

Sprachtechnologie in Suchmaschinen

Übersicht

Masterseminar Suchmaschinen
Sommersemester 2021

Stefan Langer
stefan.langer@cis.uni-muenchen.de

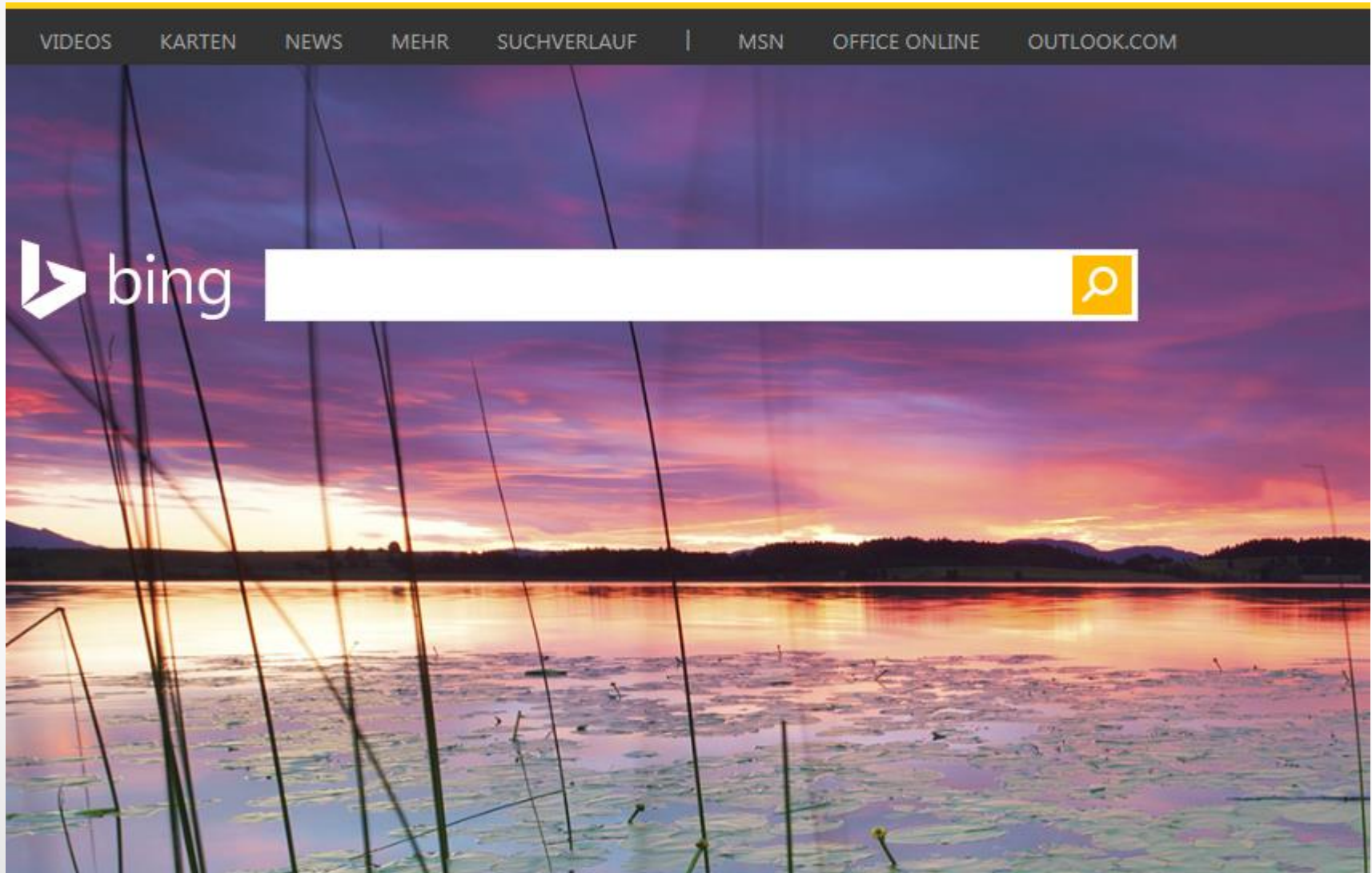
Suchmaschinen

Beispiele

Info zum Seminar

- Kontakt Stefan Langer
- stefan.langer@cis.uni-muenchen.de
- Schein: Implementierung einer Suchmaschine + ausführliche Darstellung (Referat)

Websuche - Bing



Websuche - Google



DuckDuckGo



DuckDuckGo



Die Suchmaschine, die Sie nicht verfolgt. [Mehr erfahren.](#)

