

Apache Lucene og Solr

Jana Kunert
feymour@gmail.com

The Camp 2015

July 21, 2015

Table of Contents

Hvem er jeg

Lucene

Søgning

Indeksering

Querying

Solr

Hvorfor Solr?

Demo

Hvad skal jeg bruge?

Links

Table of Contents

Hvem er jeg

Lucene

Søgning

Indeksering

Querying

Solr

Hvorfor Solr?

Demo

Hvad skal jeg bruge?

Links

Hvem er jeg

- ▶ Cand.scient. i Datalogi fra Aarhus Universitet
- ▶ Speciale i Algoritmer og Datastrukturer
- ▶ Konsulent hos Netcompany
- ▶ Gentoo Linux bruger i 9 år (tror jeg).

Table of Contents

Hvem er jeg

Lucene

Søgning

Indeksering

Querying

Solr

Hvorfor Solr?

Demo

Hvad skal jeg bruge?

Links

Lucene

- ▶ Apache Lucene (Core) er et information retrieval bibliotek skrevet i Java
- ▶ skalerbar, high-performance indeksering



Søgning

- ▶ Indeksering
- ▶ Querying

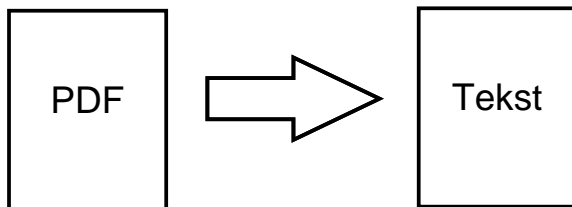


Indeksering

- ▶ Parsing
- ▶ Tokenisation
- ▶ Analyses



Indeksering – Parsing





Indeksering – Tokenisation

“Computer science is no more about computers than astronomy is about telescopes.”

- ▶ Computer
- ▶ science
- ▶ is
- ▶ no
- ▶ more
- ▶ about
- ▶ computers
- ▶ ...



Indeksering – Analysis

“Computer science is no more about computers than astronomy is about telescopes.”

- ▶ lower-case: Computer \Rightarrow computer



Indeksering – Analysis

“Computer science **is no more about** computers **then** astronomy **is about** telescopes.”

- ▶ lower-case: Computer \Rightarrow computer
- ▶ stop-words: is, no, more, about, then



Indeksering – Analysis

“Computer science **is no more about** computers **then** astronomy **is about** telescopes.”

- ▶ lower-case: Computer \Rightarrow computer
- ▶ stop-words: is, no, more, about, then
- ▶ stemming: telescopes \Rightarrow telescope



Indeksering – Analysis

“Computer science **is no more about** computers **then** astronomy **is about** telescopes.”

- ▶ lower-case: Computer \Rightarrow computer
- ▶ stop-words: is, no, more, about, then
- ▶ stemming: telescopes \Rightarrow telescope
- ▶ synonym: computer \Rightarrow computer, pc



Indeksering – Analysis

“Computer science **is no more about** computers **then** astronomy **is about** telescopes.”

- ▶ lower-case: Computer \Rightarrow computer
- ▶ stop-words: is, no, more, about, then
- ▶ stemming: telescopes \Rightarrow telescope
- ▶ synonym: computer \Rightarrow computer, pc

astronomy
computer
pc
science
telescope



Querying

- ▶ Tokenisation
- ▶ Analyses
- ▶ Scoring
- ▶ Filtrering
- ▶ Facetter



Querying – Scoring

- ▶ Boolean Model for Information retrieving



Querying – Scoring

- ▶ Boolean Model for Information retrieving
- ▶ Vector Space Model for Information retrieving
 - ▶ **hvor** i dokumentet har vi fundet et ord?
 - ▶ **hvor mange gange** har vi fundet et ord?
 - ▶ **hvor vigtigt** er dette **dokument**?
 - ▶ **hvor vigtigt** er dette ord i vores **query**?



Querying – Facetter

- ▶ mange dokumenter – et stort resultat

The screenshot shows the Amazon.co.uk search results for the query "dijkstra". The search bar at the top contains "dijkstra" and shows 1-16 of 1,596 results. The results are faceted by department, with a green sidebar on the left listing categories like Books, CDs & Vinyl, and DVD & Blu-ray. The main content area displays two book listings:

- A Discipline of Programming (Automatic Computation)** by E. Dijkstra, published 19 Mar 1976. It is available in paperback for £58.99 (with Prime) and hardcover for £44.79. It has a 5-star rating and 2 reviews.
- Idols of Perversity: Fantasies of Feminine Evil in Fin-de-Siecle Culture (Oxford Pa)** by Bram Dijkstra, published 1988. It is available in paperback for £24.00 (with Prime) and hardcover for £12.79. It has a 5-star rating and 5 reviews.

