

Lernzettel

Recherche-Systeme: Aufbau, Indizierung,
Suchlogik und Informationsdarstellung

Universität:	Technische Universität Berlin
Kurs/Modul:	Einführung in die Informatik
Erstellungsdatum:	September 19, 2025



Zielorientierte Lerninhalte, kostenlos!
Entdecke zugeschnittene Materialien für deine Kurse:

<https://study.AllWeCanLearn.com>

Einführung in die Informatik

Lernzettel: Recherche-Systeme: Aufbau, Indizierung, Suchlogik und Informationsdarstellung

(1) Aufbau des Recherchesystems.

- Client- und Server-Komponenten: Benutzeroberfläche, Request-Generator, Netzwerkkommunikation.
- Zentrale Komponente: Such- bzw. Index-Service, der Anfragen verarbeitet und Indizes verwaltet.
- Indexstruktur: Inverted Index, Posting-Listen pro Term, ggf. Segmentierung.
- Ranking- bzw. Relevanz-Modell: Bewertet die Treffermenge nach Relevanz.
- Persistenz: Dokumentation, Metadaten, Speicherformate.

(2) Indizierung.

- Tokenisierung: Zerlegung von Text in Tokens (Wörter, Phrasen).
- Normalisierung: Kleinschreibung, Stemming/Lemmatisierung, Stoppwort-Filterung.
- Inverted Index: Zu jedem Term die Liste der Dokumente, in denen er vorkommt.
- Merkmale eines Postings: {Dokument-ID, Term-Frequenz, ggf. Positionsinformationen}.
- Wichtige Größen:

$$\text{tf}(t, d) = \text{Häufigkeit von } t \text{ in } d$$

$$\text{df}(t) = |\{d : t \in d\}|$$

$$\text{IDF}(t) = \log \frac{N}{\text{df}(t)}$$

$$\text{TFIDF}(t, d) = \text{tf}(t, d) \cdot \text{IDF}(t)$$

(3) Suchlogik.

- Abbilden von Abfragen: Boolean-Retrieval (AND, OR, NOT) vs. Ranked Retrieval.
- Proximity und Phrasen: Suchanfragen mit Anführungszeichen, Nähe-Operatoren.
- Ranking-Modelle:

$$\text{BM25}(q, d) = \sum_{t \in q} \text{IDF}(t) \cdot \frac{\text{tf}(t, d) (k_1 + 1)}{\text{tf}(t, d) + k_1 (1 - b + b \frac{|d|}{\text{avgdl}})}$$

mit Parameteren $k_1 \in [1.2, 2.0]$, $b \in [0, 1]$.

(4) Informationsdarstellung.

- Snippets, Kontextvorschau, Relevanzanzeige.

- Ergebnislisten: Sortierung, Paginierung, Hervorhebung von Treffern.
- Filtern und Facetten: Zeitraum, Dokumententyp, Autor, Sprache.
- Visualisierung: Graphen, Ähnlichkeitsnetze, semantische Verknüpfungen.

(5) Architektureller Überblick und Anwendungsbeispiele.

- Typische Architektur: Frontend <-> API-Schicht <-> Such-Engine <-> Index-Speicher <-> Dokumenten-Repository.
- Workflows: Indexierung eines Korpus, Bearbeitung von Anfragen, Aktualisierung des Index.