Beispiel: Produktsuche

✓ Alles immer portofrei!* ✓ Kostenloser Rückversand ✓ Zahlung auch auf Rechnung

Anmelden



In allen Kategorien

Titel, Autor, Stichwort, ISBN

Los

Bücher eBooks Hörbücher Kinderbücher Ratgeber Schule Kalender **Musik** Filme Software Games Spielzeug E



Shopping

Suchbegriff / Artikelnr. eingeben



Inspiration . Damen . Herren . Kinder . Wäsche/Bademode .
Multimedia . Haushalt . Küche . Möbel . Heimtextilien . Bauma



Alle



Alle
Kategorien

Stefans Amazon

Angebote

Gutscheine

Verkaufen

DE

GUNNLAUGS SAGA ORMSTUNGU.

1. Þorsteinn hét maðr, hann var Egilsson Skallagrims-
sonar Kveldúlfssonar hersis ór Noregi; en Asgerðr hét móðir
Þorsteins ok var Bjarnardóttir. Þorsteinn bjó at Borg ok
Borgarfirði; hann var auðigr at fé ok höfðingi mikill, vitr
5 maðr ok hógværr ok hófsmaðr um alla hluti. Engi var
hann afreksmaðr um vöxt eða afl, sem Egill faðir hans, en
þó var hann it mesta afarmenni ok vinsæll af allri alþýðu.
Þorsteinn var vænn maðr, hvít á hár ok eygr manna bezt.
Hann átti Jófriði Gunnars dóttur Hlifarsonar. Jófriðr var
10 átján vetra, er Þorsteinn fékk hennar; hón var ekkja;
hana hafði átt fyrr Þóroddr, son Tungu-Odds, ok var þeirra
dóttir Húngerðr, er þar fœddist upp at Borg með Þorsteini.
Jófriðr var skörungur mikill; þau Þorsteinn áttu mart barna,
en þó koma fá við þessa sögu; Skúli var elztr sona þeirra,
15 annarr Kollsveinn, þriði Egill.

2. Eitt sumar er þat sagt, at skip kom af hafi í Guf-

- Download PDF - 4.7M
- Nur-Text-Format anzeigen



- Über dieses Buch
- Rezension schreiben
- Zu meiner Bibliothek hinzufügen

Inhalt

Dieses Buch kaufen

- ZVAB
- Google Produktsuche

Dieses Buch in einer Bibliothek finden.

Buchhandlungen in Ihrer Nähe suchen

Dieses Buch durchsuchen

Deutschlands Jobbörse Nr.1

62.284
Jobs in Deutschland

Was	Wo	
<input type="text" value="(Jobtitel, Firmenname oder ID)"/>	<input type="text" value="(Ort oder 5-stellige PLZ)"/>	<input type="button" value="Suchen"/>
		Erweiterte Suche


Finden Sie eine passende Stelle

* Sie suchen

Suchbegriff(e)

Arbeitsort

[Erweiterte Suche](#)





muenchen.de

Home Stadtplan Branchenbuch Hotel Webcam muenchen.de als Startseite

RATHAUS STADTLIBEN TOURISMUS WIRTSCHAFT MARKTPLATZ Kinderportal

- RATHAUS
- STADTLIBEN
- TOURISMUS
- WIRTSCHAFT
- MARKTPLATZ

Winter in München

Themen-Portale:
Veranstaltungen, Tickets
Restaurant, Café
Shopping
Verkehr, Mobilität
Stadtteile, Vereine
Finanzen
Besser leben mit M.
M// Card

Online-Services:
Hotel München
Hotels reservieren
Stadtplan
Branchenbuch München
Kino München

Grüß Gott

beim offiziellen Stadtportal für München

MÜNCHEN AKTUELL

Hier schreibt der OB

Jahresvorschau III
Baby-Boom: Über den Ausbau der Kinderbetreuung. [...](#)

Wochenendtipps
Am Wochenende noch nichts vor? Unsere Tipps... [...](#)

Fasching
Faschingskalender, Bildergalerien und viele Tipps. [...](#)

Fasching
Faschingskalender, Bildergalerien und viele Tipps. [...](#)

ANZEIGE

Die Fledermaus
Fasching mit Prinz Orlofsky und seinem Maskenball. [...](#)

Sehenswürdigkeiten
Die interessantesten oder schönsten Orte in München. [...](#)

Mitarbeiter gesucht
Das Stadtportal sucht engagierte Vertriebsmitarbeiter. [...](#)

Suche auf muenchen.de

Stadtplan München

Ticket Schnellsuche

Virtuelles Rathaus

muenchen.de Stadt-Branchenbuch

SHOPPING



Site-Suche (Bsp. Zeitung)

The screenshot shows the top section of the Süddeutsche Zeitung website. At the top left, there is a weather icon and the text "München 12°". The main header features the newspaper's name "Süddeutsche Zeitung" in a large serif font, with "SZ.de Zeitung Magazin" underneath. A navigation bar below the header contains a home icon and several menu items: "Politik", "Wirtschaft", "Panorama", "Sport", "München", "Bayern", "Kultur", "Wissen", and "Digital".

