Einführung in die Computer-linguistik und Sprachtechnologie

WiSe 2018/2019

(B-GSW-12)

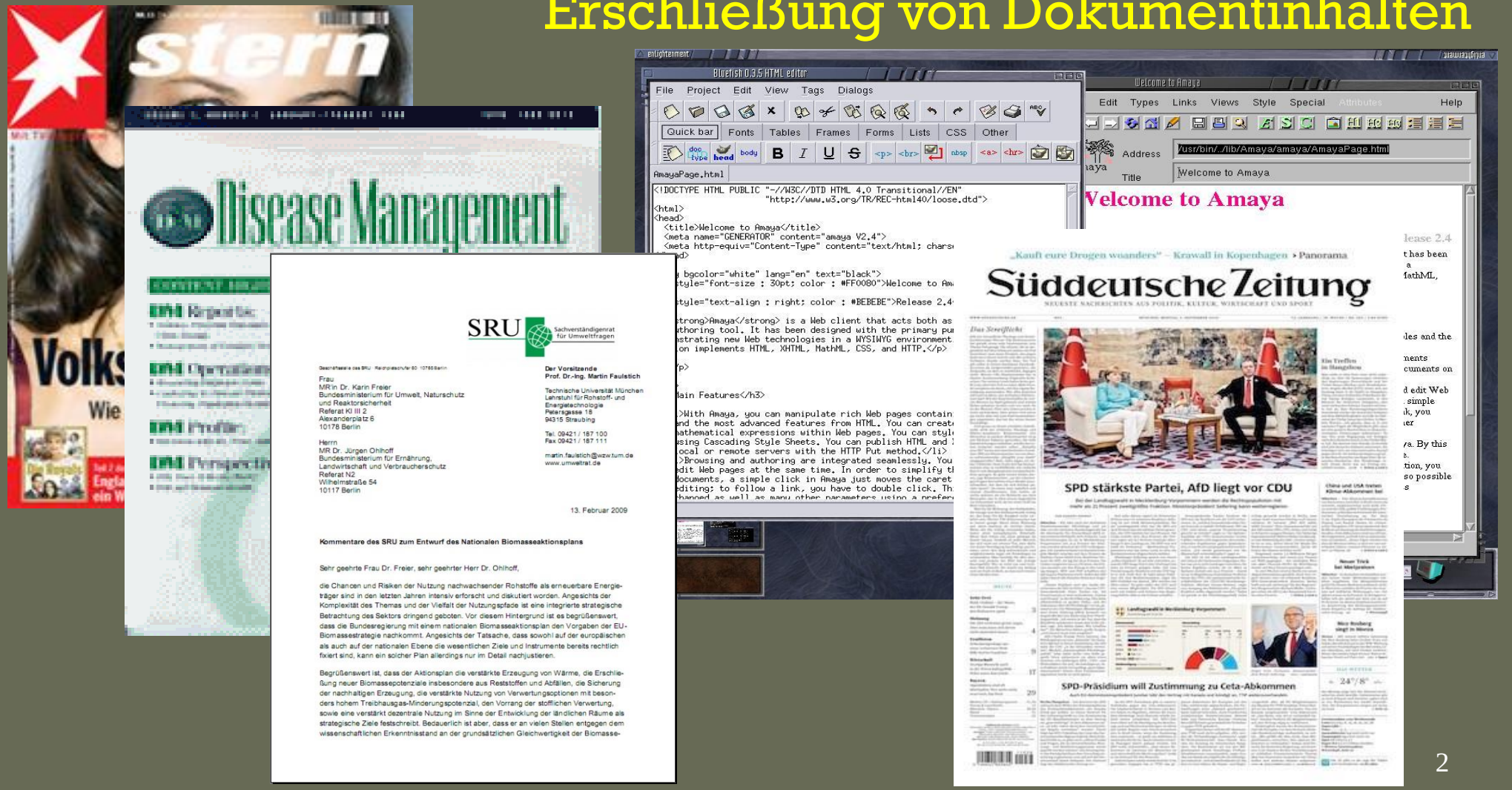
Udo Hahn

