Apache Lucene og Solr

Jana Kunert
feymour@gmail.com

The Camp 2015

July 21, 2015

Table of Contents

Table of Contents

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

Lucene

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

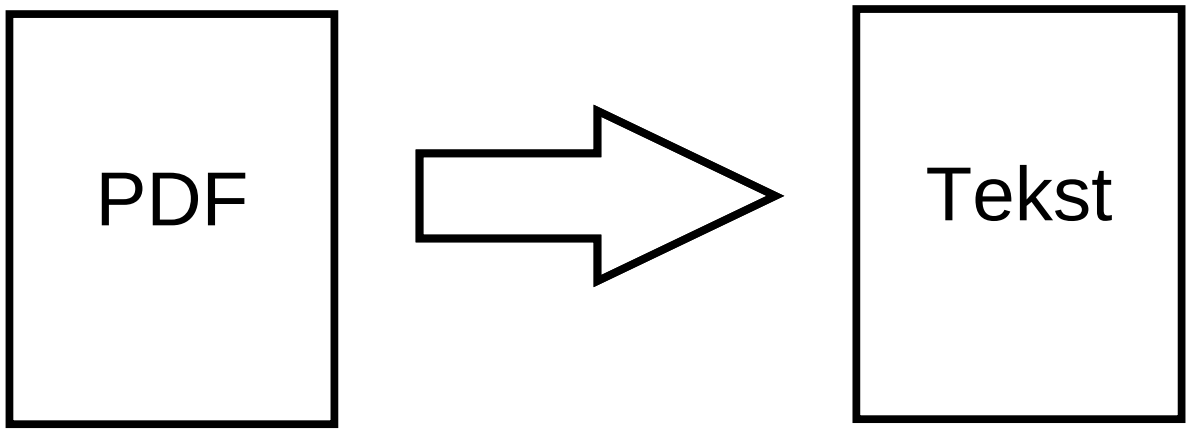
Søgning

- Indeksering
- Querying

Indeksering

- Parsing
- Tokenisation
- Analyses

Indeksering – Parsing



Indeksering – Tokenisation

"Computer science is no more about computers then astronomy is about telescopes."

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

Indeksering – Analysis

"Computer science is no more about computers then astronomy is about telescopes."

- lower-case: Computer ⇒ computer

Indeksering – Analysis

"Computer science is no more about computers then astronomy is about telescopes."

- lower-case: Computer ⇒ 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 ⇒ computer
- stop-words: ~~is~~, ~~no~~, ~~more~~, ~~about~~, ~~then~~
- stemming: telescopes ⇒ telescope

Indeksering – Analysis

"Computer science is no more about computers then astronomy is about telescopes."

- lower-case: Computer ⇒ computer
- stop-words: ~~is~~, ~~no~~, ~~more~~, ~~about~~, ~~then~~
- stemming: telescopes ⇒ telescope
- synonym: computer ⇒ computer, pc

Indeksering – Analysis

"Computer science is no more about computers then astronomy is about telescopes."

- lower-case: Computer ⇒ computer
- stop-words: ~~is~~, ~~no~~, ~~more~~, ~~about~~, ~~then~~
- stemming: telescopes ⇒ telescope
- synonym: computer ⇒ computer, pc