Below the navigation bar, the breadcrumb "Home > Schlagzeilen" is visible. There are three main content buttons: "Newsscanner", "Leser lesen aktuell" (which is highlighted), and "Leser empfehlen".

The search interface is located at the bottom of the page. It consists of a search input field with the placeholder text "Suchbegriff eingeben" and a question mark icon on the right. To the right of the input field is a dark button labeled "Finden" with a magnifying glass icon. Below the search bar, there are four filter options, each with a downward arrow: "Ressort", "Typ", "Quelle", and "Datum".

On the right side of the page, there is a mobile view overlay. It shows a smaller version of the website's search bar. A dark tooltip is positioned over the search bar, containing the text: "Fokus bewahren", "Entfernen Sie störende Elemente", and "Sie sich beim Lesen leichter".

Suchmaschinen – Weitere Anwendungsbereiche

- Mobile Suche (Smartphones)
- Suche im Intranet von Firmen und anderen Organisationen
 - Meist besondere Herausforderungen in Bezug auf Zugriffsrechte
- Desktop Suche (Suche auf dem privaten Computer)
- Soziale Netzwerke (e.g. Facebook) / professionelle Netzwerke (z.B. Xing, LinkedIn)
- Filesharing-Netzwerke

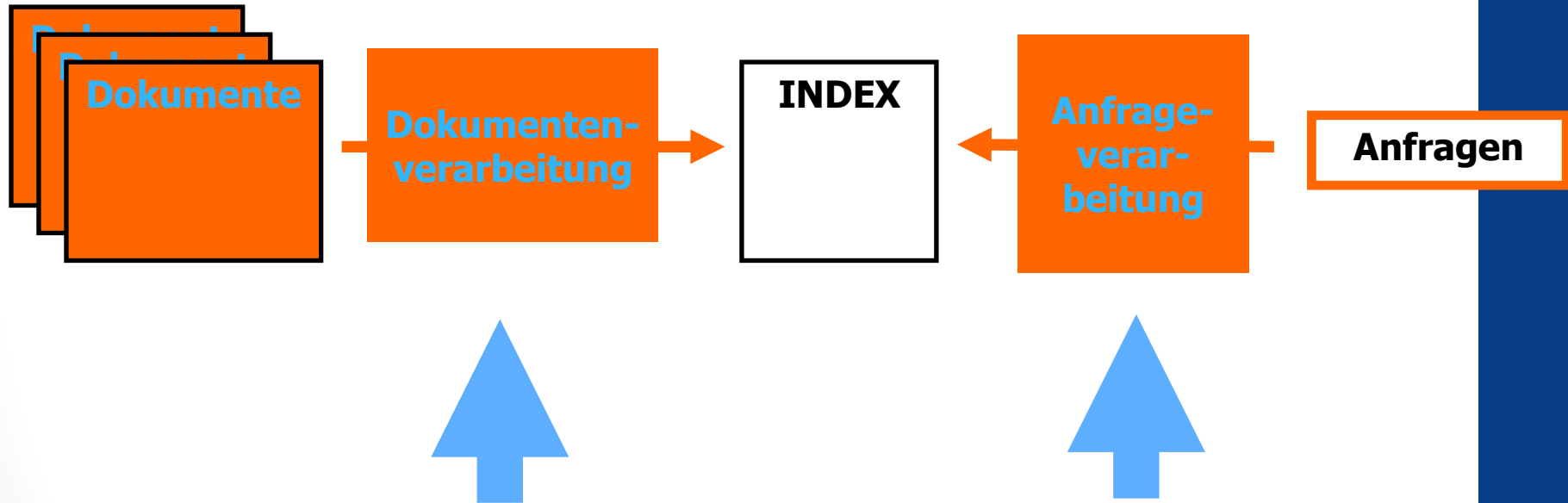
Suchmaschinen

Architektur und Anforderungen

Suchmaschinen - Software

- Webservices – siehe bisherige Beispiele
- Search Engine Software
 - Lucene
 - Elasticsearch
 - Solr
 - HP Autonomy
 - Sinequa
 - Coveo
 - Lookeen Server (Axonic)
 - FAST ESP †