FRIEDRICH-SCHILLER-UNIVERSITÄT
JENA

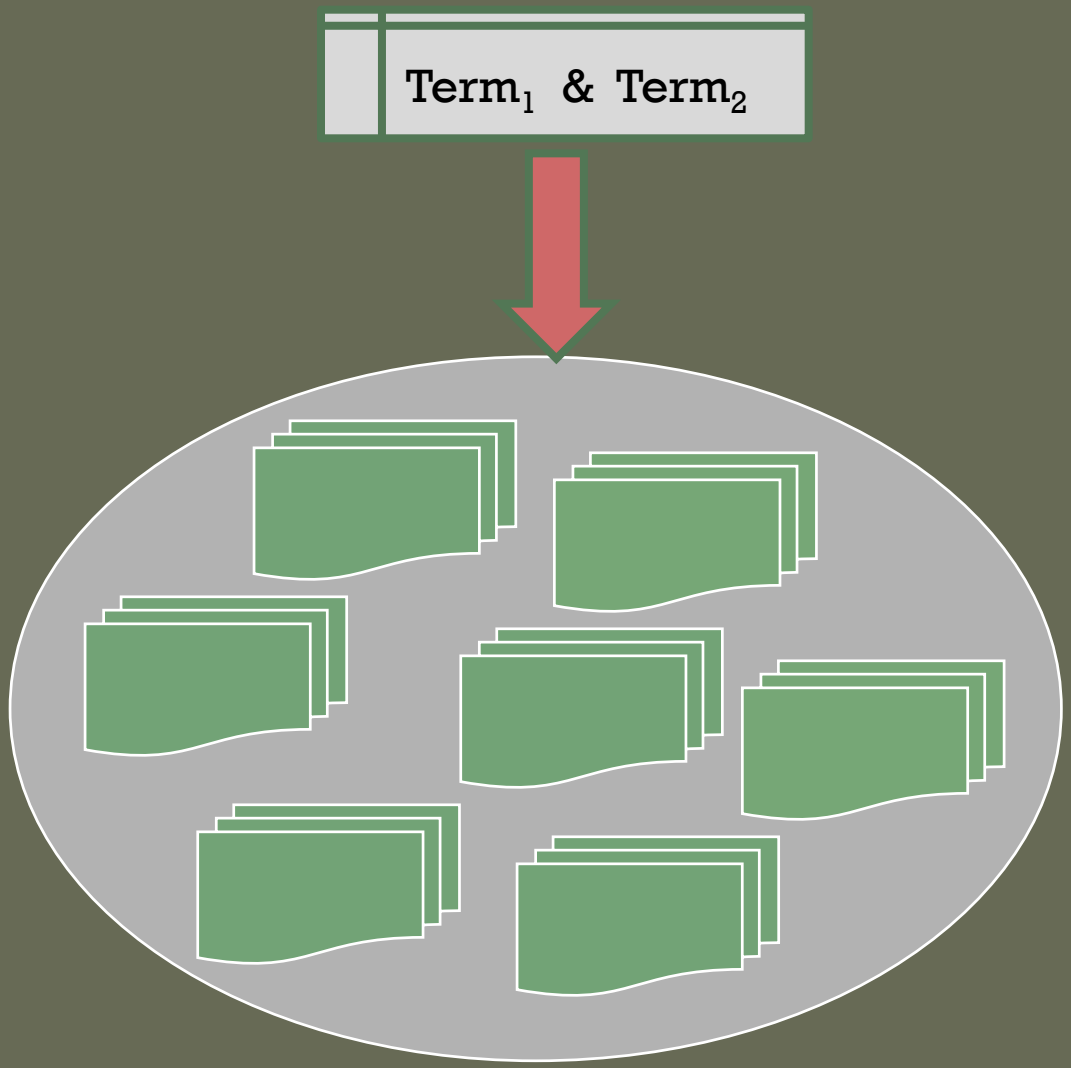
Jena University Language and Information Engineering (JULIE) Lab, Germany

