Suchmaschinen und Retrieval-Augmented Generation

Einführung

Masterseminar Suchmaschinen
Sommersemester 2024

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

Info zum Seminar

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

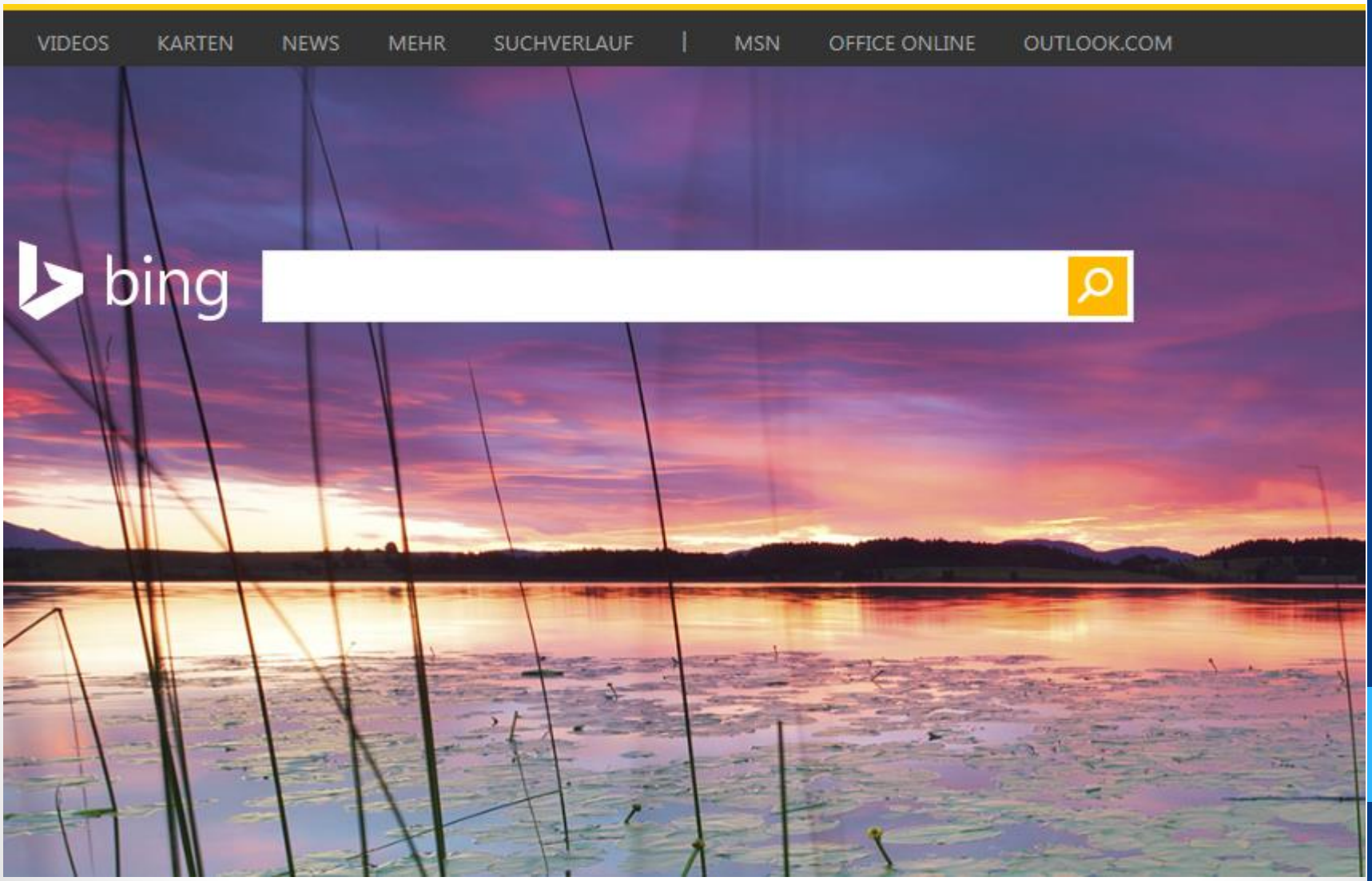
- Schein:
  - Implementierung eines RAG-Systems+ ausführliche Darstellung (Referat)
  - Vorstellung eines neueren Papers und Implementierung des Algorithmus