Grobe schematische Architektur einer Suchmaschine



Dokumentenverarbeitung

- Erkennung von Dokumenteneigenschaften
(z.B. Sprachenidentifizierung, Dokumentformat)
- Konversion in intern verwendetes Dokumentenformat
(z.B. XML mit Unicode)
- Linguistische Normalisierung
 - Tokenisierung
 - Buchstaben(sequenzen)normalisierung
 - Morphologische Analyse
- Informationsextraktion
 - (z.B. Personennamen)
- Hinzufügen von Information
 - (z.B. Synonyme)

Anfrageverarbeitung

- Erkennung von Anfrageeigenschaften
 - (z.B. Sprache)
- Parsen der Anfrage
- Linguistische Normalisierung
 - Tokenisierung
 - Buchstaben(sequenzen)normalisierung
 - Rechtschreibkorrektur
 - Morphologische Analyse
 - Stopwortentfernung
- Hinzufügen von Information
(z.B. Synonyme)

Index

- Im Index werden Terme, die auf Dokumente verweisen mit der Referenz auf die Dokumente abgespeichert
 - Term: Einzelterme, Phrasen...
- Der Zugriff muss extrem effizient sein, um schnelle Anfrageverarbeitung zu ermöglichen

Linguistische Module in Suchmaschinen – Eine Übersicht

Sprachenidentifizierung

Tokenisierung

Morphologische Analyse

Rechtschreibkorrektur

Synonyme

Informationsextraktion

Ziel computerlinguistischer Module in Suchmaschinen

- Verbesserung der Ergebnisqualität
 - Recall
 - Precision
 - Ranking
 - ...
- Vorauswahl von Ergebnissen
- Navigation in den Ergebnissen

Übung 2: Linguistik in Suchmaschinen

- Was stellen Sie sich unter linguistischen Modulen in Suchmaschinen vor? Welche Module kennen Sie, welche machen Sinn?
- Wie tragen linguistische Funktionalitäten zur Ergebnisverbesserung bei?
 - Verbesserung der Ergebnisqualität
 - Vorauswahl von Ergebnissen
 - Navigation in den Ergebnissen

Ergebnisse Übung 2

- Dokumentenklassifikation
- Rechtschreibkorrektur
- Auto-Vervollständigung
- Suche mit regulären Ausdrücken/logischen Operatoren
- Spracherkennung
- Vollformen/Stammformen
- Klein-/Großschreibung
- Satzendeerkennung
- Datumsnormalisierung/Zahlennormalisierung allgemein
- Suchmaschine als Rechenmaschine
- natürlichsprachliche Anfragen
- Phrasierung
- Phonetische Suche
- Übersetzung
- - Dokumentenseite
- - Queryseitig
- Automatische Textzusammenfassung

Sprachenidentifizierung

Automatische Erkennung der Sprache eines elektronischen Dokuments

Sprachenidentifizierung

زبان‌شناسی

زبان‌شناسی (به انگلیسی: علمي) است که به مطالعه و بررسی روشمند زبان می‌پردازد. در واقع، زبان‌شناسی می‌کوشد تا به پرسش‌هایی بنیادین همچون «زبان چیست؟»، «زبان چگونه عمل می‌کند و از چه ساخت‌هایی تشکیل شده‌است؟»، «انسان‌ها چگونه با یکدیگر ارتباط برقرار می‌کنند؟»،

fa

Lingüística

La Lingüística és la ciència que estudia totes les manifestacions de la parla humana, és a dir, l'estudi de la llengua en el seu vessant escrit i oral. En un sentit ampli la lingüística és l'estudi de les llengües humanes, analitzant el que tenen en comú i el que les diferencia. Un lingüista és, per tant, una persona que estudia les llengües.

ca

Yezhoniezh

Ez-ledan e c'heller lâret ez eo ar yezhoniezh studi yezhoù mab-den.

Deskrivañ en un doare objektivel ha dielfennañ mont-en-dro ar yezhoù dres ma vezont implijet gant an dud hep en em soursial da varnañ

br

Identifiziere die Sprache eines Textes (Dokumententext, Anfrage ...)