http://www.julielab.de

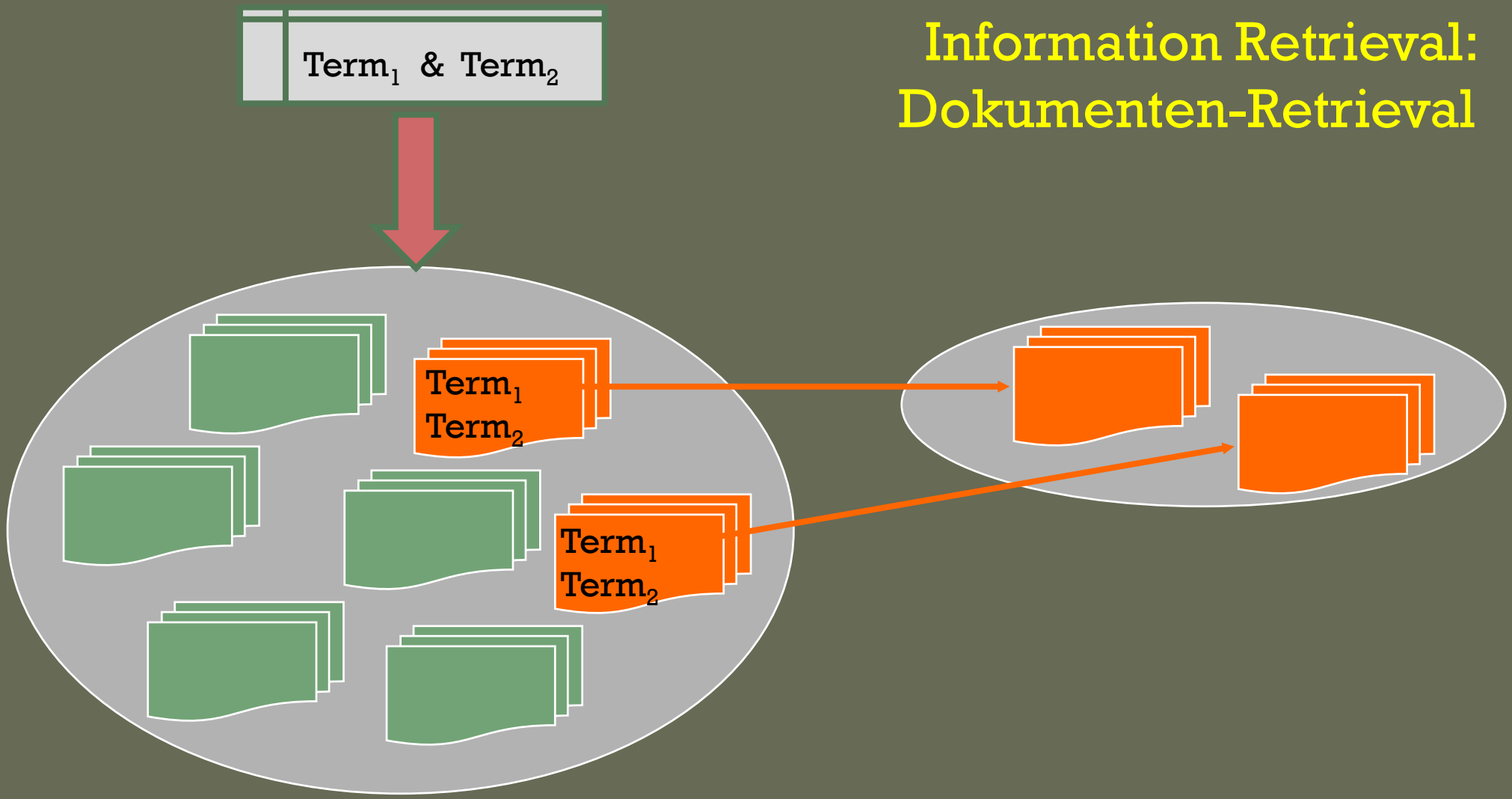
Sammeln von Dokumentkollektionen vs. Erschließung von Dokumentinhalten



Information Retrieval: Dokumenten-Retrieval



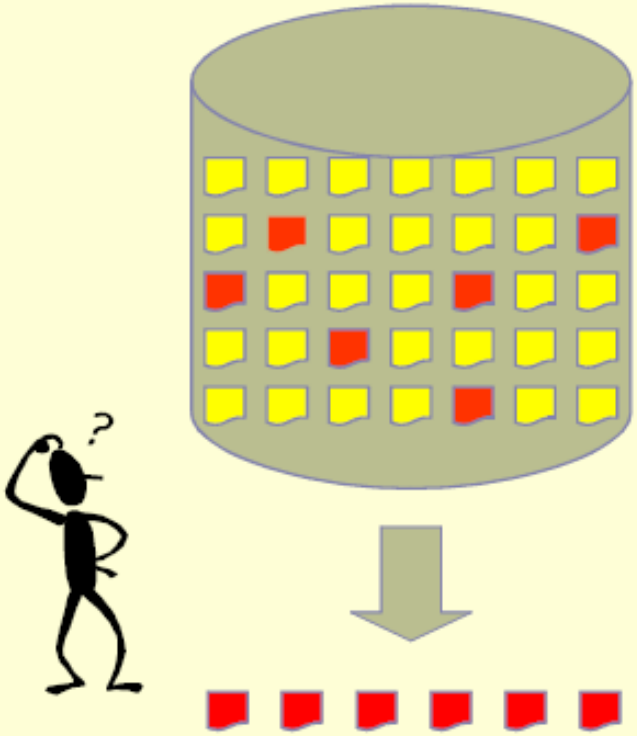
Information Retrieval: Dokumenten-Retrieval