ChatGPT

Suchmaschinen

Beispiele

Websuche - Bing

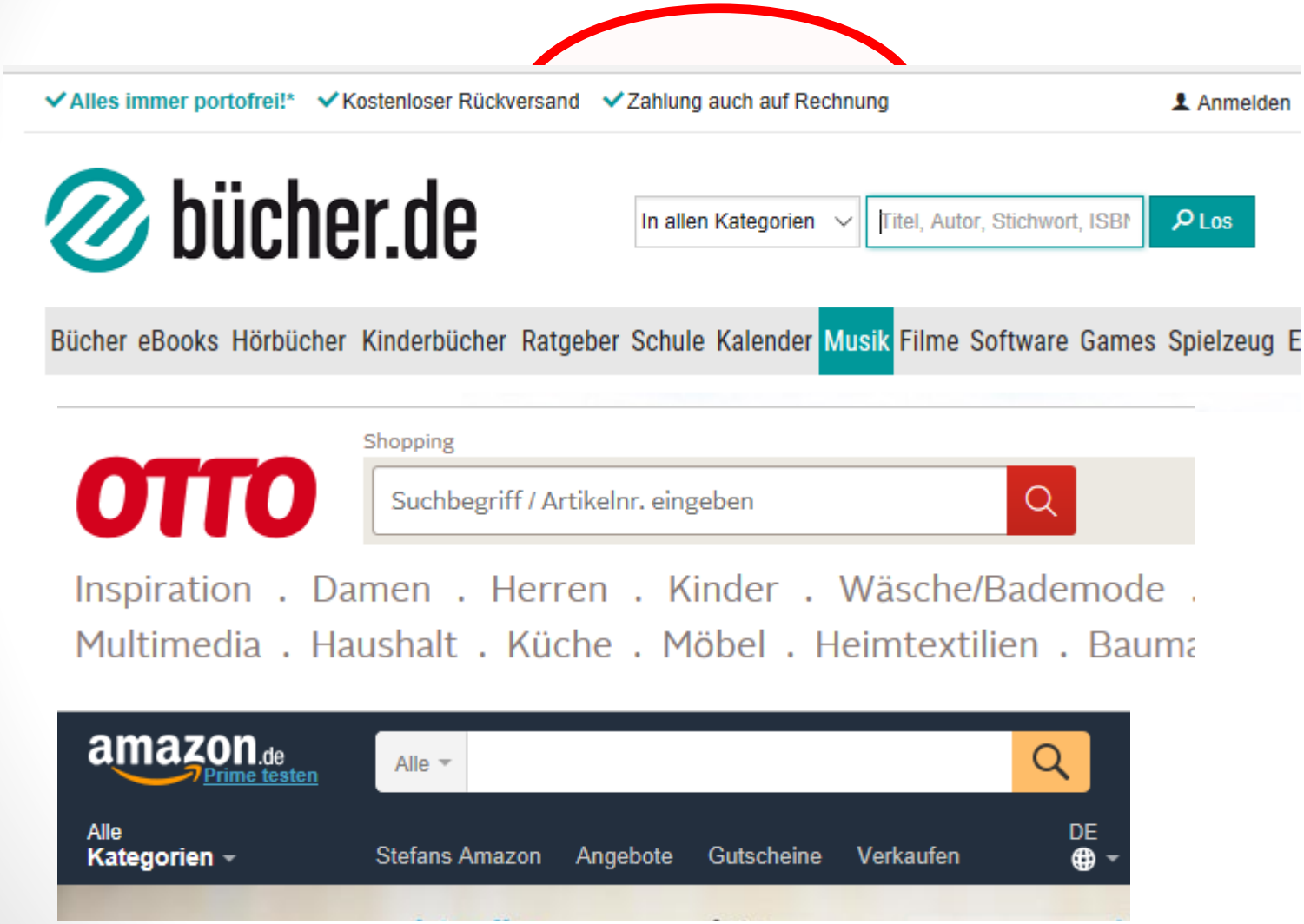


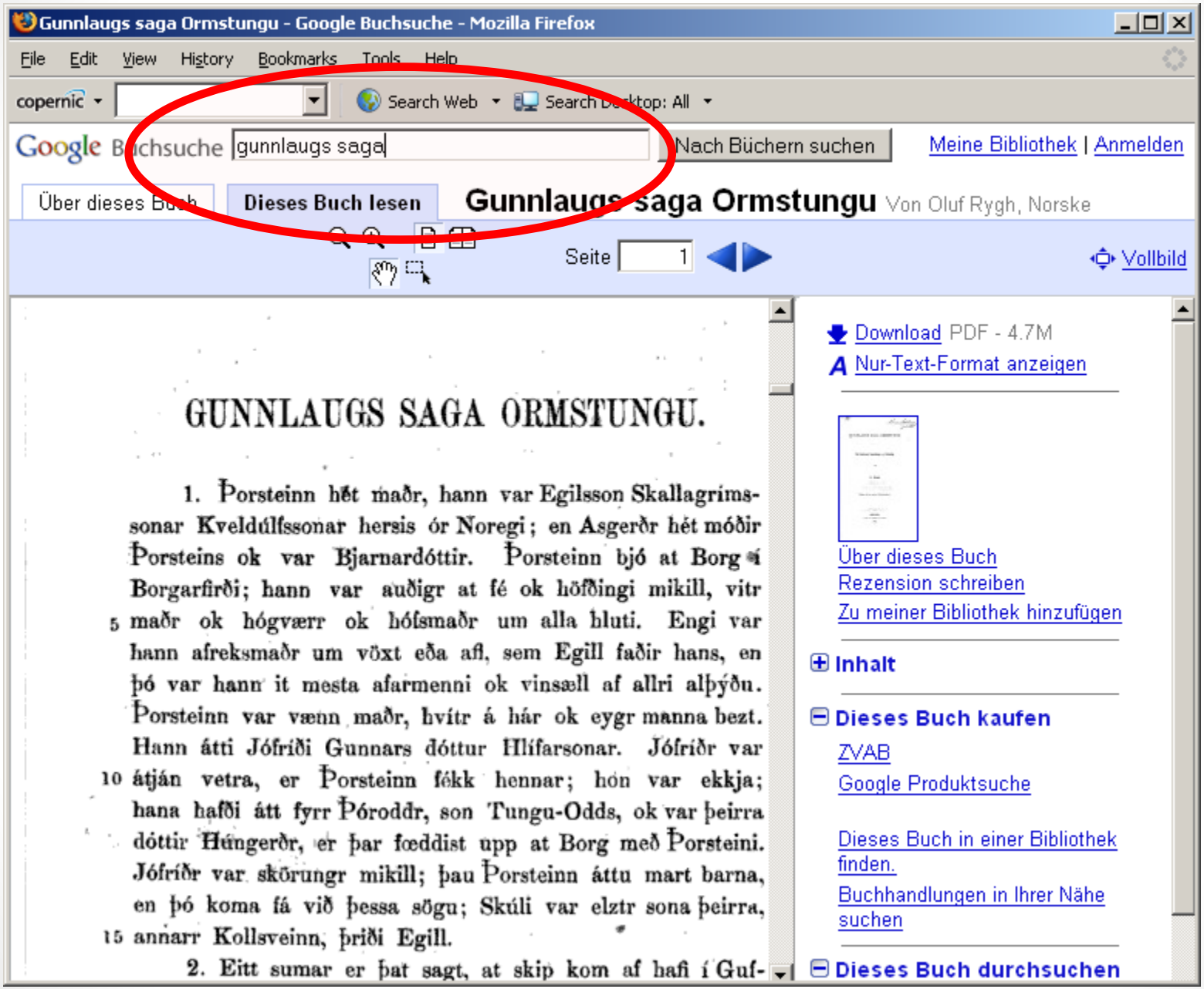
Websuche - Google

DuckDuckGo


DuckDuckGo
Die Suchmaschine, die Sie nicht verfolgt. Mehr erfahren.