Tokenisierung & Normalisierung

Tokenisierung

- Aufteilen eines Textes in indizierbare Token
- Recht trivial für westliche Sprachen; schwierig für Chinesisch, Japanisch, Thai

Normalisierung

- Groß- Kleinschreibung
- Akzente é → e
- Umlaute ä → a / ae
- (asiatische) Schriftzeichen in voller Breite/halber Breite
- □ ←→ □
 - Entsprechend auch lateinische Schriftzeichen im asiatischen Kontext
- Andere Zeichen
 - Scharfes ß u.ä.
 - Ohm-Zeichen, Angström-Zeichen

Morphologische Analyse

Grundformenreduzierung
Kompositasegmentierung

Grundformenreduzierung & Verwandtes

shop
shops

- kauppa NOM SG
- kauppa-ko NOM SG KO
- kauppa-kin NOM SG KIN
- kauppa-kaan NOM SG KAAAN
- kauppa-han NOM SG HAN
- kauppa-pa NOM SG PA
- kauppa-ko-han NOM SG KO HAN
- kauppa-pa-han NOM SG PA HAN
- kauppa-pa-s NOM SG PA S
- kauppa-ko-s NOM SG KO S
- kauppa-kin-ko NOM SG KIN KO
- kauppa-kaan-ko NOM SG KAAAN KO
- kauppa-kin-ko-han NOM SG KIN KO HAN
- kauppa-ni NOM SG SG1
- kauppa-ni-ko NOM SG SG1 KO
- kauppa-ni-kin NOM SG SG1 KIN
- kauppa-ni-kaan NOM SG SG1 KAAAN
- kauppa-ni-han NOM SG SG1 HAN
- kauppa-ni-pa NOM SG SG1 PA
- kauppa-ni-ko-han NOM SG SG1 KO HAN
- kauppa-ni-pa-han NOM SG SG1 PA HAN
- kauppa-ni-pa-s NOM SG SG1 PA S
- kauppa-ni-ko-s NOM SG SG1 KO S
- kauppa-ni-kin-ko NOM SG SG1 KIN KO
- kauppa-ni-kaan-ko NOM SG SG1 KAAAN KO
- kauppa-ni-kin-ko-han NOM SG SG1 KIN KO HAN
- ETC ETC

Grundformenreduzierung

„Stemming“

*Dokument***en**

*Suchmaschinen***en**

*Rahmen***en**

*Computers***s**

*Merkels***s**

Wörterbuchbasiert

Dokumenten:Dokument

Suchmaschinen:
Suchmaschine

Rahmen:Rahmen

Computers:Computer

Merkels:?

