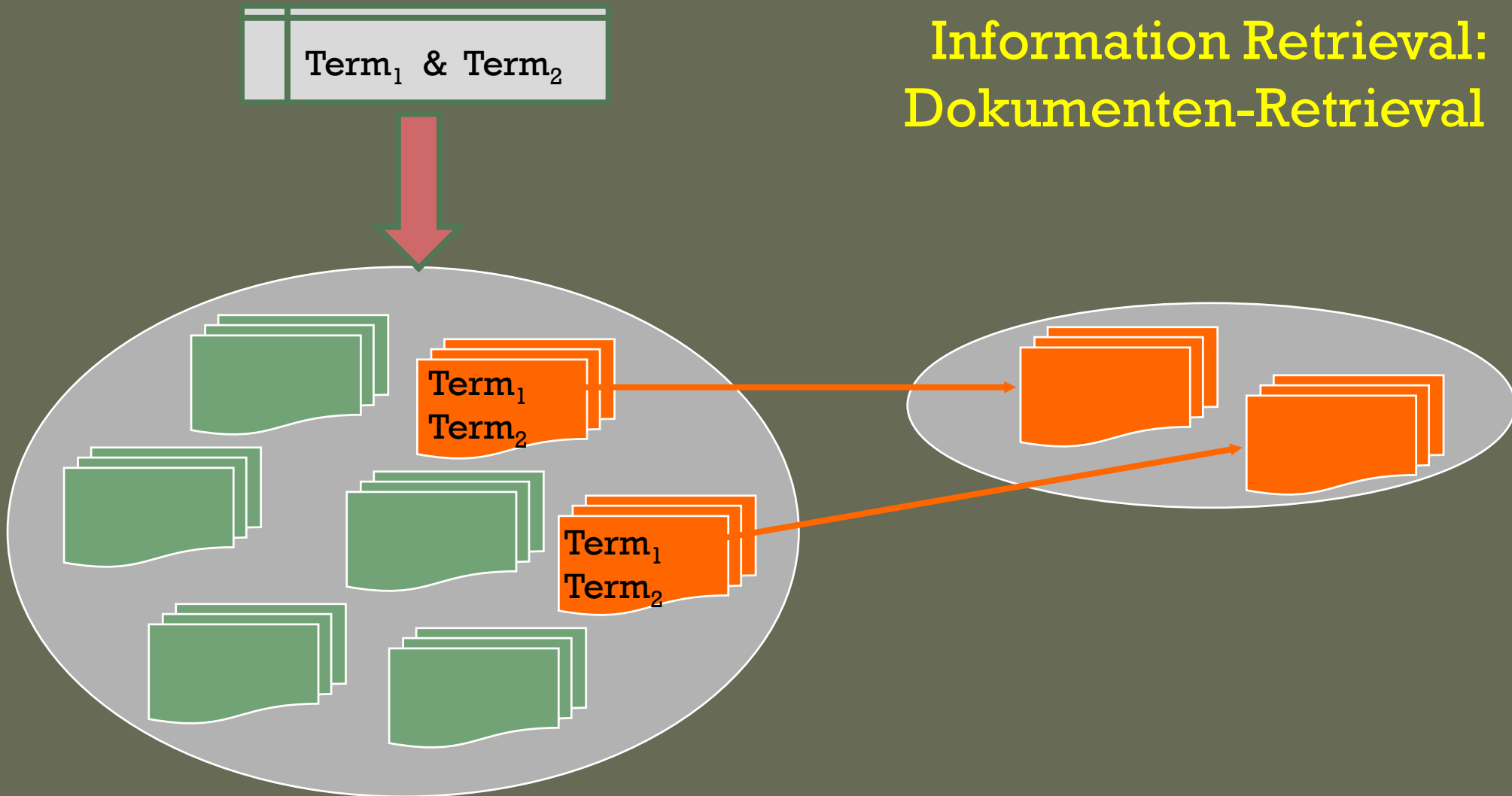
SPD-Präsidium will Zustimmung zu Ceta-Abkommen

2

Information Retrieval:  
Dokumenten-Retrieval



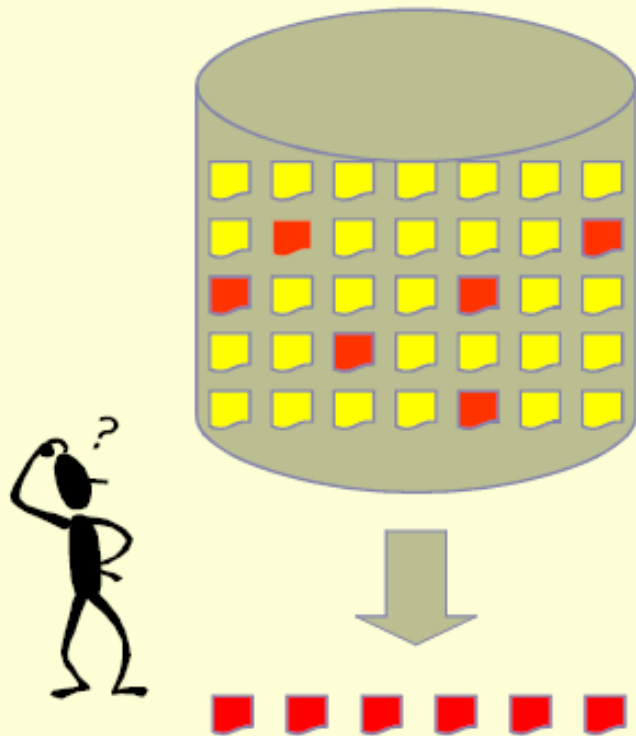
## Information Retrieval: Dokumenten-Retrieval



# Flavors of Information (Document) Retrieval (1/2)

## Ad-hoc retrieval

One time queries (e.g. Web search)



## Filtering/Routing

Constant search profile (e.g. Spam filtering)



# Flavors of Information (Document) Retrieval (2/2)

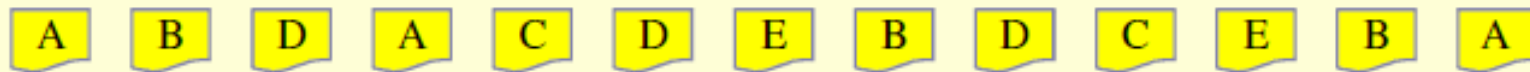
- **Categorization/Clustering:**

Group documents into predefined classes/ adaptive clusters



- **Topic Detection and Tracking:**

Cluster news in stream



# INDEXING

- ◆ Indexing by Derivation

- Index terms are derived from the document (and possibly morphologically normalized)

- ◆ Indexing by Assignment

- Index terms are assigned to a document using an authoritative terminology (usually, a thesaurus)

# INDEX TERMS

- ◆ Nouns (singletons, compounds)
  - Cell, dataset,
- ◆ Noun phrases
  - Hot spot, regulation of cells
- ◆ Avoid too complex terms (pre-coordination)
  - The regulation of cells under laser beam exposure in vitro



