

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

stern

Disease Management

Volks

SRU Sachverständigenrat für Umweltfragen

13. Februar 2009

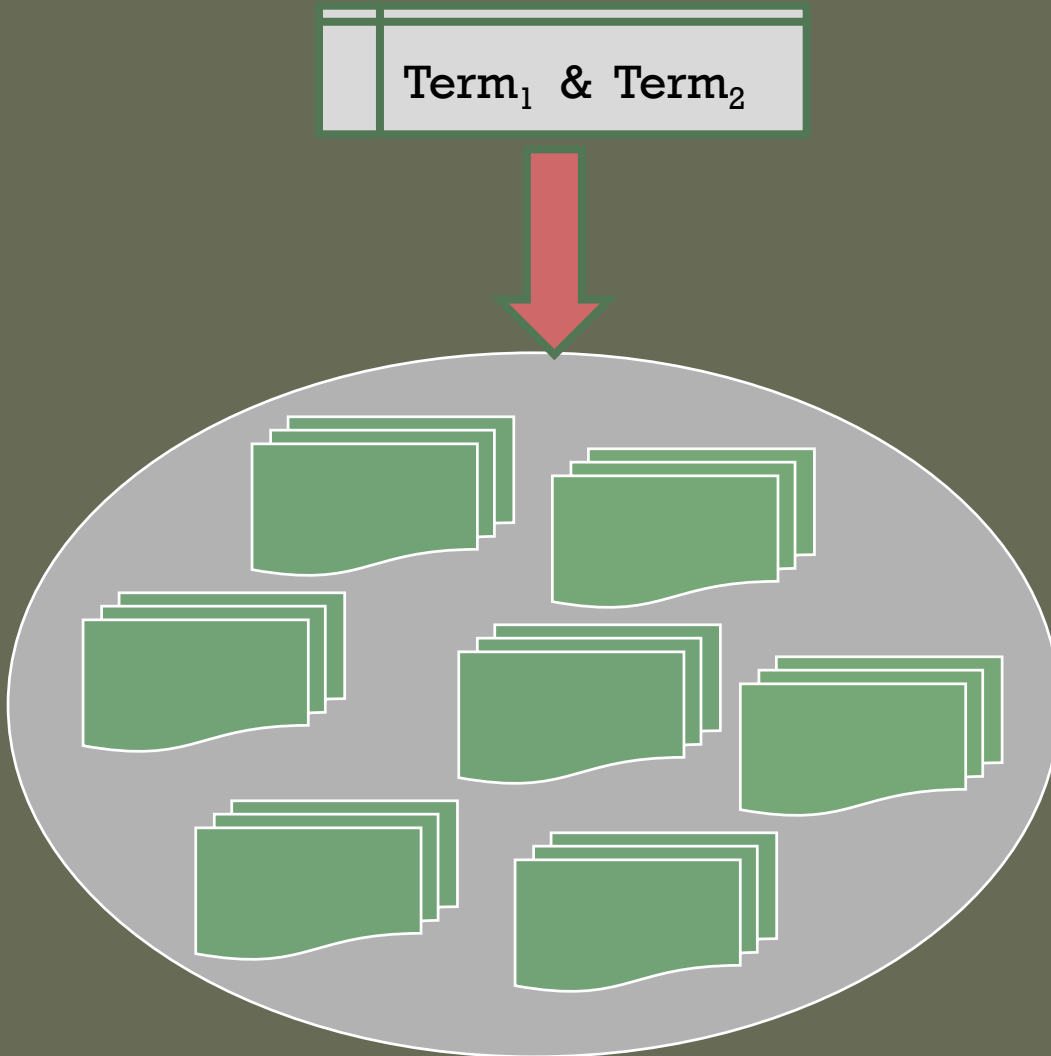
Bluefish 0.9.5 HTML editor

Amaya

