eksponent

LUCENE FOR VINDERE

niels.kuhnel@eksponent.com

eksponent

Lucene™

eksponent

DAGENS PRÆSENTATION

- Lucene i CMS-sammenhæng
- Praktiske anvendelser af Lucene
- Og gotchas
- Russian Mail Order Pets

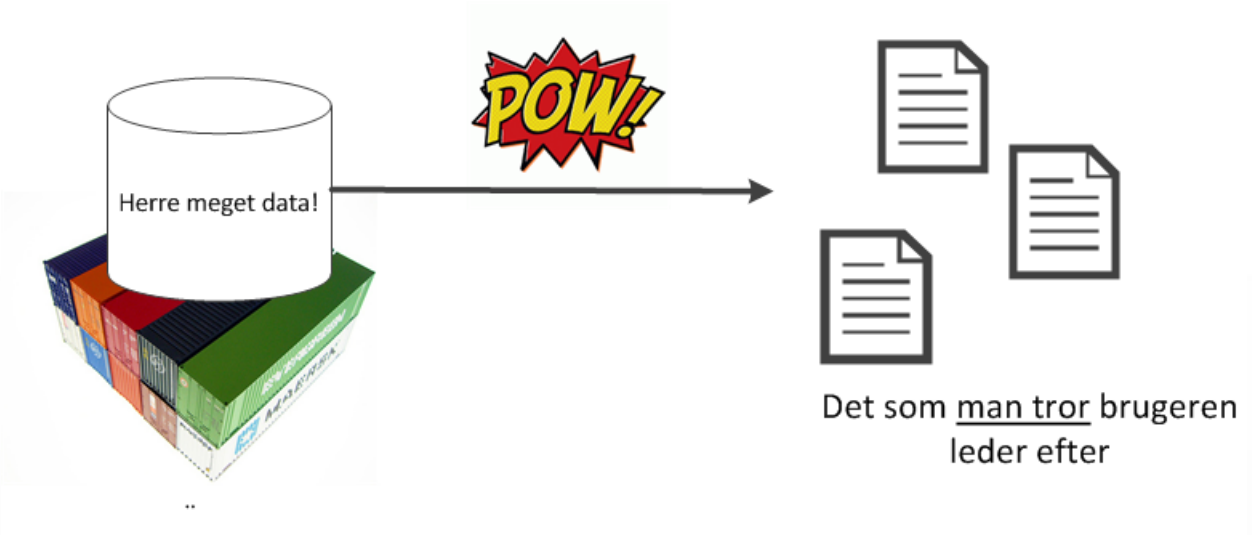
eksponent

DAGENS PRÆSENTATION

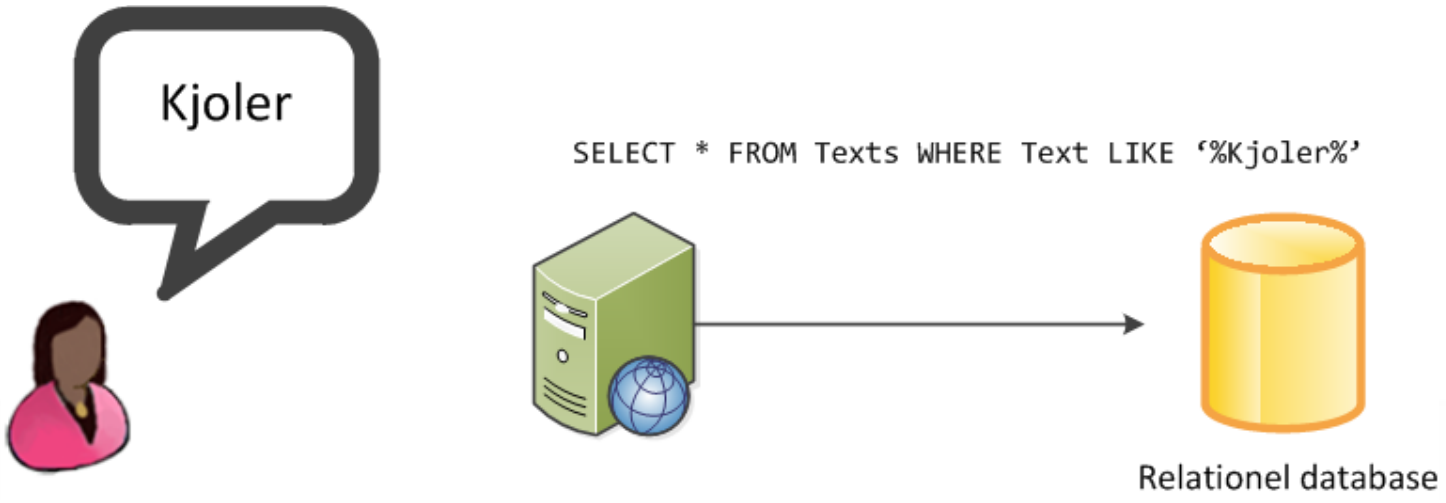
- Lucene er
  - Ikke raketvidenskab
  - Platformagnostisk og sprogneutralt*
    - Indices er binært kompatible
  - Et af de mest vellykkede open source-projekter
  - En MILF
    - SOLR, snaprojects.com og meget mere
  - API baseret på nedarvning af klasser(!)