# MANUAL INDEXING

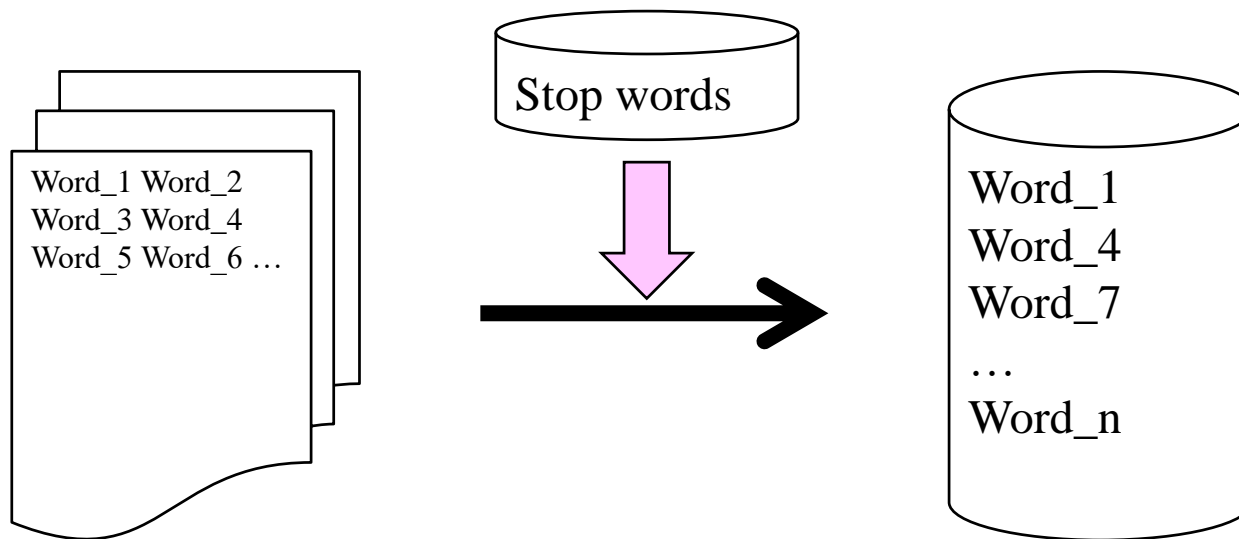
- ◆ Determine main topic(s)
- ◆ What's a relevant issue?
- ◆ Based on human (speed) reading and understanding of the document

# AUTOMATIC INDEXING

- ◆ Absolute vs. relative frequency
  - Per document
  - Relative to document collection
  - Bag-of-words (BOW)

# BAG OF WORDS

- ◆ Eliminate sequential structure of texts



# AUTOMATIC INDEXING

- ◆ Absolute vs. relative frequency
  - Per document
  - Relative to document collection
  - Bag-of-words (BOW)
  - Eliminate stop words (high occurrence frequency!)

# Lexikalische Frequenzanalyse: Stoppwörter höchstfrequent

File Global Settings Tool Preferences About

## Thema 1: Wortschatz

Corpus Files  
Leipzig-Korpus-2  
Leipzig-Korpus-3  
Leipzig-Korpus-4

Concordance Concordance Plot File View Clusters Collocates Word List Keyword List

Hits Total No. of Word Types: 108034 Total No. of Word Tokens: 937245

Rank	Freq	Word	Lemma	Word Forms
1	29691	der	der	
2	27972	die	die	
3	19618	und	und	
4	16046	in	in	
5	11392	den	den	
6	8912	von	von	
7	8599	zu	zu	
8	8043	das	das	
9	7682	mit	mit	
10	7432	sich	sich	
11	7156	ist	ist	
12	7147	auf	auf	
13	6853	im	im	
14	6743	nicht	nicht	
15	6712	für	für	
16	6629	Die	Die	
17	6164	des	des	

Type-Token-Ratio (hier: 108034:937245  $\approx 0,115$ )

Wortliste (mit Rang und Frequenzangabe)

Suche: Frequenzliste aller Wortformen und Type-Token-Ratio in einem Ausschnitt der Leipzig Corpus Collection (Sätze aus Zeitungen).

Search Term ☒ Words ☐ Case ☐ Regex

Display Options ☐ Treat all data as lowercase

Start (kein Suchausdruck)

Hit Location Search Only

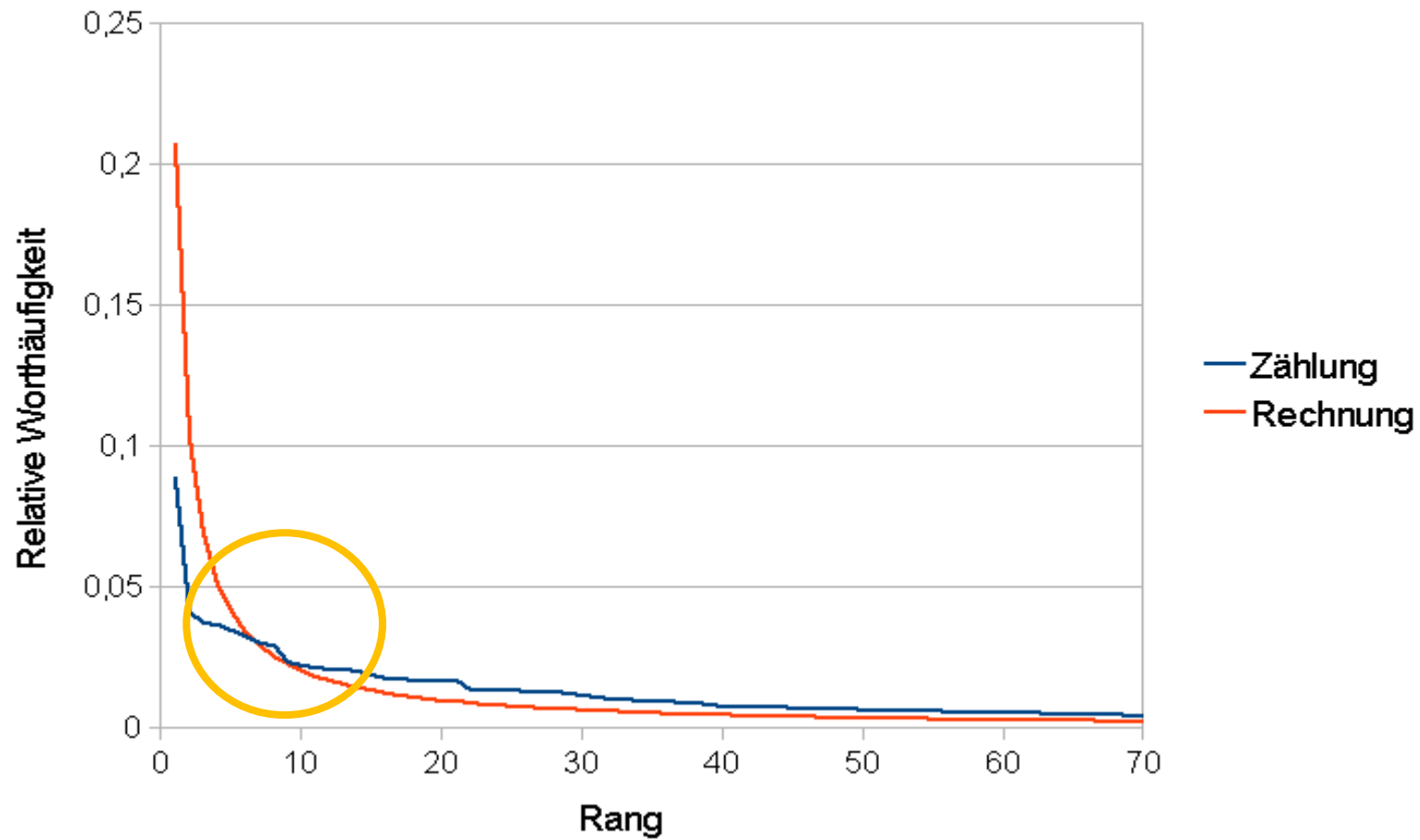
Sort by

Sortierung (hier: nach Frequenz)

<http://www1.ids-mannheim.de/fileadmin/lexik/lehre/engelberg/>

Webseite\_Korpusanalyse/Korpusanalyse\_4\_Methoden\_AntConc.pdf

# Zipf's Law



# AUTOMATIC INDEXING

- ◆ Absolute vs. relative frequency
  - Per document
  - Relative to document collection
  - Eliminate stop words (high occurrence frequency!)
- ◆ Assumption: frequency is positively correlated with relevance (denotation of main topics)
- ◆ Term frequency – inverse document frequency metric (TF-IDF)