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

Ad-hoc retrieval

One time queries (e.g. Web search)



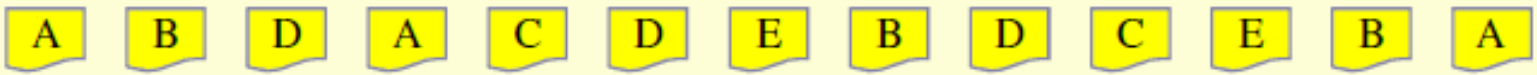
Filtering/Routing

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

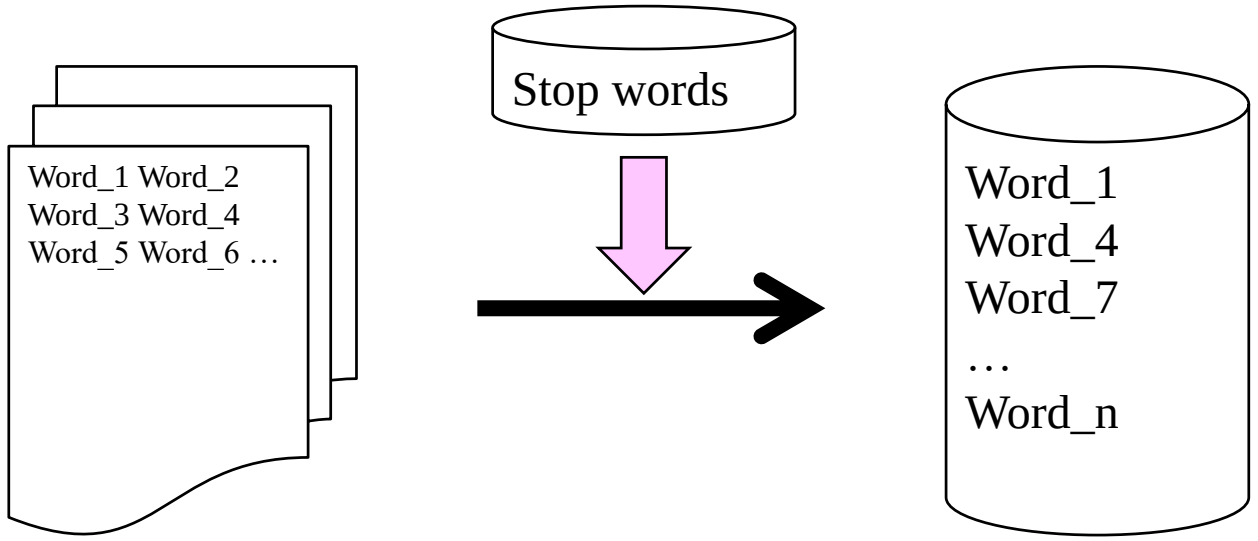
- Determine main topic(s)
- What‘s a relevant issue?
- Based on human (speed) reading and understanding of the document