SØGEFUNKTION, DEFINITION

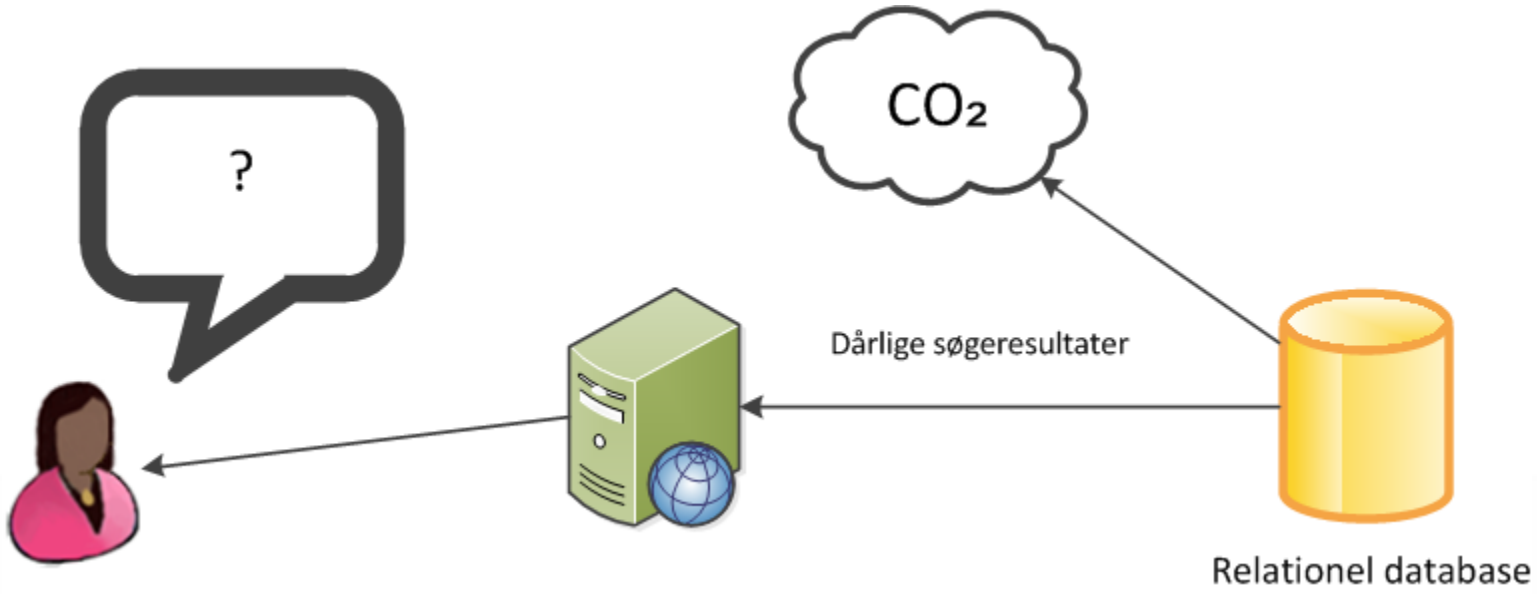
$$f: \mathbb{C} \to \mathbb{C}', \quad \mathbb{C}' = \{c \in \mathbb{C}' | P(c \in \mathbb{U}) \geq \alpha\}, \mathbb{C}' \subseteq \mathbb{C}$$