$w_{ij}$  = weight of term  $t_j$  in document  $d_i$

$tf_{ij}$  = frequency of term  $t_j$  in document  $d_i$

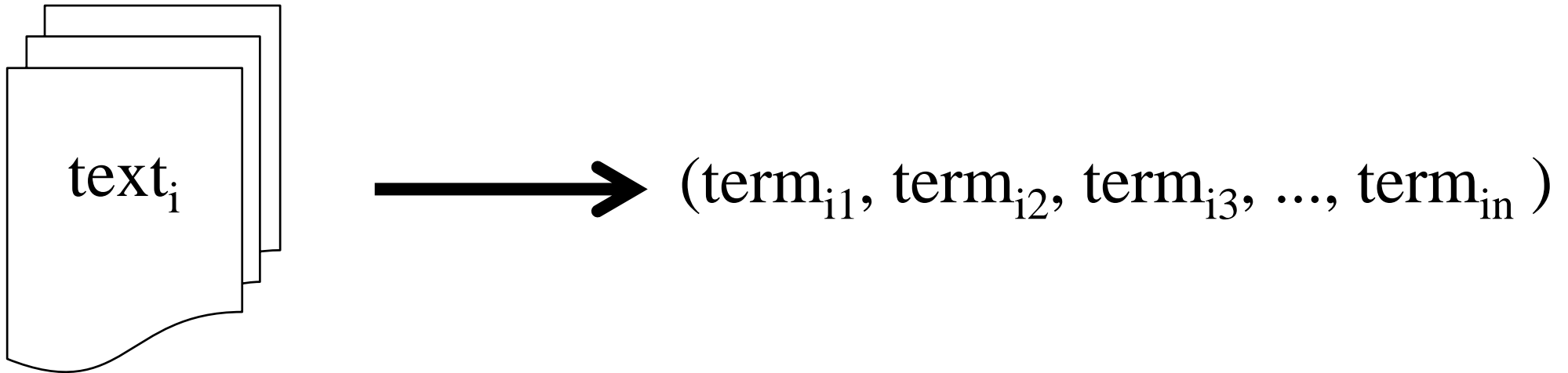
$N$  = number of documents in collection

$n$  = number of documents where term  $t_j$  occurs at least once

$$w_{ij} = tf_{ij} * \log_2 \frac{N}{n}$$

# VECTORIZATION OF TEXTS

- ◆ Transform text into n-dim vector (n=size of *collection* vocabulary)



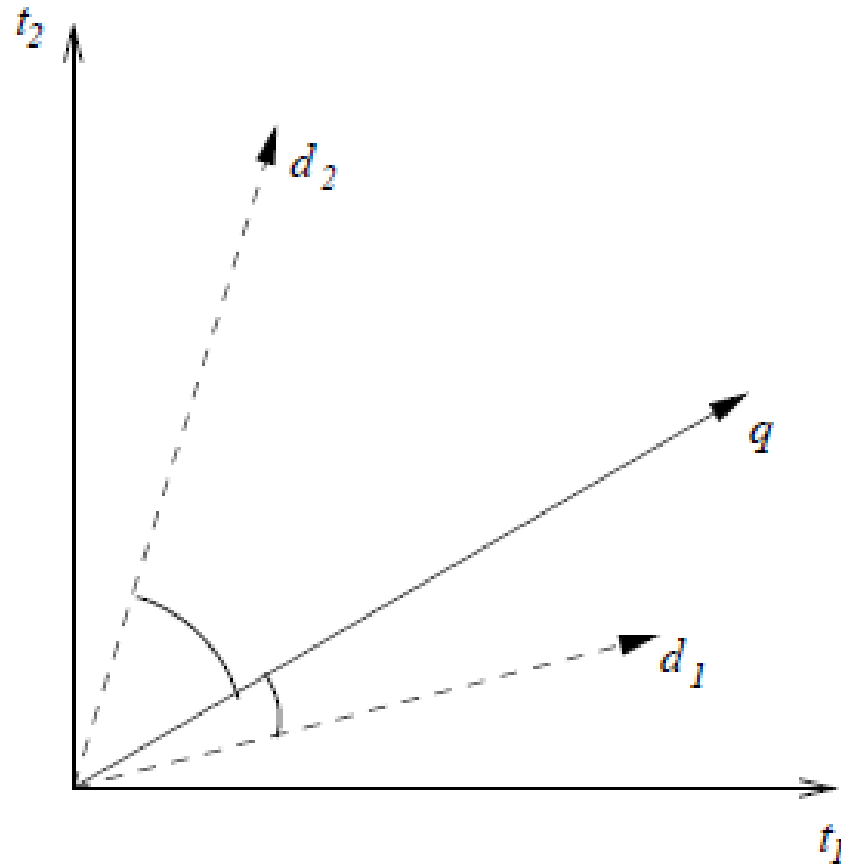


# AUTOMATIC INDEXING (Vector Space Model)

- ◆ Bag of words: remove all stop words from a doc and normalize all terms morphologically
- ◆ Create a document term matrix from the remaining terms for each document ( $n$  being the max number of terms in the document collection)
  - $\text{doc}_i = (\text{term}_{i1}, \text{term}_{i2}, \text{term}_{i3}, \dots, \text{term}_{in})$ 
    - Each component  $\text{term}_{ik}$  is either ,0‘ (absent) or ,1‘ (realized)
- ◆ Compute the association between a document term and a query term vector ( $\text{query} = (\text{query}_1, \text{query}_2, \text{query}_3, \dots, \text{query}_n)$ ,  $n$  as above), e.g., using the cosine measure

$$\text{SIM}(\text{doc}_i, \text{query}) = \frac{\sum_{k=1}^t (\text{term}_{ik} \bullet \text{query}_k)}{\sqrt{\sum_{k=1}^t (\text{term}_{ik})^2 \bullet \sum_{k=1}^t (\text{query}_k)^2}}$$