AUTOMATIC INDEXING

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

BAG OF WORDS

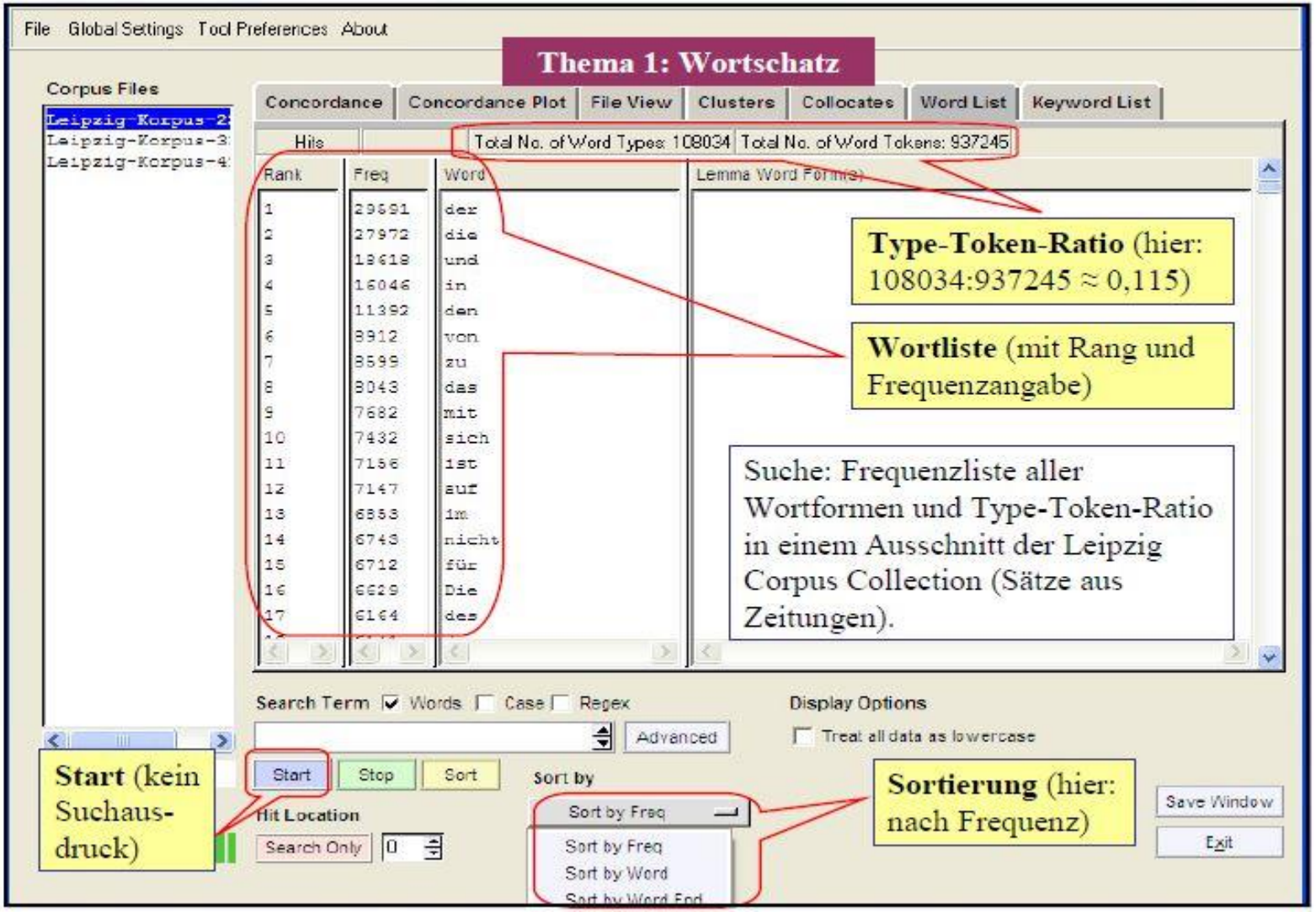
- Eliminate sequential structure of texts