Wörterbuch + Regeln

Dokumenten:Dokument+en

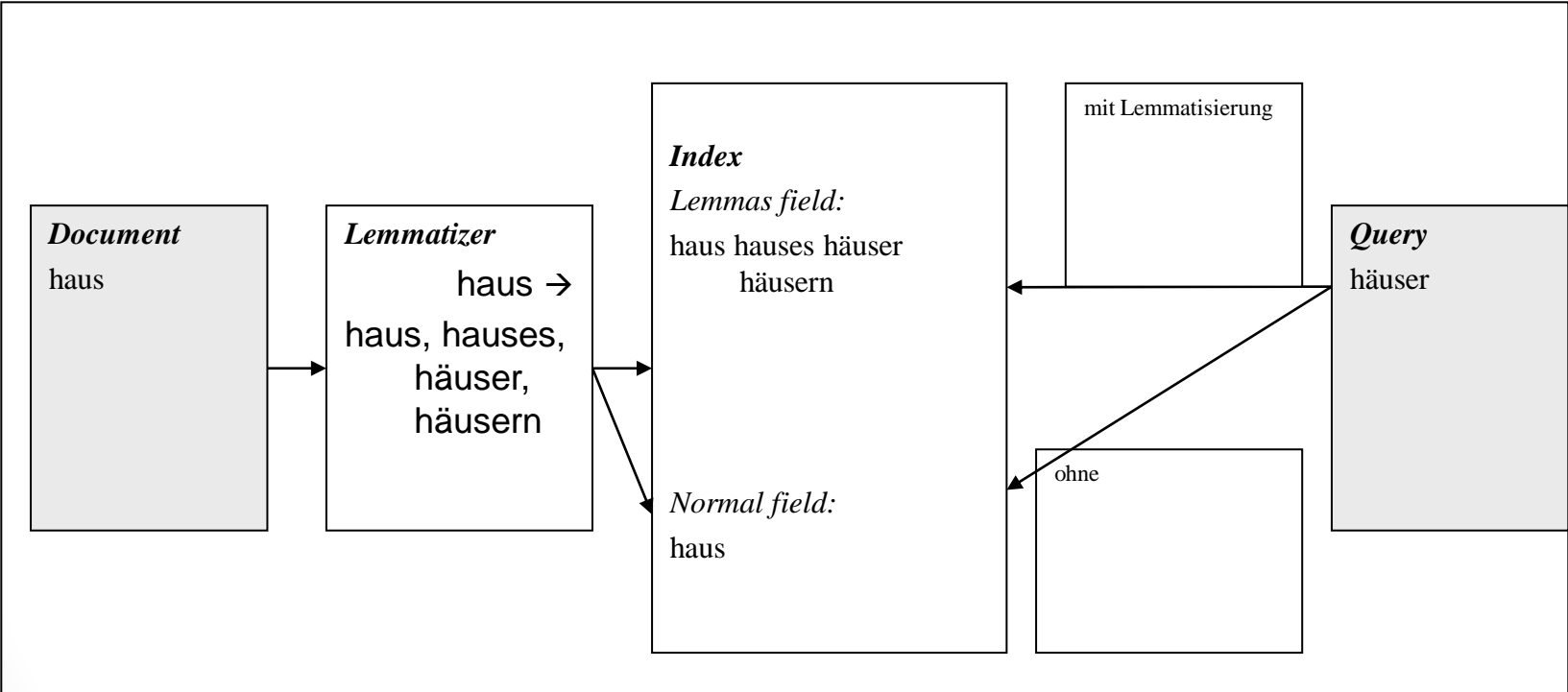
Suchmaschinen:
Suchmaschine+n

Rahmen:Rahmen+

Computers:Computer+s

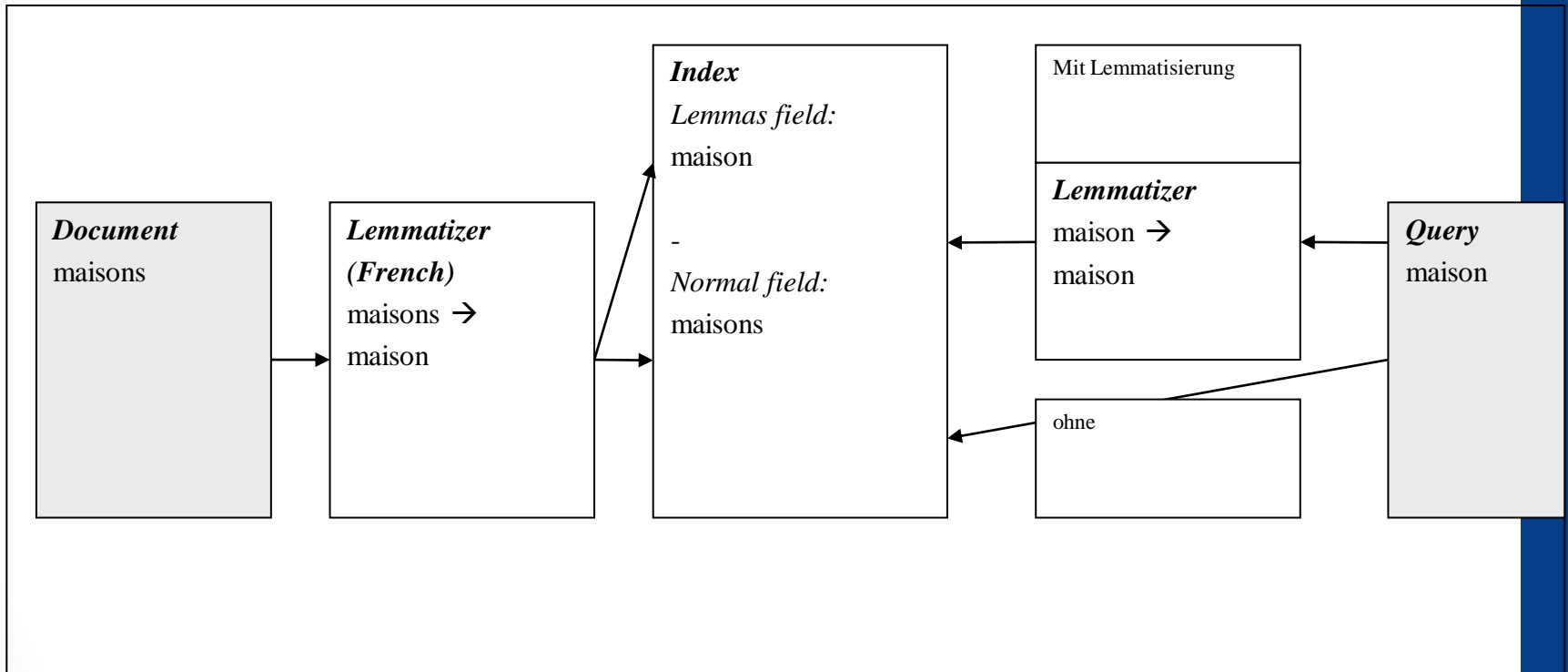
Merkels:Merkel+s

Lemmatisierung durch Expansion von Dokumententermen



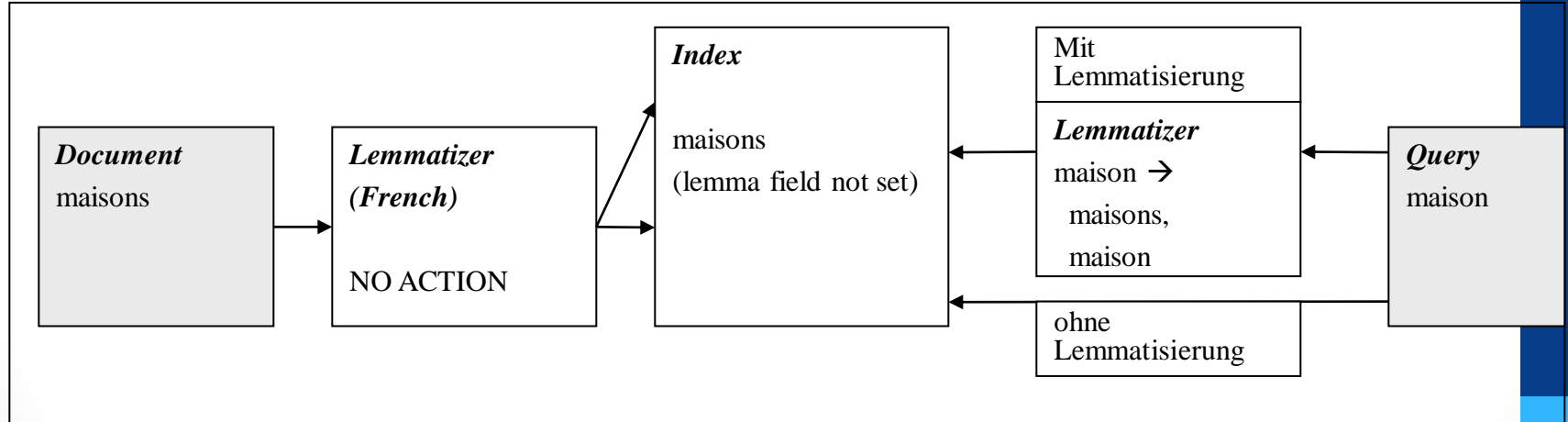
Alle Wortformen der Wörter im Dokument werden in den Index geschrieben. Die Sprache der Anfrage muss nicht bekannt sein

Lemmatisierung durch Reduktion



Wörter in Anfrage und Dokument werden auf die Grundform(en) reduziert. Dazu muss die Sprache der Anfrage bekannt sein

Lemmatisierung durch Anfrageexpansion



Nominalkompositanalyse

Blumen | versand

Internet | such | maschine

Fuchs | schwanz

Bahn | hof

Tisch | fuß | ball

Synonyme

Übung 3

- Was sind Synonyme?
- Was für Typen von bedeutungsähnlichen sprachlichen Einheiten, die in Suchmaschinen relevant sein könnten, gibt es außerdem?
- Welche Optionen gibt es, um Synonyme in die Suche einzubeziehen?

Synonyme und Verwandtes: Ergebnisse der Übung I

- Synonyme sind sprachliche Ausdrücke, die ohne Bedeutungsveränderung austauschbar sind.