# GRAPHICAL INTERPRETATION



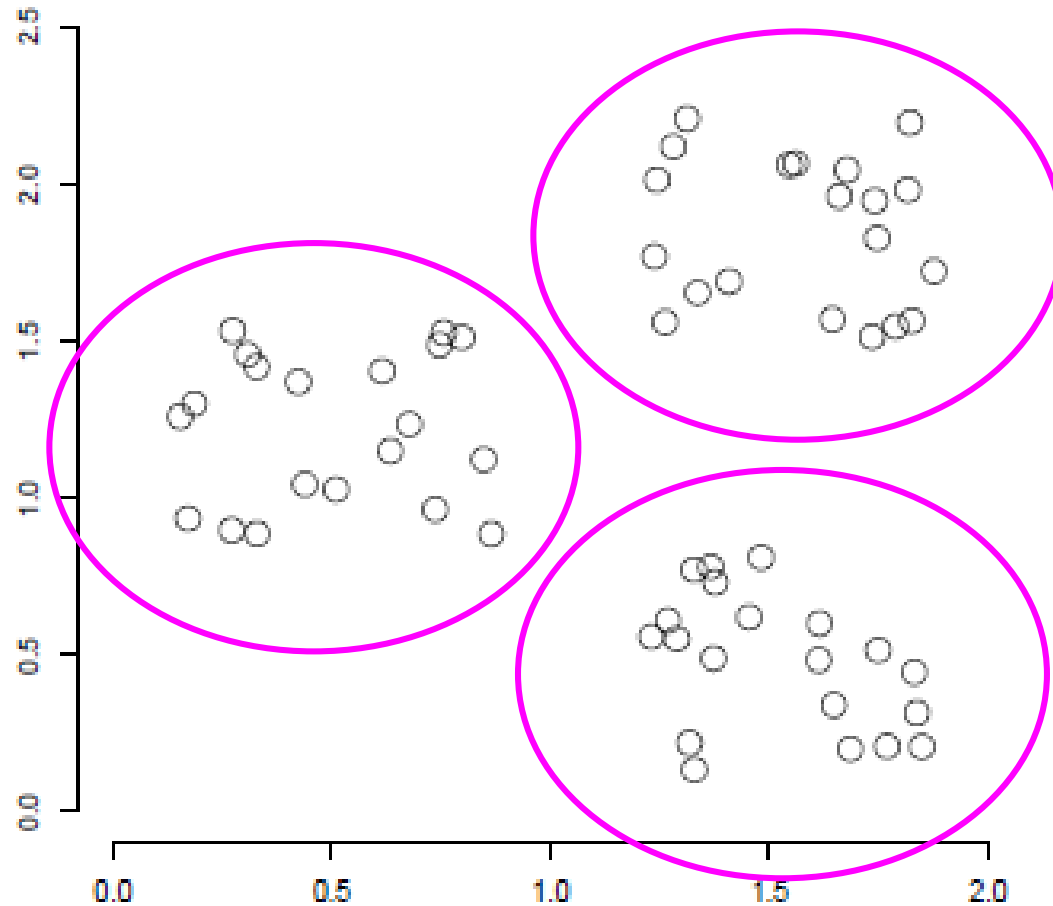
# CLASSIFICATION

- ◆ Manual classification
  - Manual assignment of docs to pre-defined categories (classes)
- ◆ Automatic classification
  - Automatic assignment of docs to pre-defined categories (classes)
  - Grouping of docs around automatically determined (unnamed) clusters

# Clustering

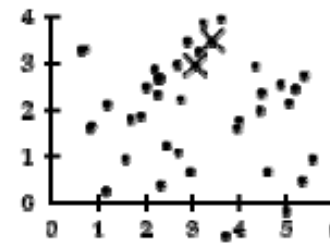
- (Document) clustering is the process of grouping a set of documents into clusters of similar documents.
- Documents within a cluster should be similar.
- Documents from different clusters should be dissimilar.
- Clustering is the most common form of **unsupervised** learning.
- Unsupervised = there are no labeled or annotated data.

# Data Set with Clear Clustering Structure

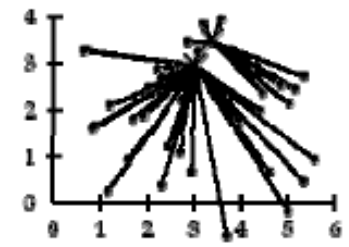


# Cluster-Modelle

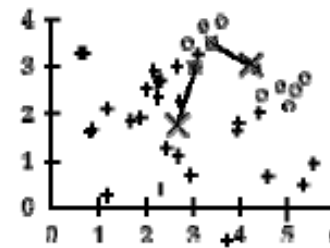
- k-Means Clustering
  - flaches Clustering
  - $k$  ist vorher bekannt
  - Dokumente werden als Vektoren repräsentiert
  - Ziel: Abstand zum Cluster-Zentrum minimieren
- Centroid
  - künstliches Zentrum eines Clusters – Mittelwert der Vektoren der Dokumente im Cluster
- Algorithmus
  - Initialisierung: wähle zufällig  $k$  Dokumente als Centroiden
  - Iteration: ordne Dokumente nächstem Centroid zu, Centroid im Cluster neu berechnen



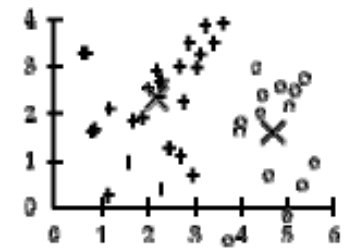
selection of seeds



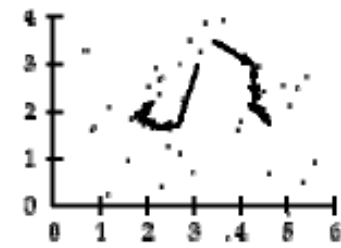
assignment of documents (iter. 1)



recomputation/movement of  $\mu$ 's (iter. 1)



$\mu$ 's after convergence (iter. 9)



movement of  $\mu$ 's in 9 iterations

Quelle: Manning, Raghavan, Schütze,  
Introduction to Information Retrieval, 2008.

# K-means Clustering

- Each cluster in  $K$ -means is defined by a centroid.
- Objective/partitioning criterion: minimize the average squared difference from the centroid
- Recall definition of centroid:

$$\vec{\mu}(\omega) = \frac{1}{|\omega|} \sum_{\vec{x} \in \omega} \vec{x}$$

where we use  $\omega$  to denote a cluster.

- We try to find the minimum average squared difference by iterating two steps:
  - reassignment: assign each vector to its closest centroid
  - recomputation: recompute each centroid as the average of the vectors that were assigned to it in reassignment

# K-means Clustering Algorithm

```
K-MEANS( $\{\vec{x}_1, \dots, \vec{x}_N\}, K$ )  
  1  ( $\vec{s}_1, \vec{s}_2, \dots, \vec{s}_K$ )  $\leftarrow$  SELECTRANDOMSEEDS( $\{\vec{x}_1, \dots, \vec{x}_N\}, K$ )  
  2  for  $k \leftarrow 1$  to  $K$   
  3  do  $\vec{\mu}_k \leftarrow \vec{s}_k$   
  4  while stopping criterion has not been met  
  5  do for  $k \leftarrow 1$  to  $K$   
  6  do  $\omega_k \leftarrow \{\}$   
  7  for  $n \leftarrow 1$  to  $N$   
  8  do  $j \leftarrow \arg \min_{j'} |\vec{\mu}_{j'} - \vec{x}_n|$   
  9  do  $\omega_j \leftarrow \omega_j \cup \{\vec{x}_n\}$  (reassignment of vectors)  
 10  for  $k \leftarrow 1$  to  $K$   
 11  do  $\vec{\mu}_k \leftarrow \frac{1}{|\omega_k|} \sum_{\vec{x} \in \omega_k} \vec{x}$  (recomputation of centroids)  
 12  return  $\{\vec{\mu}_1, \dots, \vec{\mu}_K\}$ 
```



# Idee des Relevance Feedback

- Relevance feedback: user feedback on relevance of docs in initial set of results
  - *User issues a (short, simple) query*
  - *The **user** marks some results as relevant or non-relevant.*
  - *The **system** computes a better representation of the information need based on feedback.*
  - *Relevance feedback can go through one or more **iterations**.*
- **Idea: it may be difficult to formulate a good query when you don't know the collection well, so iterate**

# Relevance Feedback (Rocchio-Algorithmus)

- ▶ unterschiedliche Gewichtung positiver und negativer Beispiele
- ▶ Berücksichtigung der ursprünglichen Anfrage

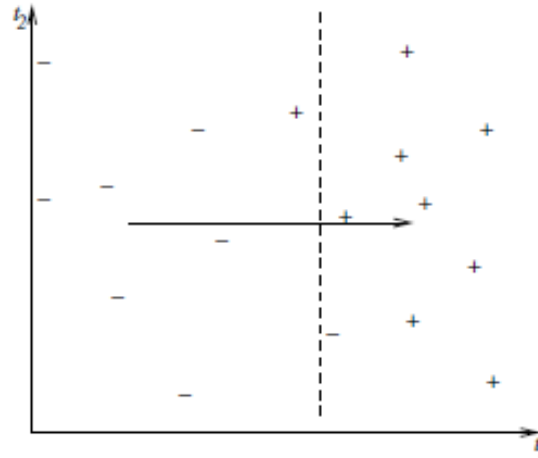
$$\vec{q}_k' = \vec{q}_k + \alpha \frac{1}{|D_k^R|} \sum_{d_j \in D_k^R} \vec{d}_j - \beta \frac{1}{|D_k^N|} \sum_{d_j \in D_k^N} \vec{d}_j$$