AUTOMATIC INDEXING

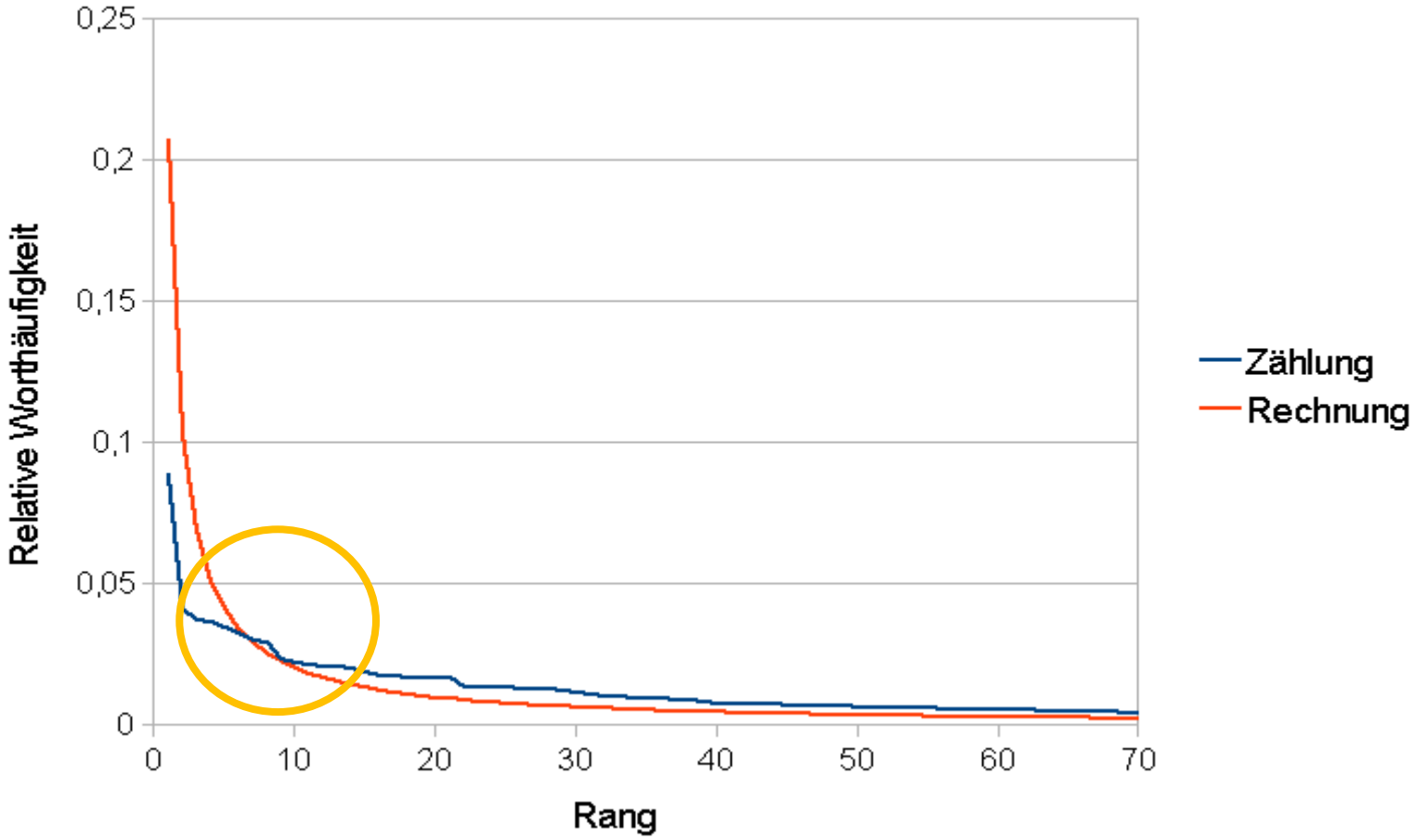
- Absolute vs. relative frequency
  - Per document
  - Relative to document collection
  - Bag-of-words (BOW)
  - Eliminate stop words (high occurrence frequency!)

Lexikalische Frequenzanalyse: Stoppwörter höchstfrequent



http://www1.ids-mannheim.de/fileadmin/lexik/lehre/engelberg/Webseite_Korpusanalyse/Korpusanalyse_4_Methoden_AntConc.pdf

Zipf's Law


0,25
0,2
0,15
0,1
0,05
0
Relative Worthäufigkeit
0
10
20
30
40
50
60
70
Rang
Zählung
Rechnung

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

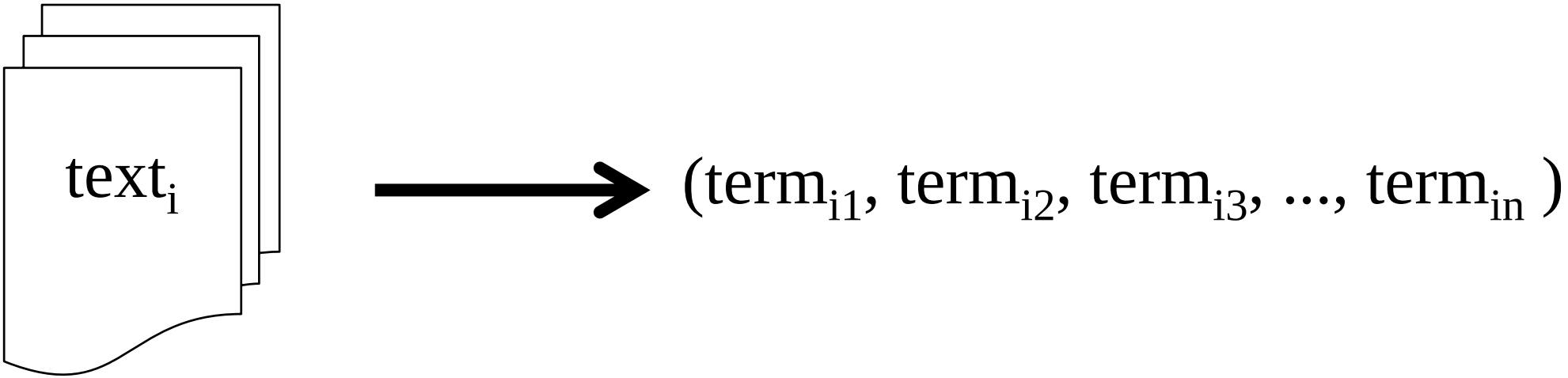
N = number of documents in collection

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)