Beispiel: Produktsuche




Gunnlaugs saga Ormstungu - Google Buchsuche - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Search Web
gunnlaugs saga
Nach Büchern suchen
Meine Bibliothek | Anmelden
Über dieses Buch
Dieses Buch lesen
Gunnlaugs saga Ormstungu Von Oluf Rygh, Norske
Seite 1
Vollbild
GUNNLAUGS SAGA ORMSTUNGU.
1. Þorsteinn hét maðr, hann var Egilsson Skallagríms-
sonar Kveldúlfssonar hersis ór Noregi; en Ásgerðr hét móðir
Þorsteins ok var Bjarnardóttir. Þorsteinn bjó at Borg á
Borgarfirði; hann var auðigr at fé ok höfðingi mikill, vitr
maðr ok hógværr ok höfsmaðr 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, hvítr á hár ok eygr manna bezt.
Hann átti Jófríði Gunnars dóttur Hlífarsonar. Jófríðr var
átján vetra, er Þorsteinn fékk hennar; hón var ekkja;
hana hafði átt fyrr Þóroddr, son Tungu-Odds, ok var þeirra
dóttir Hungerðr, er þar fæddist upp at Borg með Þorsteini.
Jófríðr var skörungr mikill; þau Þorsteinn áttu mart barna,
en þó koma fá við þessa sögu; Skúli var elztr sona þeirra,
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 (Jobtitel, Firmenname oder ID)

Wo (Ort oder 5-stellige PLZ)