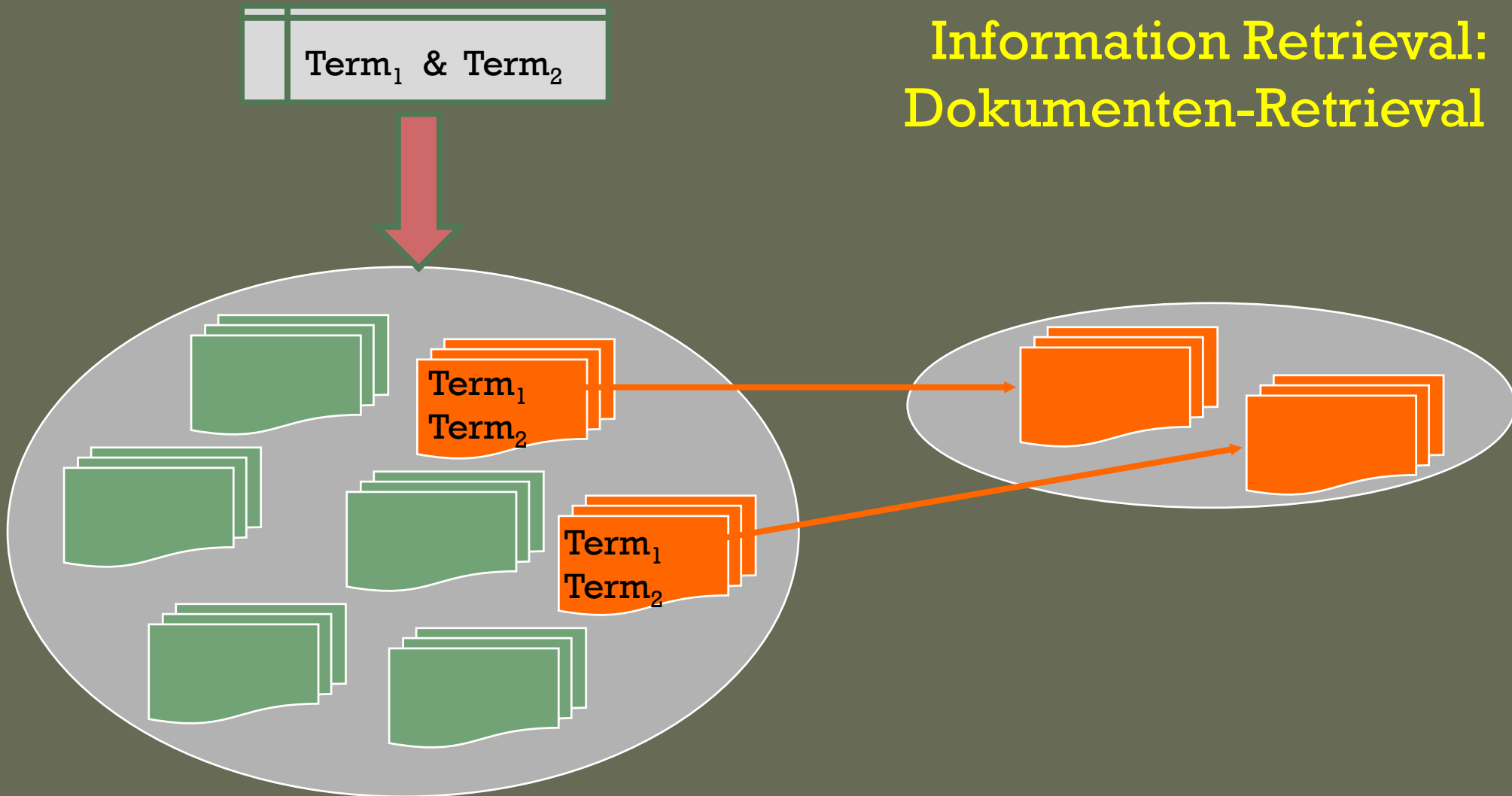
Süddeutsche Zeitung

SPD stärkste Partei, AfD liegt vor CDU

Information Retrieval:
Dokumenten-Retrieval



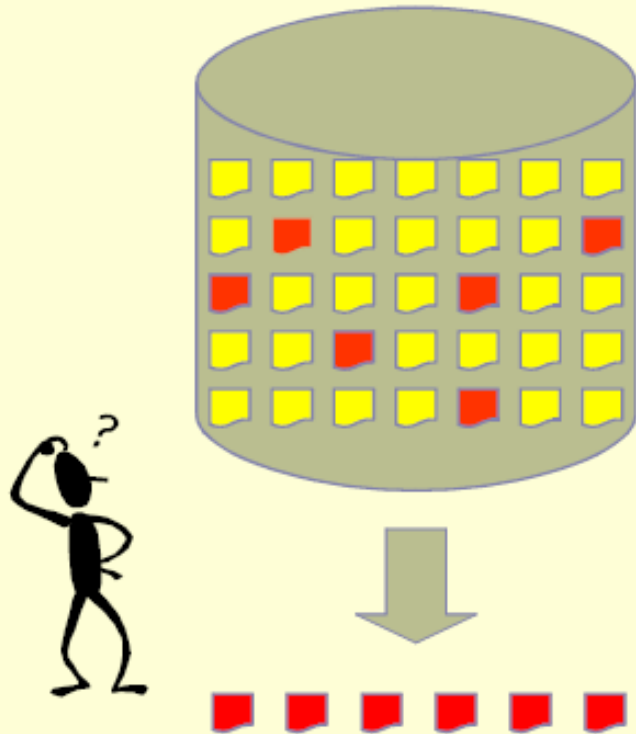
Information Retrieval: Dokumenten-Retrieval