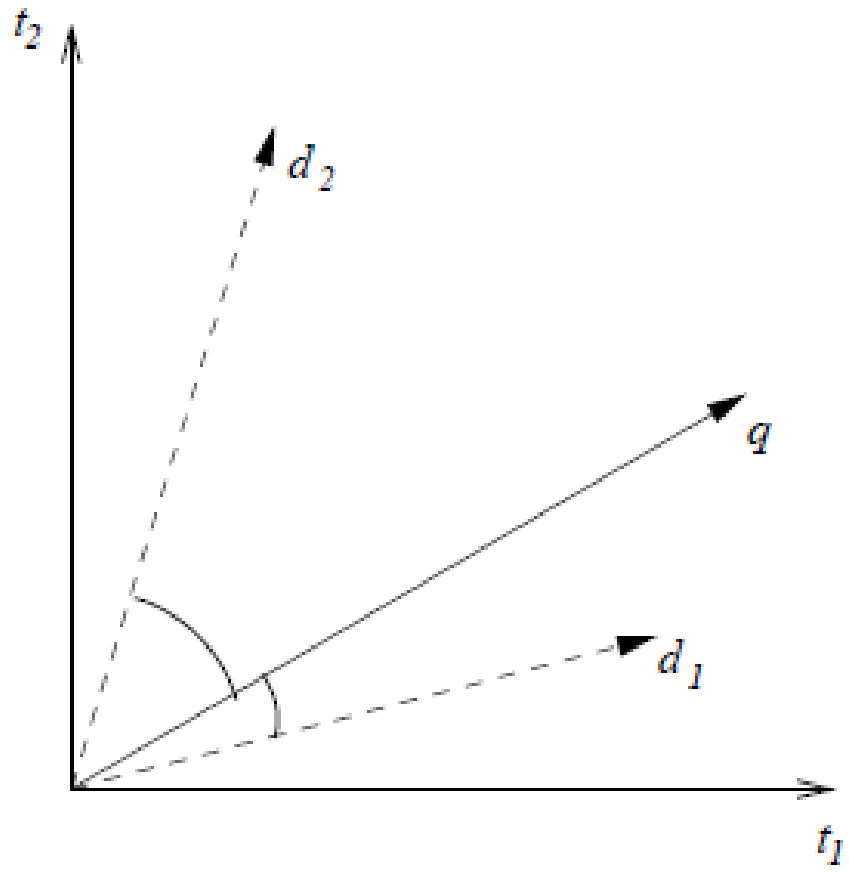


AUTOMATIC INDEXING (Vector Space Model)

- 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)
  - $doc_i = (term_{i1}, term_{i2}, term_{i3}, ..., term_{in})$
    - Each component $term_{ik}$ is either ‚0‘ (absent) or ‚1‘ (realized)
- Compute the association between a document term and a query term vector (query = $(query_1, query_2, query_3, ..., query_n)$, *n* as above), e.g., using the cosine measure

$$SIM(doc_i, query) = \frac{\sum_{k=1}^{t}(term_{ik} \bullet query_k)}{\sqrt{\sum_{k=1}^{t}(term_{ik})^2 \bullet \sum_{k=1}^{t}(query_k)^2}}$$