Suchen

Erweiterte Suche

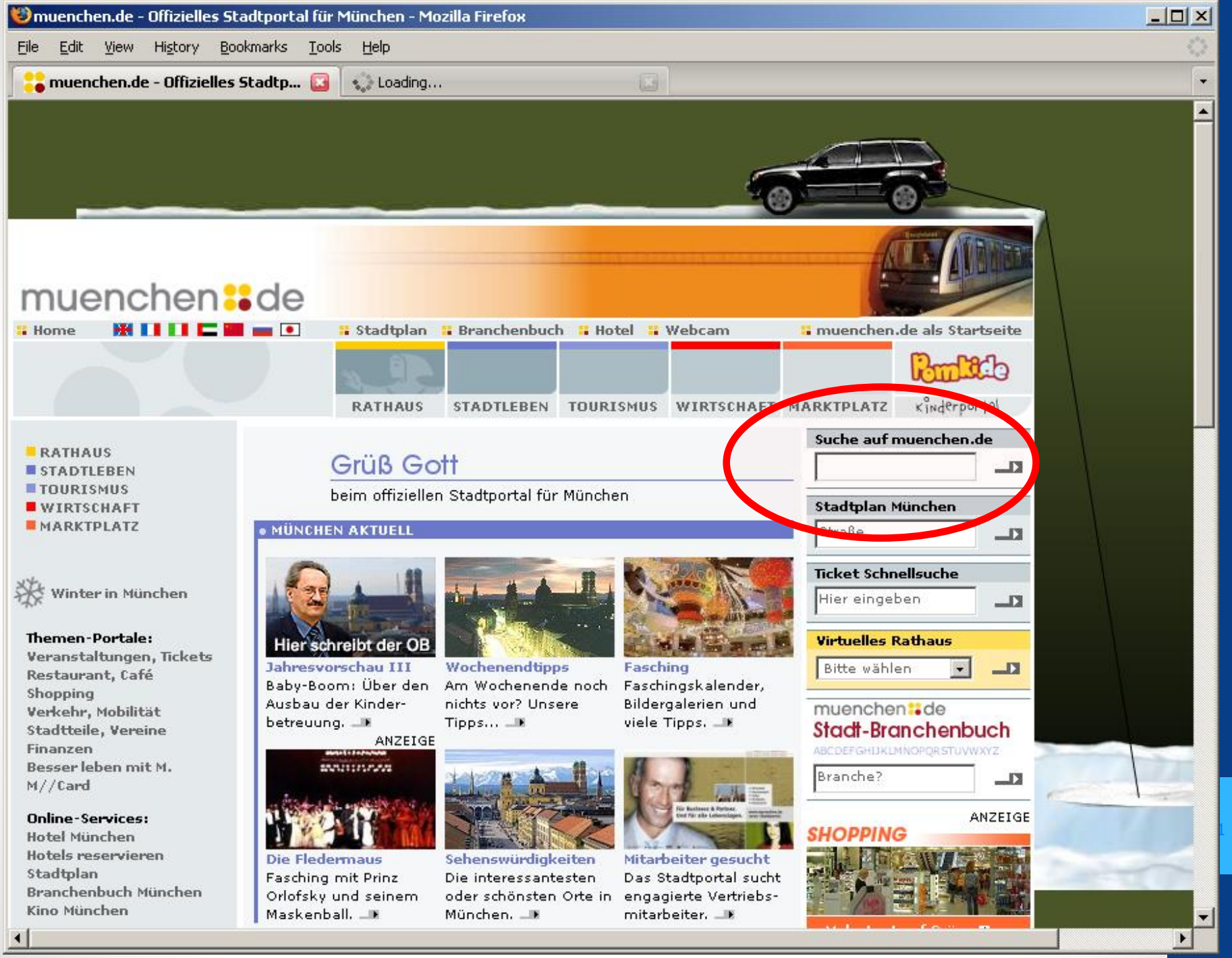
Finden Sie eine passende Stelle

* **Sie suchen**

Suchbegriff(e)

Arbeitsort

Suchen Erweiterte Suche


muenchen.de - Offizielles Stadtportal für München - Mozilla Firefox
File Edit View History Bookmarks Tools Help
muenchen.de - Offizielles Stadtp...
Loading...
muenchen.de
Home
Stadtplan
Branchenbuch
Hotel
Webcam
muenchen.de als Startseite
RATHAUS
STADTLEBEN
TOURISMUS
WIRTSCHAFT
MARKTPLATZ
RATHAUS
STADTLEBEN
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 Kinder-betreuung.
Wochenendtipps
Am Wochenende noch nichts vor? Unsere 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 Vertriebs-mitarbeiter.
Suche auf muenchen.de
Stadtplan München
Ticket Schnellsuche
Hier eingeben
Virtuelles Rathaus
Bitte wählen
muenchen.de
Stadt-Branchenbuch
Branche?
ANZEIGE
SHOPPING

Site-Suche (Bsp. Zeitung)

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 – QnA - Chatbot

	Suchmaschine	QnA-System	Chatbot
Anfrage	Keyword oder Satz/Frage	Frage	Frage
Antwort	Dokument + Kontext	Antwort aus Antwortliste	Antwort oder Rückfrage
	Einstufig	Eher einstufig	Mehrstufig
	Keyword-basiert Embeddings	Keyword Embedding	Keyword, Embeddings, Regeln, Entscheidungsbaum