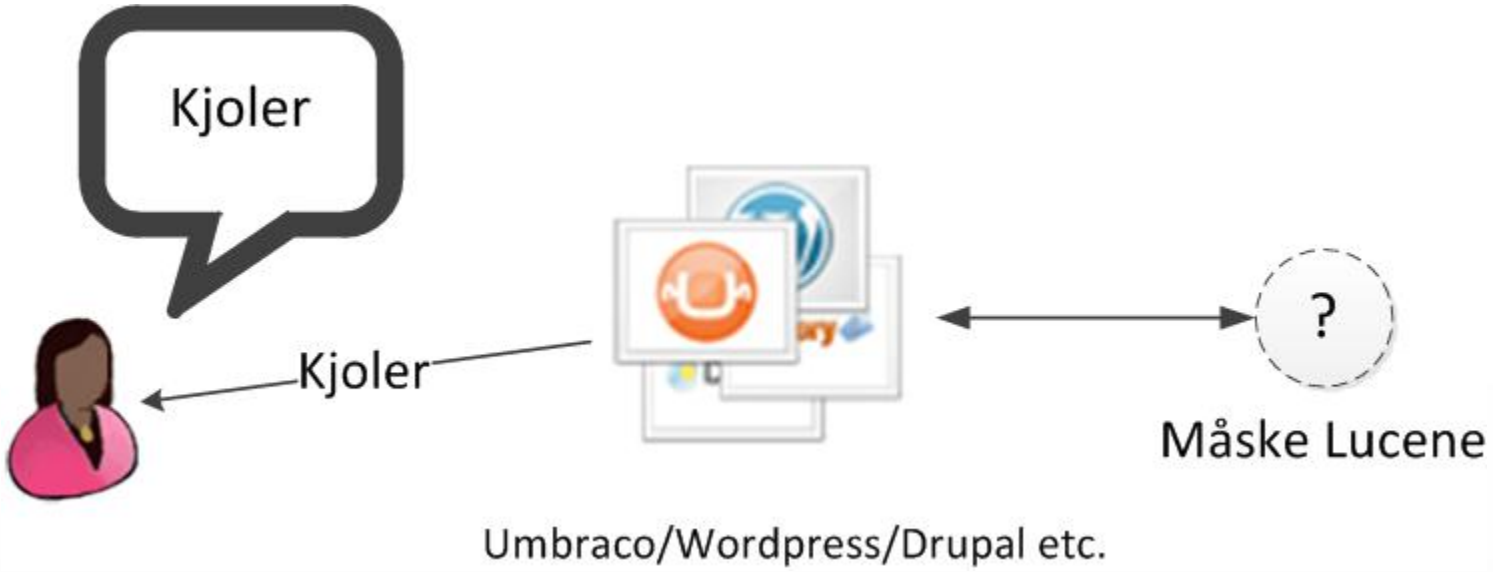
GAMLE DAGE / NEM ”LØSNING”