GRAPHICAL INTERPRETATION



CLASSIFICATION

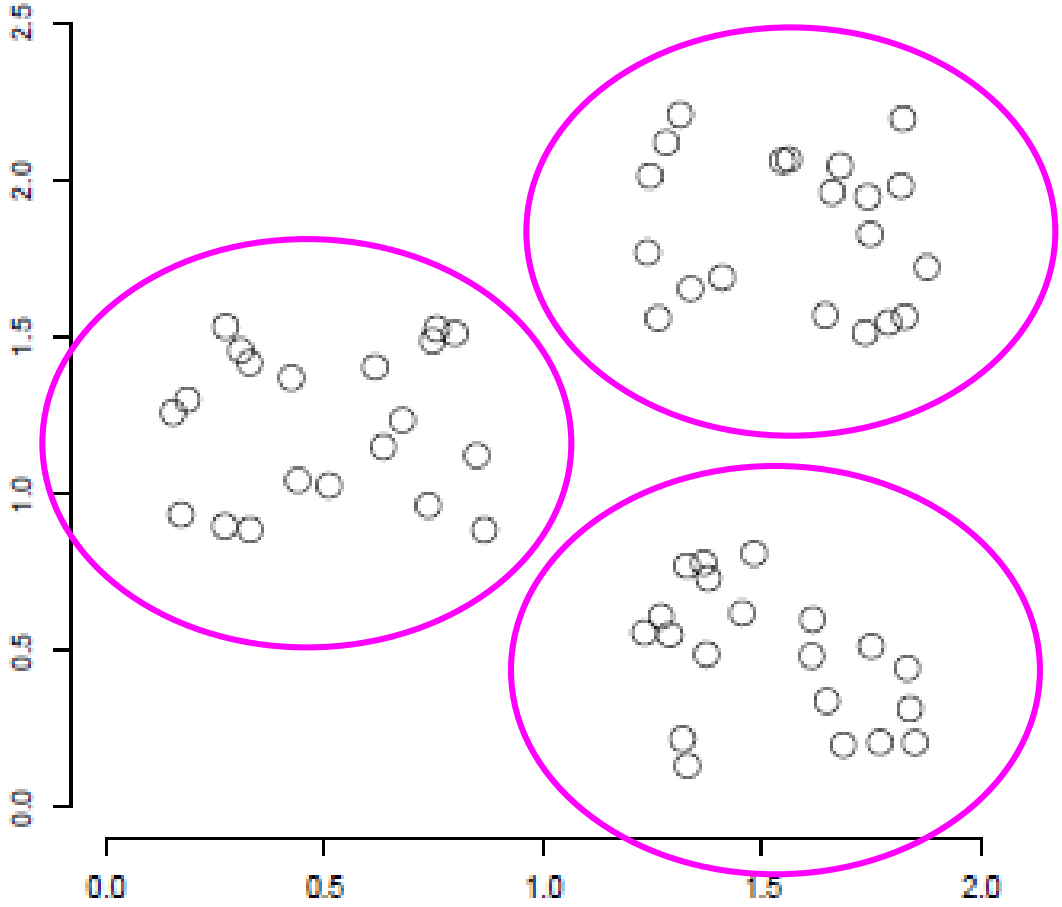
- Manual classification
  - Manual assignment of docs to pre-defined categories (classes)

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

Clustering

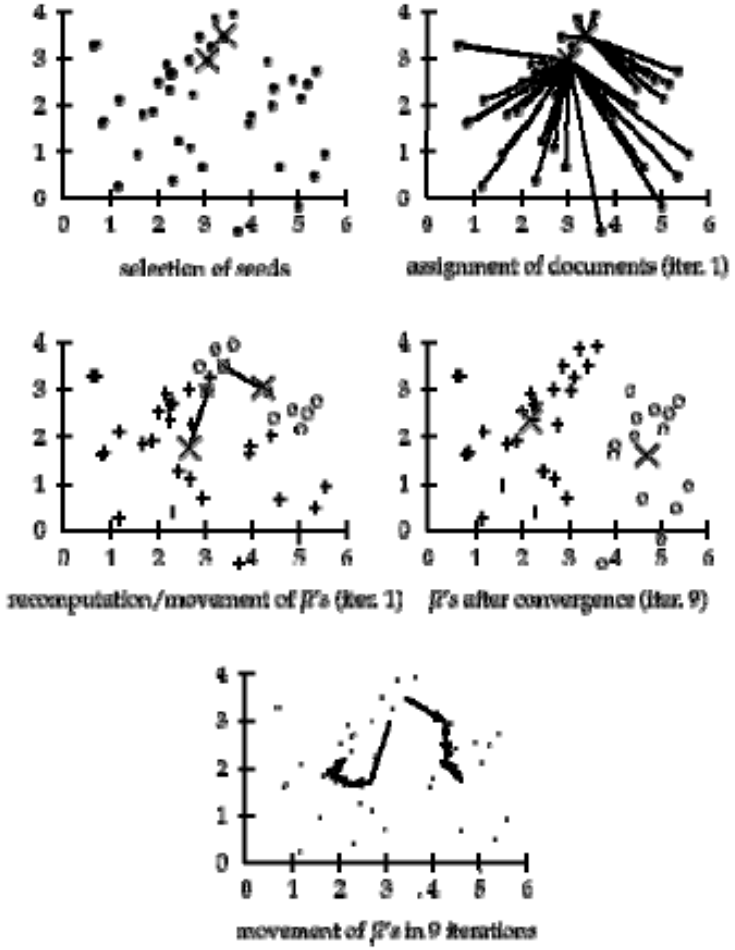
- (Document) clustering is the process of grouping a set of documents into clusters of similar documents.
- Documents within a cluster should be similar.
- Documents from different clusters should be dissimilar.
- Clustering is the most common form of unsupervised learning.
- Unsupervised = there are 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

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