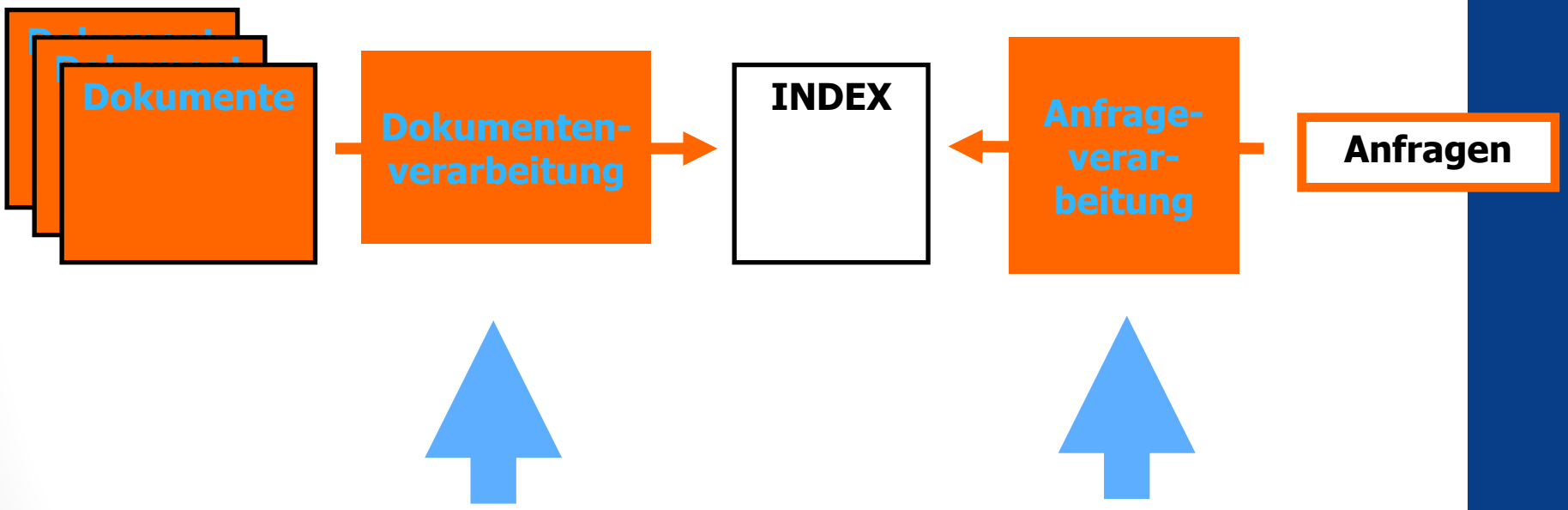
Suchmaschinen

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

Sprachenidentifizierung

Automatische Erkennung der Sprache eines elektronischen Dokuments

Sprachenidentifizierung

زبان‌شناسی
زبان‌شناسی (به انگلیسی:(علمی 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 KAAN
- 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 KAAN 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 KAAN
- 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 KAAN KO
- kauppa-ni-kin-ko-han NOM SG SG1 KIN KO HAN
- ETC ETC

Grundformenreduzierung

„Stemming“

Dokumenten

Suchmaschinen

Rahmen

Computers

Merkels

Wörterbuchbasiert

Dokumenten:Dokument

Suchmaschinen:
Suchmaschine

Rahmen:Rahmen

Computers:Computer

Merkels:?