The sidebar on the left lists various departments under "Show results for":

- Books >**
 - Computing & Internet
 - Reference
 - Computer Information Systems
 - Philosophy
 - Amazon Online Shopping
 - + See more
- CDs & Vinyl >**
 - Choral
 - Chamber Music
 - Religious Mass
 - Classical Song
 - Motets & Anthems
 - Classical Opera
- DVD & Blu-ray >**
 - Musicals & Classical
- + See All 11 Departments



Søgning

- ▶ Indeksering
 - ▶ Parsing
 - ▶ Tokenisation
 - ▶ Analyses
- ▶ Querying
 - ▶ Tokenisation
 - ▶ Analyses
 - ▶ Scoring
 - ▶ Filtrering
 - ▶ Facetter

Table of Contents

Hvem er jeg

Lucene

Søgning

Indeksering

Querying

Solr

Hvorfor Solr?

Demo

Hvad skal jeg bruge?

Links

Solr

- ▶ Apache Solr er en standalone enterprise search server som bruger Lucene
- ▶ er en del af Lucene-projektet
- ▶ kan parse nogle af de mest udbredte filtyper: xml, json, csv, pdf, doc, docx, ppt, pptx, xls, xlsx, odt, odp, ods, ott, otp, ots, rtf, htm, html, txt, log
- ▶ cloud-mode: load-balancing
- ▶ jetty-job: recovery



Hvorfor Solr?

- ▶ man behøver ikke at kunne programmere
- ▶ web-baseret brugerinterface
- ▶ al konfiguration fortages i xml-filer
- ▶ paging!

The screenshot shows the Amazon.co.uk search results for the query "dijkstra". The search bar at the top contains "dijkstra" and a magnifying glass icon. Below the search bar, there are navigation links for "Shop by Department", "Today's Deals", "Gift Cards", "Sell on Amazon", and "Help". The search results are displayed in a grid format. The first result is the book "A Discipline of Programming (Automatic Computation)" by E. Dijkstra, published in 1976. It is available in paperback for £58.99 (with a Prime discount) and has a 2-star rating. The second result is "Idols of Perversity: Fantasies of Feminine Evil in Fin-de-Siecle Culture" by Sram Dijkstra, published in 1989. It is also available in paperback for £24.00 (with a Prime discount) and has a 5-star rating. The page includes a sidebar with department filters and a footer with navigation icons.

amazon.co.uk

Shop by Department

Amazon.co.uk Today's Deals Warehouse Deals Outlet Subscribe & Save Vouchers Amazon Family Amazon Prime Amazon Student Amazon Instant Vi

1-16 of 1,596 results for "dijkstra"

Show results for

Books

- Computing & Internet
- Reference
- Computer Information Systems
- Philosophy
- Amazon Online Shopping
- See more

CDs & Vinyl

- Choral
- Chamber Music
- Religious Mass
- Classical Song
- Motets & Anthems
- Classical Opera

DVD & Blu-ray

- Musicals & Classical

See All 11 Departments

A Discipline of Programming (Automatic Computation) 19 Mar 1976
by E. Dijkstra

Paperback
£58.99
Only 2 left in stock - order soon.
More buying choices
£52.72 used & new (30 offers)

Hardcover
£44.79 used & new (9 offers)

★★★★☆ 2
Trade-in eligible for an Amazon
Eligible for FREE UK Delivery
Books: See all 1,104 items

Idols of Perversity: Fantasies of Feminine Evil in Fin-de-Siecle Culture (Oxford Pa
by Sram Dijkstra

Paperback
£24.00
Get it by **Wednesday, Jul 22**
More buying choices
£12.79 used & new (41 offers)

★★★★☆ 5
Trade-in eligible for an Amazon
Eligible for FREE UK Delivery
Books: See all 1,104 items

Hardware

Hvorfor Solr?

- ▶ schema.xml
 - ▶ indexed?
 - ▶ stored?
 - ▶ dynamicField
 - ▶ copyField

Hvorfor Solr?

- ▶ schema.xml
 - ▶ indexed?
 - ▶ stored?
 - ▶ dynamicField
 - ▶ copyField
 - ▶ ... eller lad den selv finde ud af det

Hvorfor Solr?

- ▶ schema.xml
 - ▶ indexed?
 - ▶ stored?
 - ▶ dynamicField
 - ▶ copyField
 - ▶ ... eller lad den selv finde ud af det
- ▶ solrconfig.xml
 - ▶ tilpas requesthandleren

Hvorfor Solr?

- ▶ Man kan tilføje, opdatere og slette dokumenter ved at sende xml-filer over http
- ▶ Queries sendes som GET eller POST request
- ▶ Svar modtages fx. på XML-format



Demo

Table of Contents

Hvem er jeg

Lucene

Søgning

Indeksering

Querying

Solr

Hvorfor Solr?

Demo

Hvad skal jeg bruge?

Links

Hvad skal jeg bruge?

- ▶ Der eksisterer Lucene-implementationer i andre programmeringssprog
- ▶ Brug Solr, hvis du ikke vil slås med Java
- ▶ “You can't drive an engine, but you can drive a car.”

Table of Contents

Hvem er jeg

Lucene

Søgning

Indeksering

Querying

Solr

Hvorfor Solr?

Demo

Hvad skal jeg bruge?

Links

Links

- ▶ <https://lucene.apache.org/core/>
- ▶ <https://lucene.apache.org/solr/>