$\alpha, \beta$  — positive Konstanten, heuristisch festzulegen (z.B.  
 $\alpha = 0.75, \beta = 0.25$ )

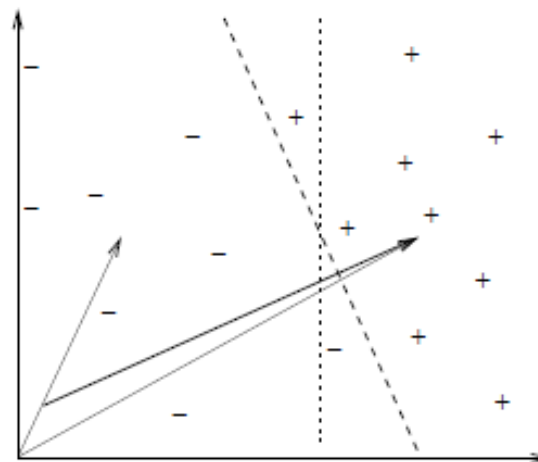
## Vorgehensweise:

1. Retrieval mit Fragevektor  $\vec{q}_k$  vom Benutzer
2. Relevanzbeurteilung der obersten Dokumente der Rangordnung
3. Berechnung eines verbesserten Fragevektors  $\vec{q}_k'$  aufgrund der Feedback-Daten
4. Retrieval mit dem verbesserten Vektor
5. Evtl. Wiederholung der Schritte 2-4

# Idee des Relevance Feedback (Rocchio-Algorithmus)



unterschiedliche Gewichtung positiver und negativer Beispiele:



# Rechenbeispiel zum Relevance Feedback

## ■ Beispiel:

Original query 

0	4	0	8	0	0
---	---	---	---	---	---

 $\alpha = 1.0$ 

0	4	0	8	0	0
---	---	---	---	---	---

2	4	8	0	0	2
---	---	---	---	---	---

 $\beta = 0.5$ 

1	2	4	0	0	1
---	---	---	---	---	---

 (+)

8	0	4	4	0	16
---	---	---	---	---	----

 $\gamma = 0.25$ 

2	0	1	1	0	4
---	---	---	---	---	---

 (-)

---

New query 

-1	6	3	7	0	-3
----	---	---	---	---	----

# ANTWORTEN VON INFORMATIONSSYSTEMEN

**Datenbanksysteme** liefern stets korrekte und vollständige Antwort auf Anfragen

- ▶ im Sinne eines Beweisverfahrens
- ▶ i.a. *nicht* bezüglich der realen Welt

→ Betrachtung von Effektivität hier nicht sinnvoll

**IR-Systeme** können wegen Vagheit und Unsicherheit i.a.

- ▶ weder korrekte (alle gefundenen Dokumente relevant)
- ▶ noch vollständige (alle relevanten Dokumente)

Antworten liefern.

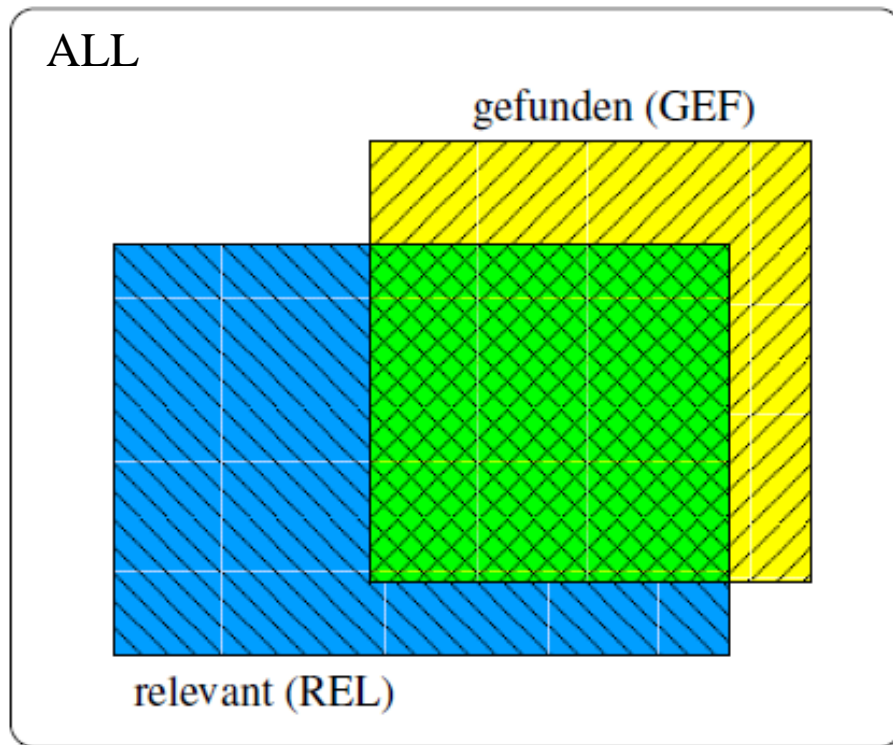
→ Effektivität als wichtiges Qualitätskriterium

# EVALUATIONSMETRIKEN

GEF: Menge der gefundenen Antwortdokumente

REL: Menge der relevanten Dokumente in der Datenbank

ALL: Menge aller Dokumente in der Datenbank



$$\text{Precision } p = \frac{|REL \cap GEF|}{|GEF|}$$

$$\text{Recall } r = \frac{|REL \cap GEF|}{|REL|}$$

$$\text{Fallout } f = \frac{|GEF - REL|}{|ALL - REL|}$$

# INTEGRATION IM F-MASS

Abbildung von  $(r, p)$ -Paar auf einzelnes Maß  
(definiert Kurve zur Aufteilung des 'Unentschieden-Bereichs')

Grundidee:

harmonisches Mittel aus Recall und Precision

$$F = \frac{2}{\frac{1}{p} + \frac{1}{r}}$$

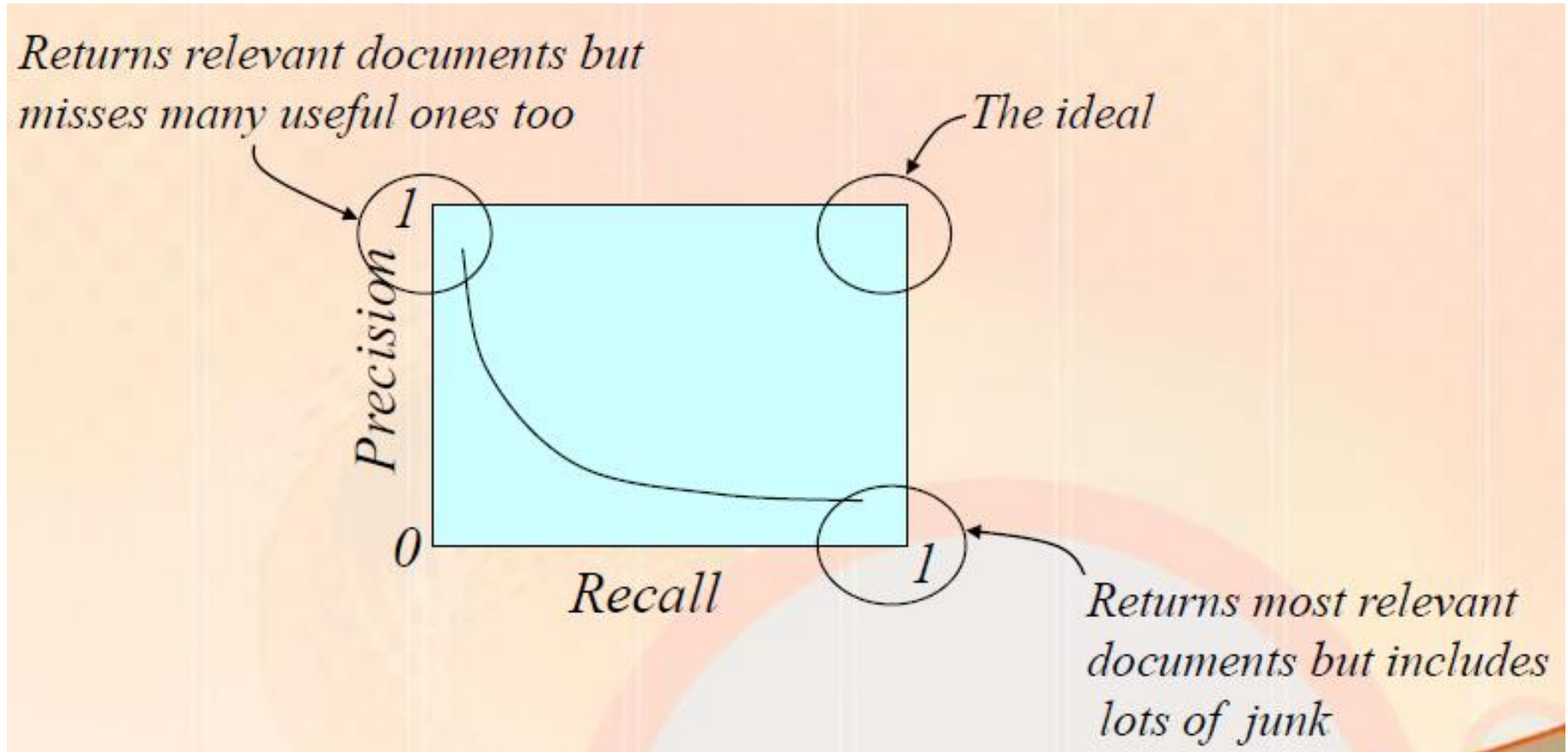
Unterschiedliche Gewichtung von Recall und Precision:

Gewichtungsfaktor  $\beta$  für Recall

$$F_{\beta} = \frac{1 + \beta^2}{\frac{1}{p} + \beta^2 \frac{1}{r}}$$

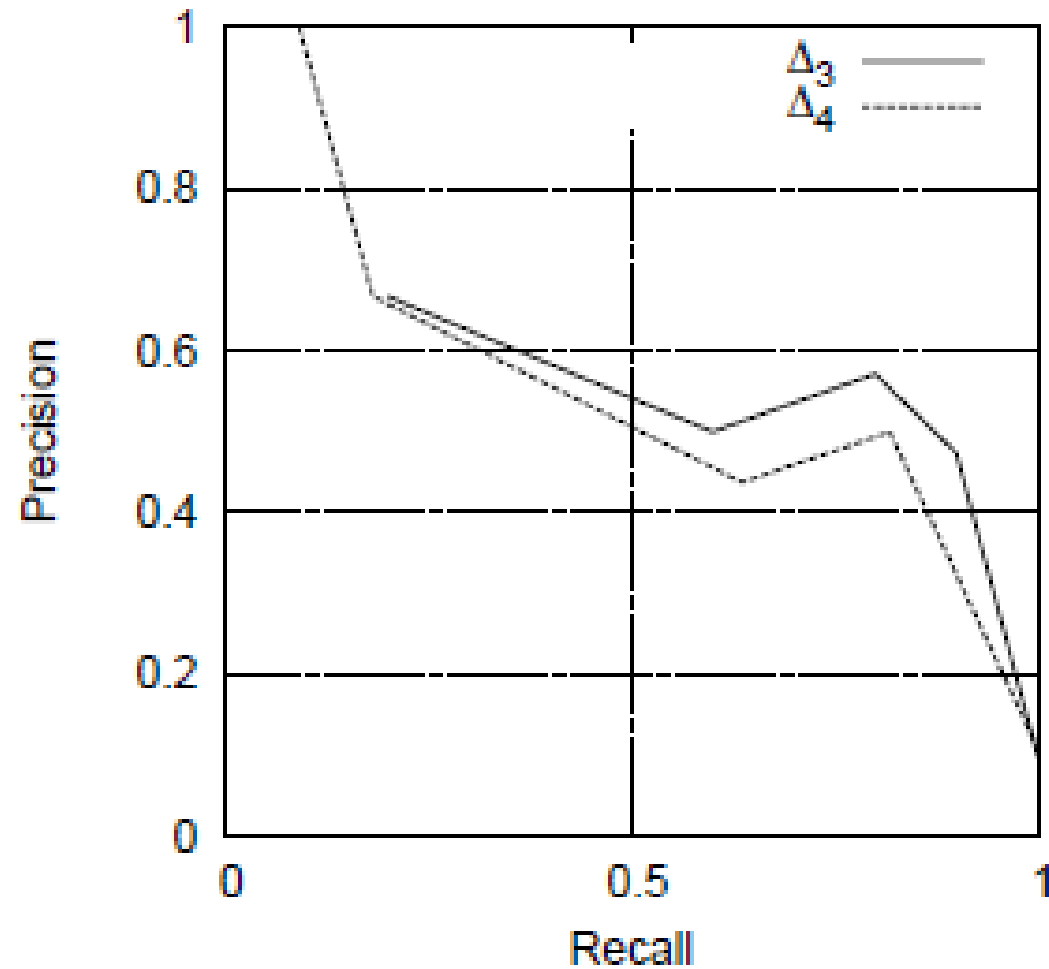


# Trade-off between Precision and Recall





# „NATURGESETZ“ DER INVERSEN P-R-BEZIEHUNG



# EVALUIERUNGSINITIATIVEN: TREC, CLEF, NTCIR, INEX, ...

Standardumgebung für die Evaluierung von IR-Methoden:

- ▶ Dokumentkollektionen im Umfang praktischer Anwendungen (z.B. Newsticker-, Zeitungs-, Magazinartikel, Web-Kollektionen)
- ▶ vordefinierte Anfragen (*Topics*)
- ▶ verschiedene Aufgaben (*Tracks*)