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)

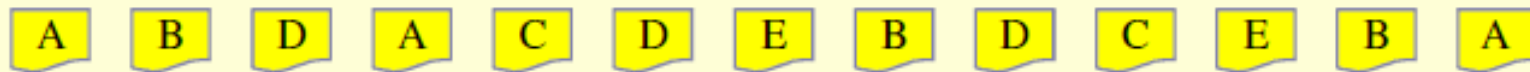
- **Categorization/Clustering:**

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

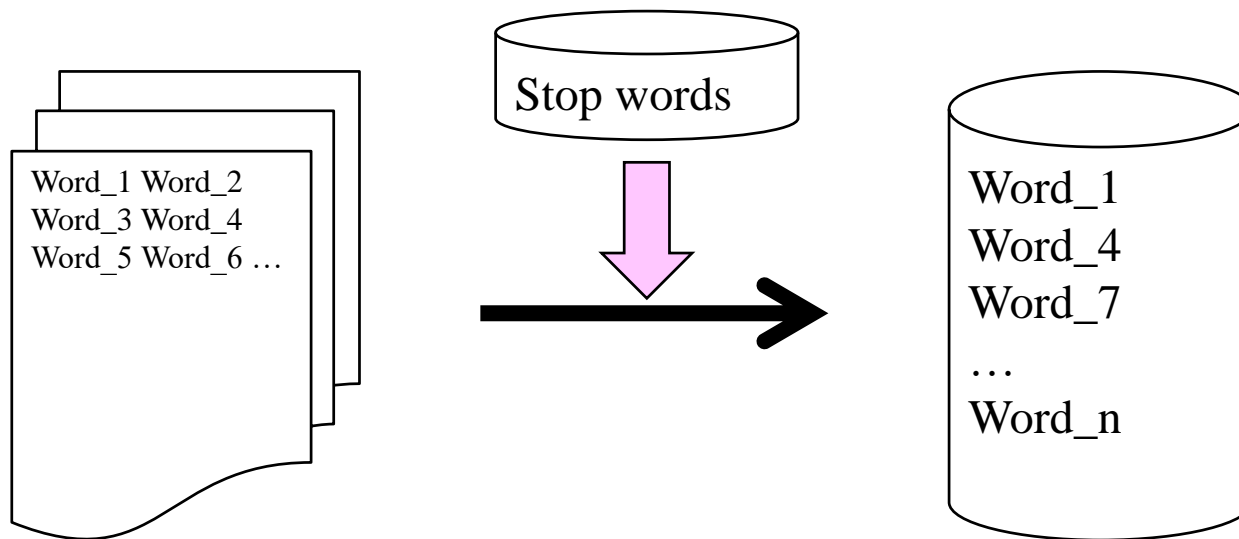
- ◆ Determine main topic(s)
- ◆ What's a relevant issue?
- ◆ Based on human (speed) reading and understanding of the document