Wörterbuch + Regeln

Dokumenten:Dokument+en

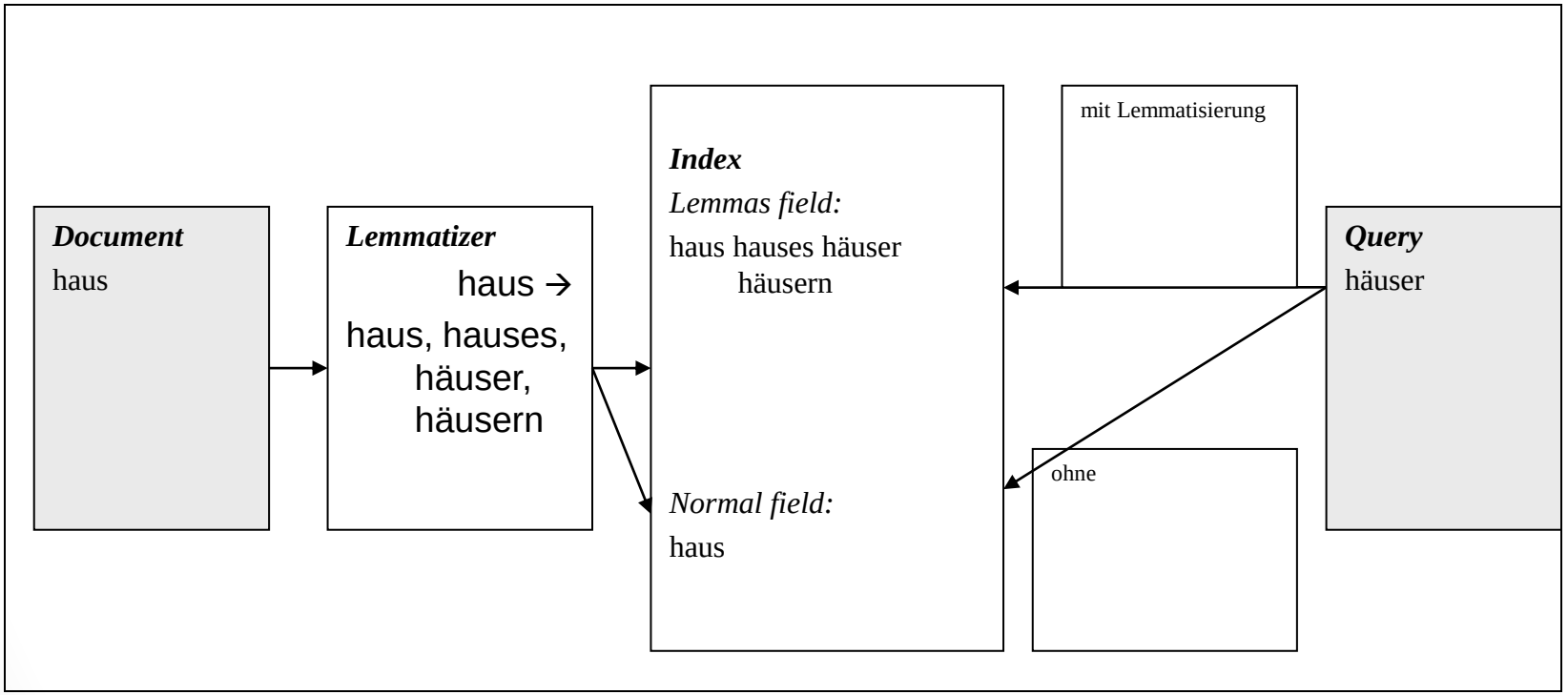
Suchmaschinen:
Suchmaschine+n

Rahmen:Rahmen+

Computers:Computer+s

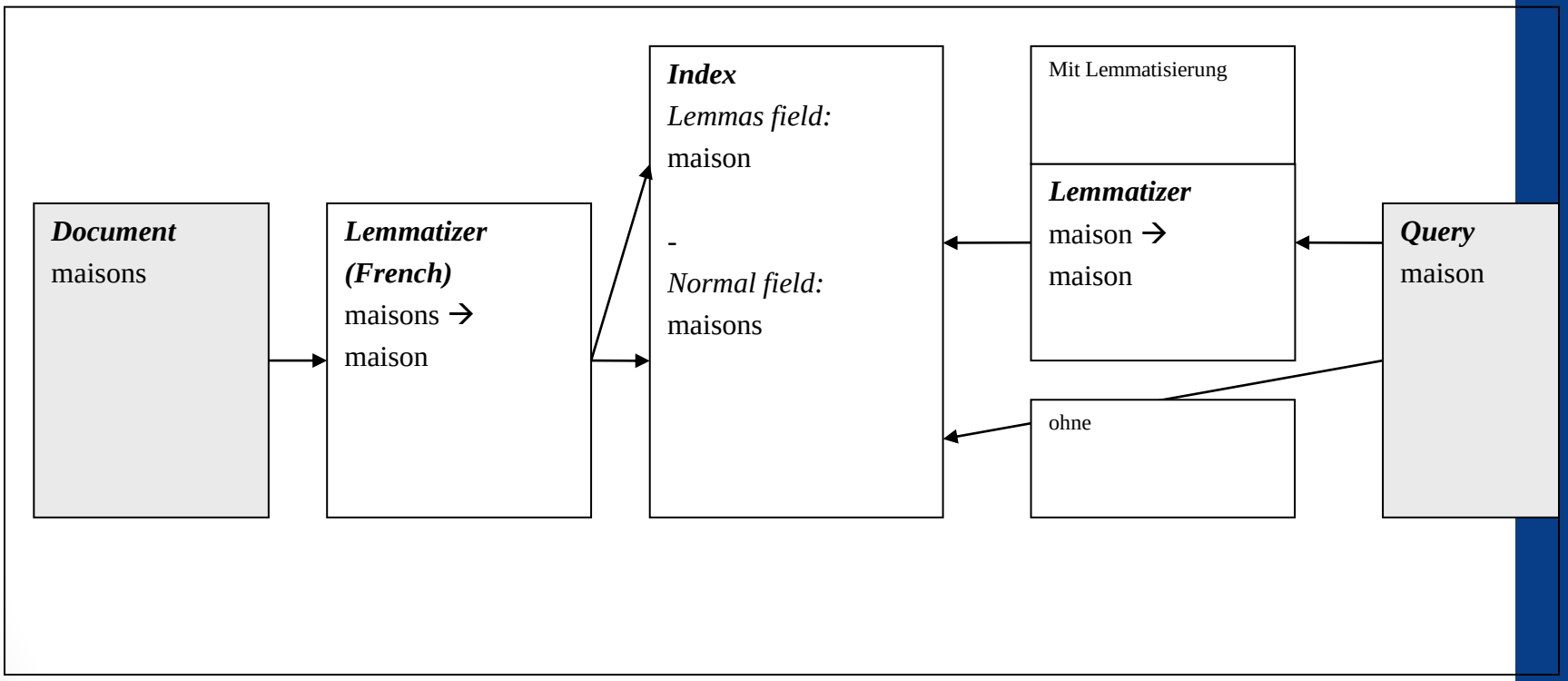
Merkels:Merkel+s

Lemmatisierung durch Expansion von Dokumententermen



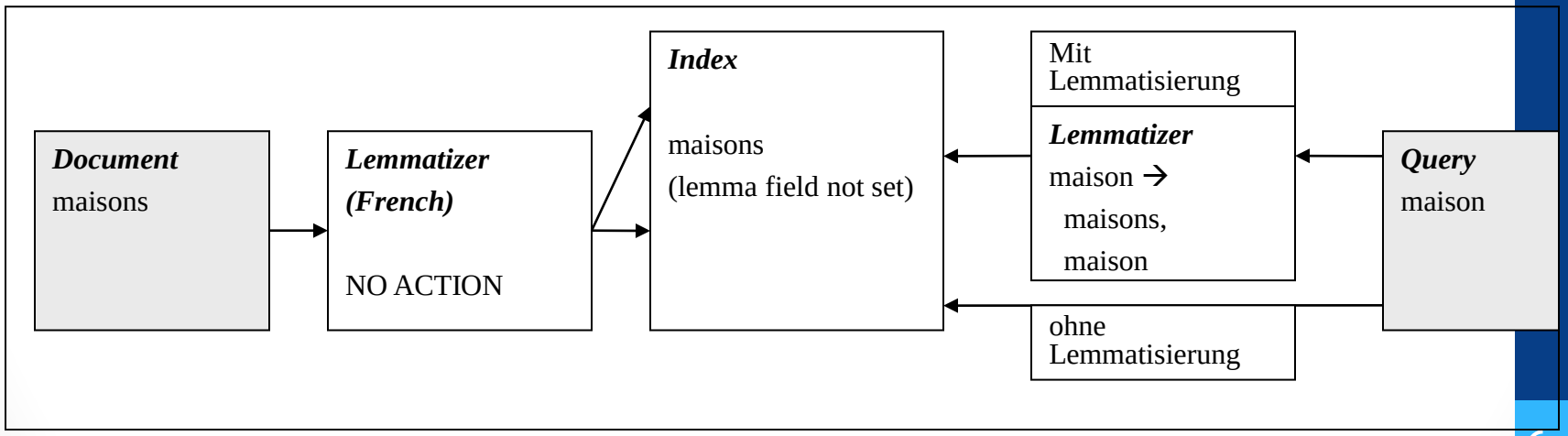
Alle Wortformen der Wörter im Dokument werden in den Index geschrieben. Die Sprache der Anfrage muss nicht bekannt sein

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

Synonyme 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:

Andere Bedeutungsähnlichkeiten:

- Alle Sinnrelationen: Hyponymie, Hyperonymie, Meronymie/Holonymie
- Abkürzungen und Akronyme (z.B. UNO United Nations Organisation)
- Paraphrasen
- Übersetzungen
- Umschreibungen
- Komposita ← → Kompositatteile

- Technische Umsetzung von Synonymexpansion:
  - Expansion der Anfrage
  - Expansion der Terme im Dokument (→ 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

SUUCH.DE

Ibsen Geburtstag

Suuchen

1024 Treffe

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

Keywortsuche vs. semantische Suche Suche vs. Chat

- Klassische Suchmaschinen arbeiten mit Keywortsuche (invertierte Dokument / inverted index)
- State of the Art: Dokumenten-Embedding für Retrieval und Re-Ranking.
- Frage-Antwort und Chatsysteme
- RAG-System (Retrieval Augmented Generation)

Question-answering architecture (Retrieval Augmented Generation) for large language models / GPT