astronomy
computer
pc
science
telescope

Querying

- Tokenisation
- Analyses
- Scoring
- Filtrering
- Facetter

Querying – Scoring

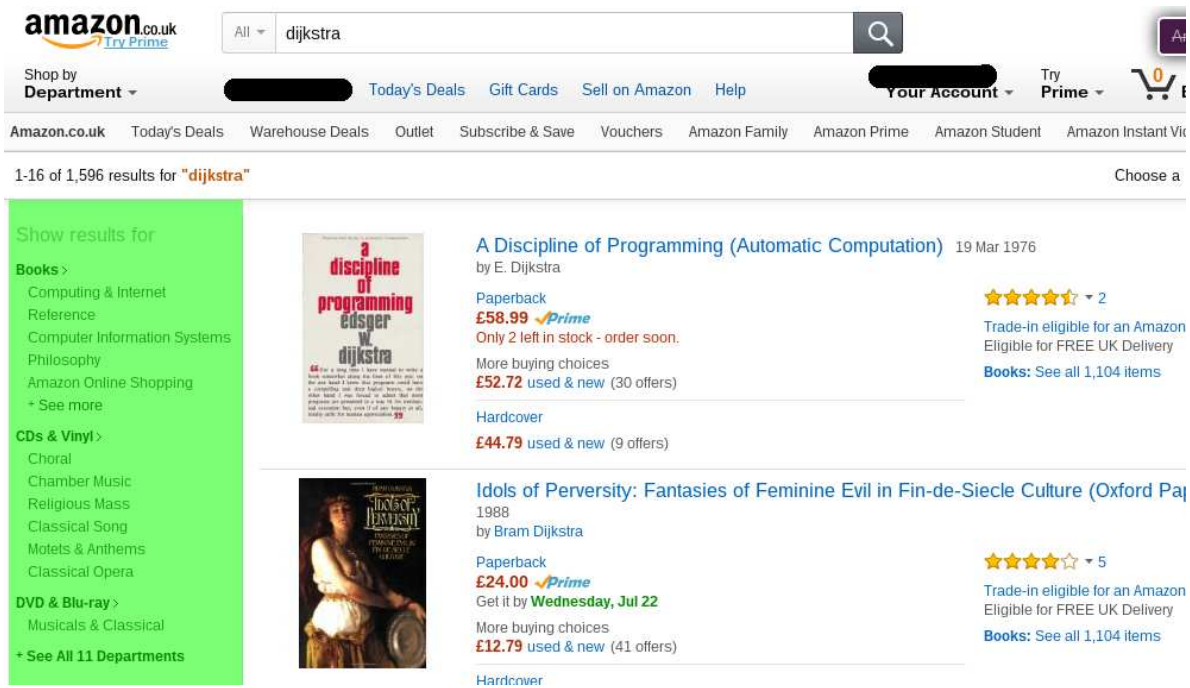
- Boolean Model for Information retrieving

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



Søgning

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

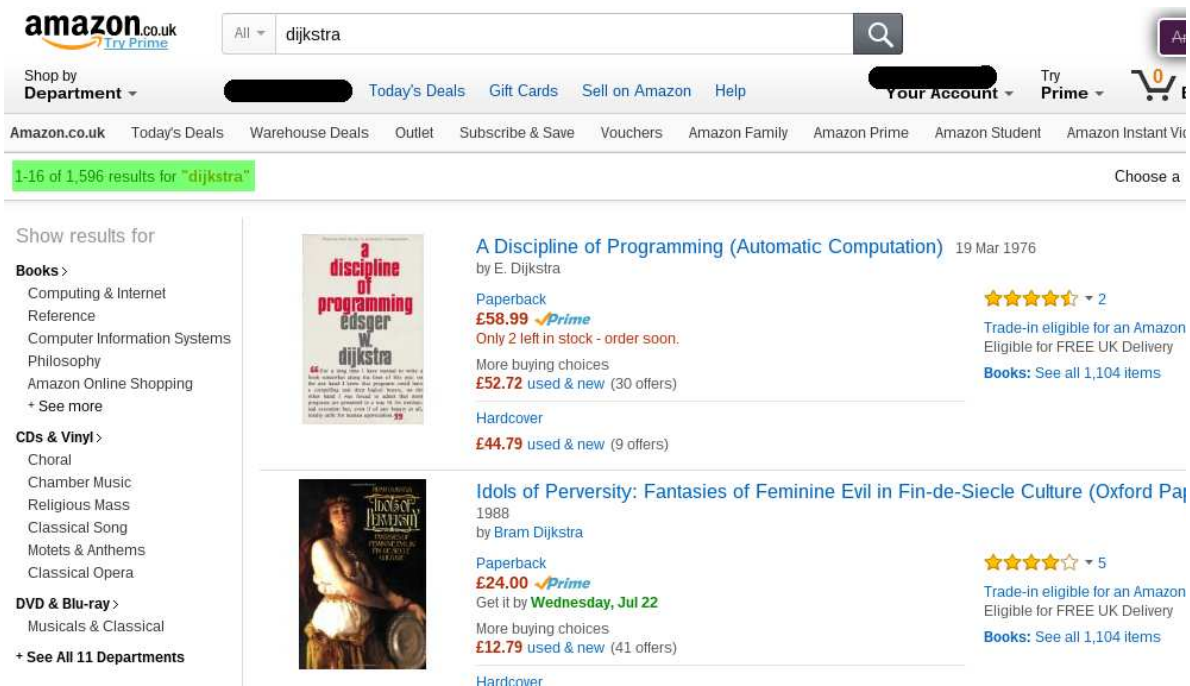
Table of Contents

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!



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

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

Links

- https://lucene.apache.org/core/
- https://lucene.apache.org/solr/