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

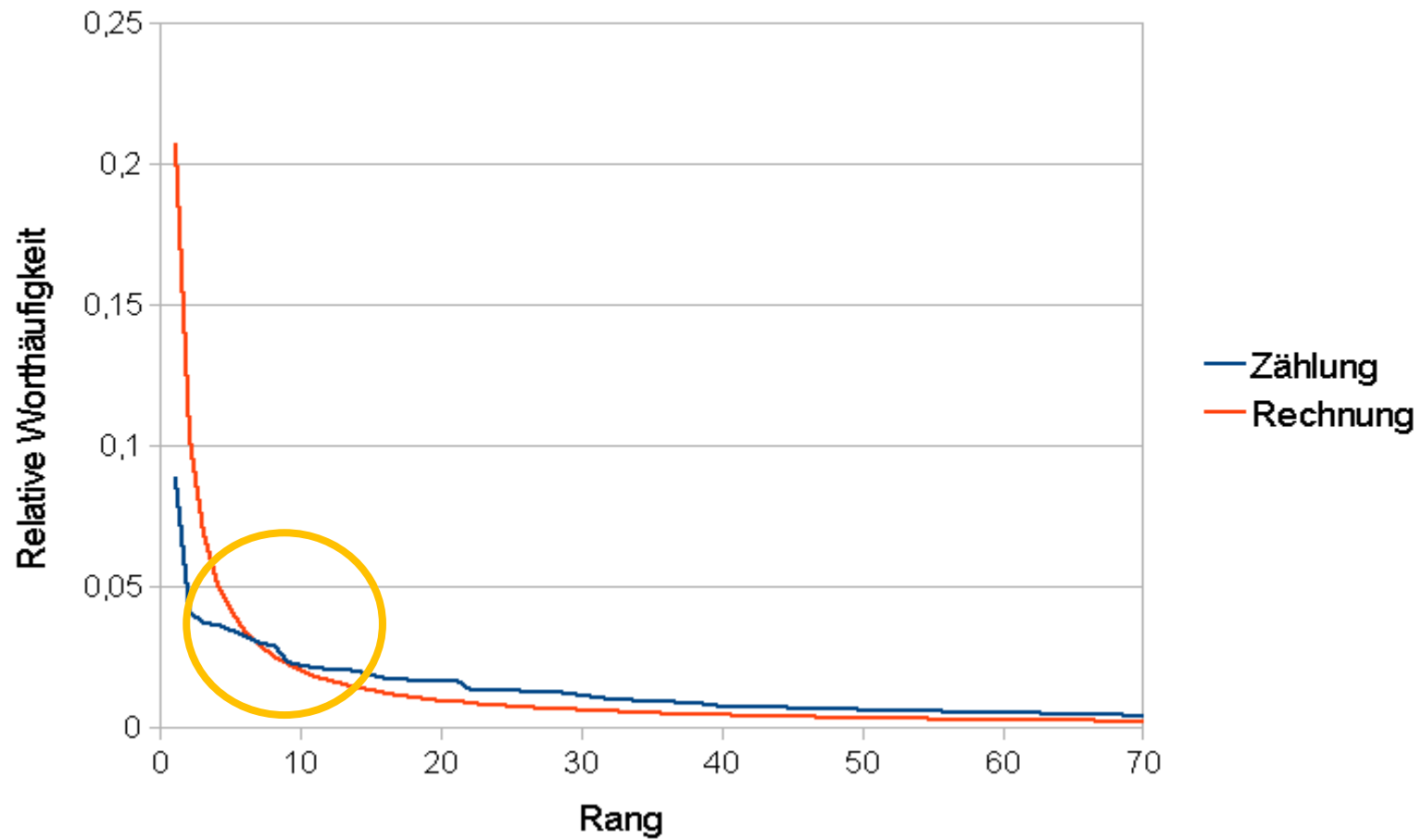
Sort by

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

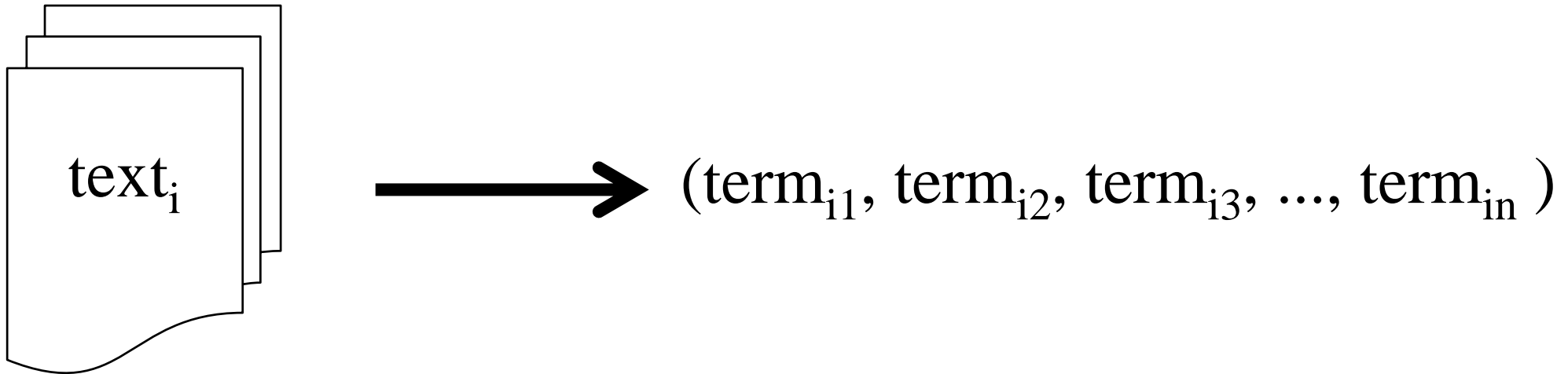
N = number of documents in collection