# TREC EVALUATIONSMETRIKEN

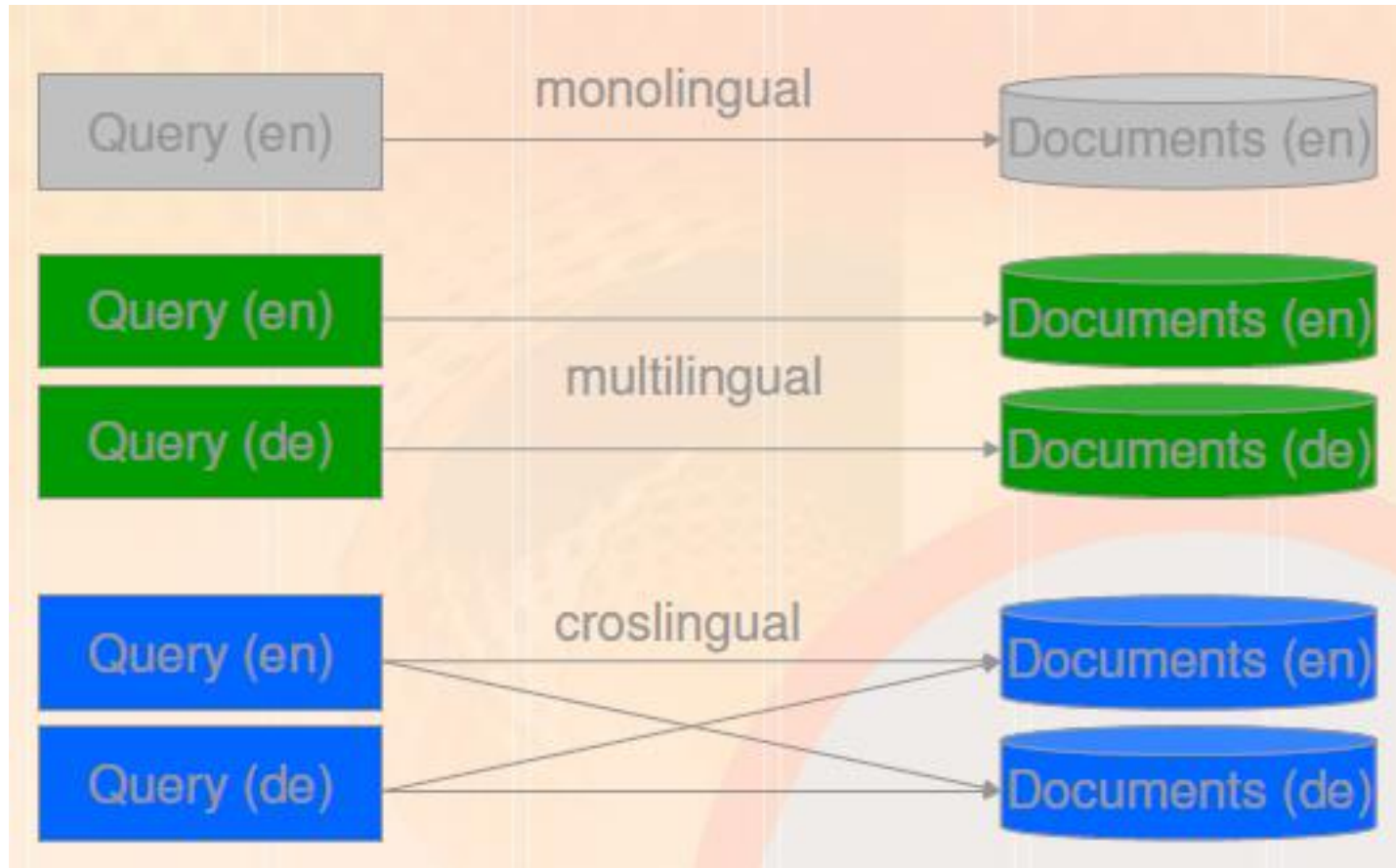
## Benutzerorientierte Maße:

- ▶ **Prec@5, Prec@10, Prec@30, Prec@100**  
(jeweils als Makro-Mittelwert über alle Fragen)
- ▶ **Mean reciprocal rank:**  
Annahme: Benutzer ist nur an einem relevanten Dokument interessiert
  1. Bestimme Rang  $k$  des ersten relevanten Dokumentes
  2. Bilde Kehrwert  $1/k$
  3. Mittele über alle Fragen

## Systemorientiertes Maß: **Mean average precision (MAP)**

Mittelwert der Precision nach jedem relevanten Dokument einer Rangliste  
(anschließend arithmetisches Mittel über alle Fragen)

# Wichtige Forschungsfragen (1/2)



# Wichtige Forschungsfragen (2/2)

## Multi-media Retrieval

- Text
- Grafiken
- Tabellen
- Fotos
- Filme
- Musik

# TREC Medicine

- **Genomics Track (2004-08)**
  - Retrieving information about genes
- **Clinical Decision Support Track (2014-16)**
  - Retrieving information from the Electronic Health Record
    - Evidence- based information (in the form of full-text literature articles) to clinicians for a specific patient (represented as a case description or admission note)

# TREC Precision Medicine

- Precision Medicine Track (2017-2018)
  - Precision medicine paradigm
    - Personalized treatment for patients based on their genetic, environmental and life style characteristics
  - Focus on genetic mutations of cancer
  - Retrieving scientific abstracts (Medline) relevant for patient's case
  - Retrieving clinical trials documents (ClinicalTrials.gov) most similar to patient's case

# TREC PM 2017/2018

- TREC-PM 2017/2018
  - Initialized 2017, largely repeated in 2018
  - 30 synthetically created topics
  - each topic is described by 4 items
    - disease (e.g., type of cancer)
    - genetic variants (primarily the genetic variants in the tumors themselves as opposed to the patient's DNA)
    - demographic information (e.g., age, sex), and
    - other factors (which could impact certain treatment options)

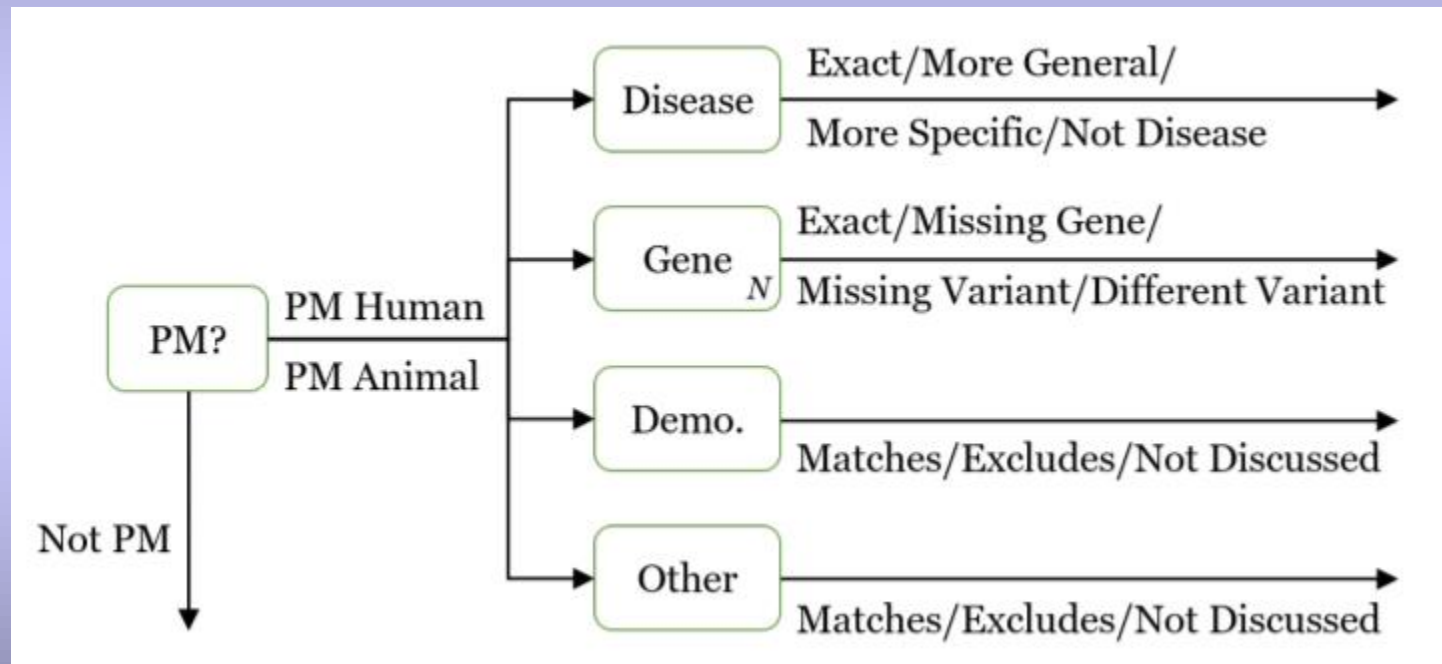


# TREC-PM Topics

<b>Disease:</b> Liposarcoma <b>Variant:</b> CDK4 Amplification <b>Demographic:</b> 38-year-old male <b>Other:</b> GERD
<b>Disease:</b> Colon Cancer <b>Variant:</b> KRAS (G13D), BRAF (V600E) <b>Demographic:</b> 52-year-old male <b>Other:</b> Type II Diabetes, Hypertension
<b>Disease:</b> Cervical Cancer <b>Variant:</b> STK11 <b>Demographic:</b> 26-year-old female <b>Other:</b> None
<b>Disease:</b> Cholangiocarcinoma <b>Variant:</b> IDH1 (R132H) <b>Demographic:</b> 64-year-old male <b>Other:</b> Neuropathy

# TREC PM 2017

## Result Assessment



Roberts, Kirk, & Demner-Fushman, Dina, & Voorhees, Ellen M., & Hersh, William R., & Bredrik, Steven, & Lazar, Alexander J., & Pant, Shubham (2017). Overview of the TREC 2017 Precision Medicine Track. in: TREC 2017 – Proceedings of the 26th Text REtrieval Conference. Gaithersburg, Maryland, USA, November 15-17, 2017, 1-13.