 - Z.B. Zündholz/Streichholz
- Synonyme in Suchmaschinen: sollten gleichbedeutende Ausdrücke zu gleichen Suchergebnissen führen

Synonyme und Verwandtes: Ergebnisse der Übung II

Andere Bedeutungsähnlichkeiten:

- Alle Sinnrelationen: Hyponymie, Hyperonymie, Meronymie/Holonymie
- Abkürzungen und Akronyme (z.B. UNO United Nations Organisation)
- Paraphrasen
- Übersetzungen
- Umschreibungen
- Komposita $\leftarrow \rightarrow$ Kompositateile

- Technische Umsetzung von Synonymexpansion:
 - Expansion der Anfrage
 - Expansion der Terme im Dokument (\rightarrow Synonyme im Index)
 - Andere Einsatzmöglichkeiten: Zur Disambiguierung von Anfragen

Rechtschreibkorrektur

Rechtschreibkorrektur

- Vergleiche Anfrageterme mit bekannten Termen:
 - Mauresegler → Mauersegler
 - Merkel → Mergel

Voraussetzung:

- Abstandsmaß zwischen Termen
- Algorithmus zum schnellen Abgleich zwischen Lexikon und Anfrageterm

Zusätzlich:

Erstellung des Lexikons auf Basis der indizierten Terme

Phrasen-Rechtschreibkorrektur

Britnay Speers → Britney Spears

Rechtschreibkorrektur:

Verwandtes

- Phonetische Korrektur
- Phonetische Suche
- Anfragevervollständigung

Stopwörter

Stoppwörter und Stoppphrasen

- *Wo finde ich Informationen über Eric Rohmer*
- Eric Rohmer *und* Godard

Informationsextraktion

Extraktion von Eigennamen und weitergehende Ansätze

SCIRUS
for scientific information only

FEBS
Letters

Sign up to receive **FREE e-mail alerts** when new journal content becomes available to this site

About Us

Newsroom

Advisory Board

Submit Web Site

Help

Contact Us

Basic Search

[Advanced Search](#) [Search Preferences](#)

computational

Search

Journal sources Preferred Web sources Other Web sources Exact phrase

Searched for: All of the words **computational**

Found: **3,860,245 total** | [359,364 journal results](#) | [225,201 preferred web results](#) | [3,275,680 other web results](#)

Sort by: **relevance** | [date](#)

[Save checked results](#)

[Email checked results](#)

[Export checked results](#)

1. [institute for computational biomedicine \[15K\]](#)

Apr 2007

...Bin Abdulaziz Alsaud Institute for **Computational** Biomedicine houses the David A. Cofrin...Archives Mailing List Events May 2-4, 2007 **Computational** Biophysicists to Systems Biology Conference...science to discuss current trends in **computational** biophysics and systems biology. [CBSB07...
<http://icb.med.cornell.edu/>

[similar results](#)

2. [Computational Linguistics and Phonetics \[6K\]](#)

Jul 2007

...College -- Groups Bill Barry Matthew W. Crocker Manfred Pinkal Hans Uszkoreit -- Research Research Projects Publications **Computational** Linguistics Colloquium Phonetics Colloquium Software and Corpora Related Institutions -- Last modified 2007-06-05 by webmaster1...
<http://www.coli.uni-saarland.de/>

[more hits from](#) [<http://www.coli.uni-saarland.de/>]

Refine your search using these keywords found in the results:

[applied mathematics](#)

[bioinformatics](#)

[biophysics](#)

[computational biology](#)

[computational chemistry](#)

[computational fluid dynamics](#)

[computational geometry](#)

[computational linguistics](#)

[computational materials science](#)

[computational mathematics](#)

[computational mechanics](#)

[computational physics](#)

Henrik Johan Ibsen (* 20. März 1828 in Skien/Norwegen; † 23. Mai 1906 in Kristiania, damaliger Name von Oslo) war ein norwegischer Schriftsteller, der für den Naturalismus in Deutschland und Norwegen bedeutend war...

SUUCH.DE

Ibsen Geburtstag

Suuchen

1024 Treffer

Zusammenfassung

Henrik Ibsen wurde am 20. März 1828 in Skien/Norwegen geboren.

Quellen: wikipedia.de ; lexikon.meyers.de;

Treffer 1: Wikipedia...

Auch ausgereifte Suchmaschinen wie Google setzen Computerlinguistik ein (ein Sprachtechnologieprodukt der Firma Canoo, Basel). ...

Maschinelle Übersetzung

Maschinelle Übersetzung in Suchmaschinen

Mögliche Strategien

- Übersetzung der Originaldokumente und Indizierung der übersetzten Dokumente
 - Langsame Dokumentenverarbeitung
- Übersetzung des Index
 - → Ambiguität, wenn Kontext nicht berücksichtigt
- Übersetzung der angezeigten Dokumenteninhalte, evt. kombiniert mit der Übersetzung des gesamten Dokuments wenn ausgewählt
 - → verlangsamte Ergebnisverarbeitung
- Übersetzung der Anfragen
 - Hier zeigt sich besonders stark das Problem der Ambiguität

Klassifikation und Clustering

Klassifizierung

Zuweisung zu vordefinierten Kategorien

Dokumentenklassifizierung

- Erfordert vordefinierte, saubere Kategorien und Trainingsdokumente oder Auswahl exemplarischer Dokumente durch den Benutzer

Mögliche Dimensionen:

- Inhaltliche Themenbereiche
- Disziplinen
- Dokumententypen (z.B. wiss. Artikel, Zeitungsartikel, Adresssammlung)

Anfrageklassifizierung

Keywordsuche vs. semantische Suche

- Klassische Suchmaschinen arbeiten mit Keywordsuche (invertierte Dokument / inverted index)
- Alternativ ist es möglich mit Dokumenten- und Wortembeddings zu arbeiten
 - Neuerer Ansatz, noch stark im Flux

Nächstes Thema – bitte vorbereiten

An Introduction to Information Retrieval, Kapitel 1

Dieses Kapitel ganz lesen bis 8. Mai, bitte.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

<http://nlp.stanford.edu/IR-book/information-retrieval-book.html>