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)

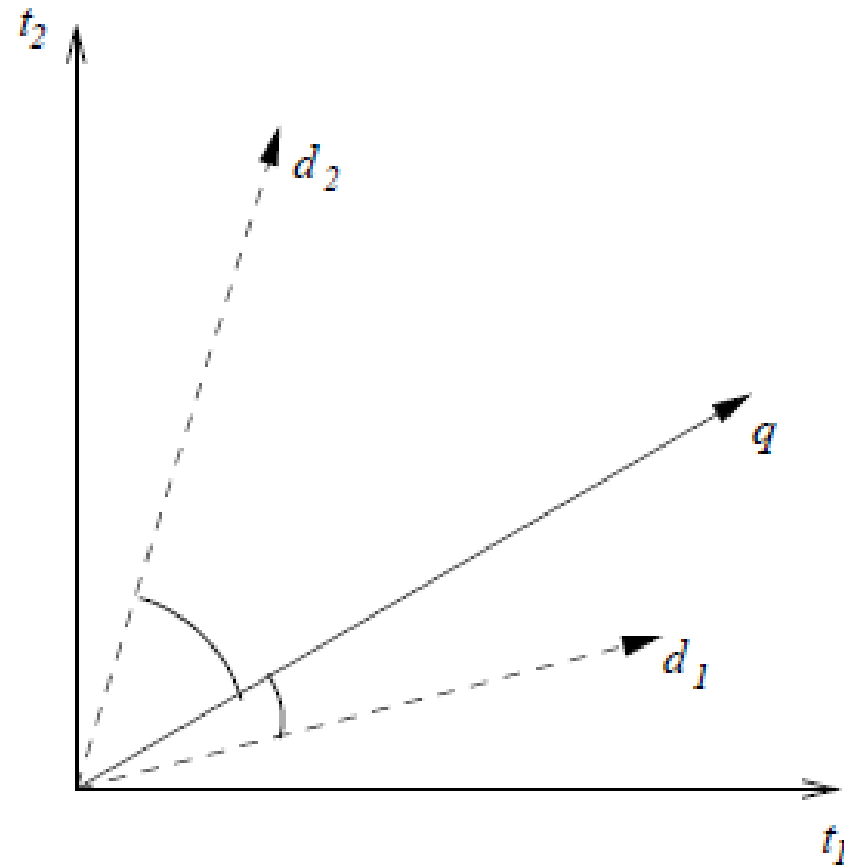


AUTOMATIC INDEXING (Vector Space Model)

- ◆ 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 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