# TREC PM 2018

## Evaluation Criteria

### Evaluation

The evaluation will follow standard TREC evaluation procedures for ad hoc retrieval tasks. Participants may submit a maximum of **five automatic or manual runs for each corpus (scientific abstracts and clinical trials)**, each consisting of a ranked list of up to one thousand IDs (**PMIDs for MEDLINE abstracts, provided IDs for extra abstracts (part of file name), and NCT IDs for trials**). The highest ranked results for each topic will be pooled and judged by physicians trained in medical informatics.

Assessors will be instructed to judge abstracts and clinical trials according to each of the four topic dimensions (disease, gene, demographic). Each of these corresponds to 3-4 categories (e.g., a disease can be an "exact", "more general", "more specific", or "not disease" match). Please read the [Relevance Guidelines](#) for more details.

**Scientific Abstracts:** The goal of retrieving scientific abstracts is to identify relevant articles for the *treatment, prevention, and prognosis* of the disease under the specific conditions for the given patient. Abstracts discussing information not useful for these goals will not be considered relevant.

**Clinical Trials:** The goal of retrieving clinical trials is to identify trials for which the given patient is eligible to enroll, or *would have been* eligible to enroll had the trial been open. *The timing and location of the trial are not factors in determining relevance, only the eligibility criteria.*

As in past evaluations of medically-oriented TREC tracks, we are fortunate to have the assessment conducted by the Department of Medical Informatics of the Oregon Health and Science University (OHSU). We are extremely grateful for their participation.

## Literature Articles

2

infNDCG		
Team	Run	Score
Cat_Garfield	MSIIP_BASE	0.5621
hpi-dhc	hpiplibnone	0.5605
UCAS	UCASSA5	0.5580
MedIER	MedIER_sa13	0.5515
SIBTextMining	SIBTMLit4	0.5410
imi_mug	imi_mug_abs2	0.5391
udel_fang	UDInfoPMSA2	0.5081
RSA_DSC	RSA_DSC_LA_5	0.4855
UTDHLTRI	UTDHLTRLNL	0.4797
IKMLAB	IKMLAB_3	0.4710

2

R-prec		
Team	Run	Score
MedIER	MedIER_sa13	0.3684
hpi-dhc	hpiplibcommon	0.3658
UCAS	UCASSA2	0.3654
imi_mug	imi_mug_abs1	0.3630
SIBTextMining	SIBTMLit3	0.3574
udel_fang	UDInfoPMSA1	0.3289
Cat_Garfield	MSIIP_PBPk	0.3257
SINAI	SINAI_Base	0.3082
FDUDMIIP	raw_medline	0.3072
cbnu	cbnuSA1	0.2992

1

P @ 10		
Team	Run	Score
hpi-dhc	hpiplibnone	0.7060
Cat_Garfield	MSIIP_BASE	0.6680
SIBTextMining	SIBTMLit5	0.6320
UVA_ART	UVAEXPBTEXT	0.6260
MedIER	MedIER_sa11	0.6220
UTDHLTRI	UTDHLTRLNL	0.6160
imi_mug	imi_mug_abs2	0.6000
UCAS	UCASSA5	0.5980
IKMLAB	IKMLAB_3	0.5960
udel_fang	UDInfoPMSA2	0.5800

## Clinical Trials

1

infNDCG		
Team	Run	Score
hpi-dhc	hpipliball	0.5545
Cat_Garfield	MSIIP_TRIAL1	0.5503
ims_unipd	IMS_TERM	0.5395
UCAS	UCASCT4	0.5347
udel_fang	UDInfoPMCT1	0.5057
NOVASearch	NS_PM_5	0.4992
Poznan	BB2_vq_nopr	0.4894
UTDHLTRI	UTDHLTRLNL	0.4794
RSA_DSC	RSA_DSC_CT_5	0.4743
IRIT	irit_prf_cli	0.4736

4

R-prec		
Team	Run	Score
Cat_Garfield	MSIIP_TRIAL1	0.4294
ims_unipd	IMS_TERM	0.4128
Poznan	BB2_vq_nopr	0.4101
hpi-dhc	hpiplibphrase	0.4081
UCAS	UCASCT4	0.4005
udel_fang	UDInfoPMCT3	0.3967
NOVASearch	NS_PM_5	0.3931
UTDHLTRI	UTDHLTRLSSST	0.3920
RSA_DSC	RSA_DSC_CT_5	0.3721
IRIT	irit_prf_cli	0.3658

7

P @ 10		
Team	Run	Score
Cat_Garfield	MSIIP_TRIAL1	0.6260
ims_unipd	IMS_TERM	0.5660
Poznan	BB2_vq_nopr	0.5580
NOVASearch	NS_PM_5	0.5520
RSA_DSC	RSA_DSC_CT_3	0.5480
UCAS	UCASCT1	0.5460
hpi-dhc	hpiplibphrase	0.5400
UTDHLTRI	UTDHLTRLNL	0.5380
udel_fang	UDInfoPMCT5	0.5240
InfoLabPM	tinfolabBF	0.5240

Team ID	Affiliation	# Runs	
		Articles	Trials
ASU_Biomedical	Arizona State University	3	0
Brown	Brown University	5	5
Cat_Garfield	Tsinghua-iFlytek Joint Laboratory	5	5
cbnu	Chonbuk National University	3	3
CSIROmed	Commonwealth Scientific and Industrial Research Organisation	3	3
ECNUica	East China Normal University	5	5
FDUDMIIP	School of Computer Science, Fudan University	5	5
hpi-dhc	Hasso Plattner Institute Med. Universität Graz, JULIE Lab	5	5
IKMLAB	Institute of Medical Informatics of National Cheng Kung Univ.	5	5
imi_mug	Medical University of Graz	5	5
ims_unipd	Information Management Systems (IMS) Group	0	3
InfoLabPM	InfoLab, Faculty of Engineering, University of Porto	4	3
IRIT	Institut de Recherche en Informatique de Toulouse	0	1
KlickLabs	Klick Inc.	4	5
MayoNLPTeam	Mayo Clinic	4	3
MedIER	University of Michigan	5	0
NOVASearch	Universidade NOVA Lisboa	0	5
PM_JBI	Integrative Biomedical Informatics Group, Barcelona	3	0
Poznan	Poznan University of Technology	1	5
RSA_DSC	Research Studios Austria / Studio Data Science	5	5
SIBTextMining	SIB Text Mining Group (HES-SO)	5	4
SINAI	Universidad de Jaen	3	0
UCAS	University of Chinese Academy of Sciences	5	5
udel_fang	InfoLab at University of Delaware	5	5
UNTIIA	University of North Texas	5	0
UTDHLTRI	The University of Texas at Dallas	5	5
UVA_ART	University of Virginia Medical Center	5	0
Total	<b>27 Teams</b>	103	90

Table 5: Participating teams and submitted runs.

# Literatur

- Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Modern Information Retrieval, 1999.
  - der Klassiker
- Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, Introduction to Information Retrieval, 2008.
  - Online verfügbar unter: <http://nlp.stanford.edu/IR-book/information-retrieval-book.html>



# Document Content Analysis Techniques I

## Indexing and Classification

Udo Hahn



FRIEDRICH-SCHILLER-UNIVERSITÄT  
JENA



Jena University Language and Information Engineering (JULIE) Lab, Germany

[www.julielab.de](http://www.julielab.de)