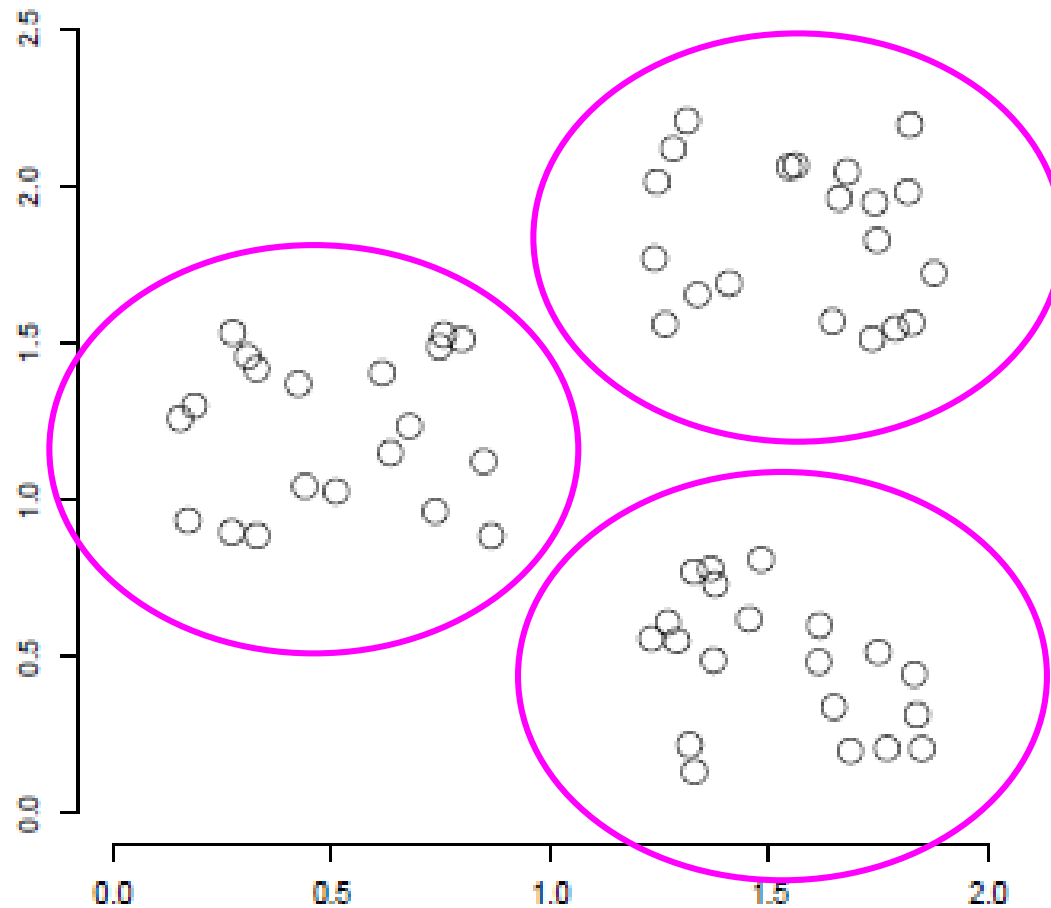
CLASSIFICATION

- ◆ 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

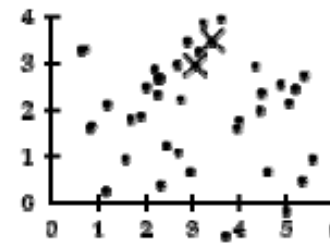
- (Document) clustering is the process of grouping a set of documents into clusters of similar documents.
- Documents within a cluster should be similar.
- Documents from different clusters should be dissimilar.
- Clustering is the most common form of **unsupervised** learning.
- Unsupervised = there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

Data Set with Clear Clustering Structure

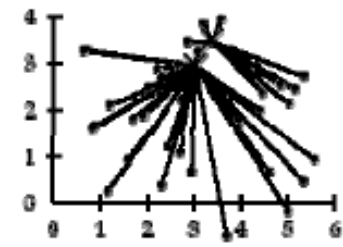


Cluster-Modelle

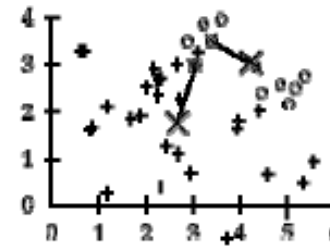
- k-Means Clustering
 - flaches Clustering
 - k ist vorher bekannt
 - Dokumente werden als Vektoren repräsentiert
 - Ziel: Abstand zum Cluster-Zentrum minimieren
- Centroid
 - künstliches Zentrum eines Clusters – Mittelwert der Vektoren der Dokumente im Cluster
- Algorithmus
 - Initialisierung: wähle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



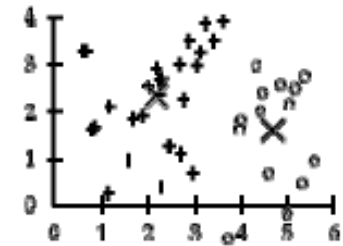
selection of seeds



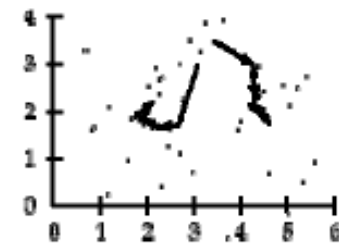
assignment of documents (iter. 1)



recomputation/movement of μ 's (iter. 1)



μ 's after convergence (iter. 9)



movement of μ 's in 9 iterations

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