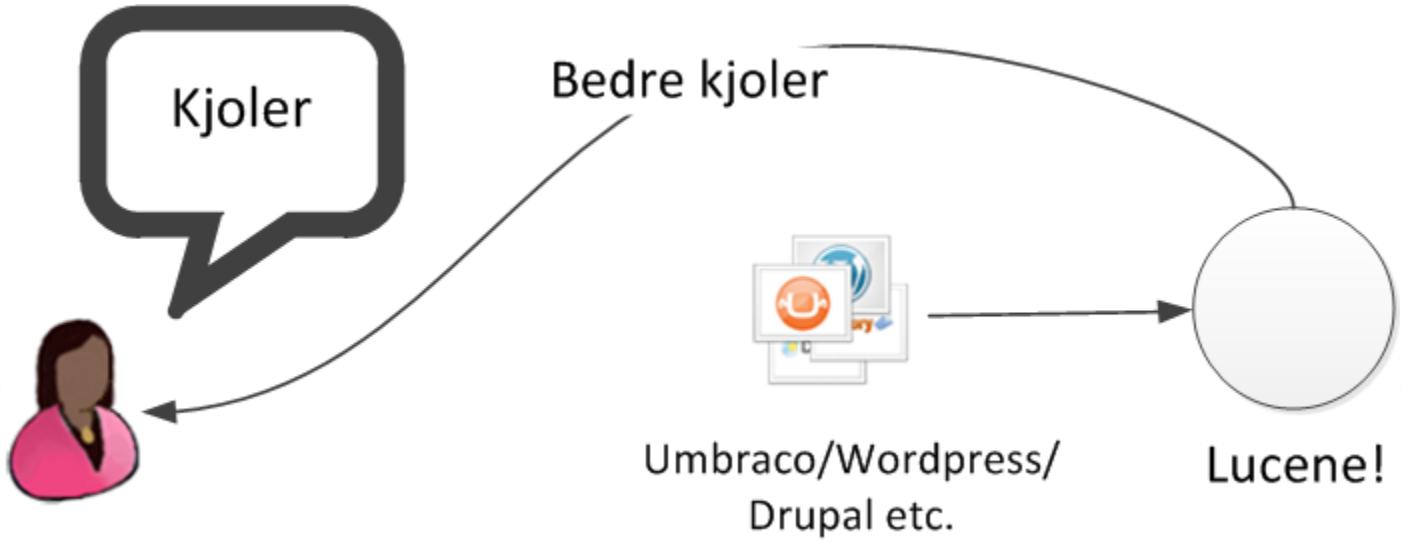
GAMLE DAGE / NEM ~~"LØSNING"~~



TYPISK LUCENE



LUCENE FOR VINDERE



HVORFOR LUCENE?

- Snedigt
- Herre hurtigt
- Passerer ”The CIO”
- Fleksibelt hinsides enhver form for rimelighed

eksponent

LUCENE I EN NØDDESKAL

- Én skriver
- Mange søger
  - Men de bruger alle den samme singleton
- Ingen opdaterer
- Moderat komplekst synkroniseringsscenarie

- Din CMS ved det typisk godt
  - Examine i Umbraco ved det godt

”TERMS/TOKENS”

- Lucene er ligeglad med din tekst
- Lucene bruger terms
- En token er en forekomst af et term i en tekst

”Hej! Jeg er en tekst <3

```
{hej, jeg, er, en, tekst}
```

”TERMS/TOKENS”

- Tokenizers omsætter tekster til terms
- Normalt:
  - Lowercase
  - En token er en sekvens af bogstaver og tal
  - Alt andet er ligemeget
- Særlige tilfælde
  - Autosuggest /N-grams:

```
{t, to, tok, toke, token}
```

INTERMEZZO

- Luke

SØGNINGER

- Lucenes kerne kender kun ”=”
  - Alt andet omsættes via præprocessering
  - ”Os*” = ”Osmose, Osama, Ostehest”
  - ”1 – 100” = ”1, 2, 3, 4, …, 99, 100”
    - Derfor Trie-indeksering med NumericField
  - Flere søgetermer -> langsommere søgning
  - Flere terms -> Større indeks
- VSM
  - Det der gør søgeresultater relevante
  - Google Summer Of Code ’11 (David Nemesky)

SØGNINGER

- Glem QueryParser og brug API
  - Ingen escaping errors og bedre kontrol
- BooleanQuery
  - Should = OR, Must = AND
- Boost
- Collectors
  - TopScoreDocCollector (relevans)
  - TopFieldCollector (sortering)

SCORING

- ”Kost”
  - ”Kost” (godt)
  - ”Koste” (rimelig godt)
  - ”Kosteskab” (ok)
- ”Intrasandt”
  - ”Intrasandt” (dårligt)
  - ”Interessant” (godt)
  - ”Interessante” (rimelig godt)
  - ”Spændende” (rimelig godt)

SCORING

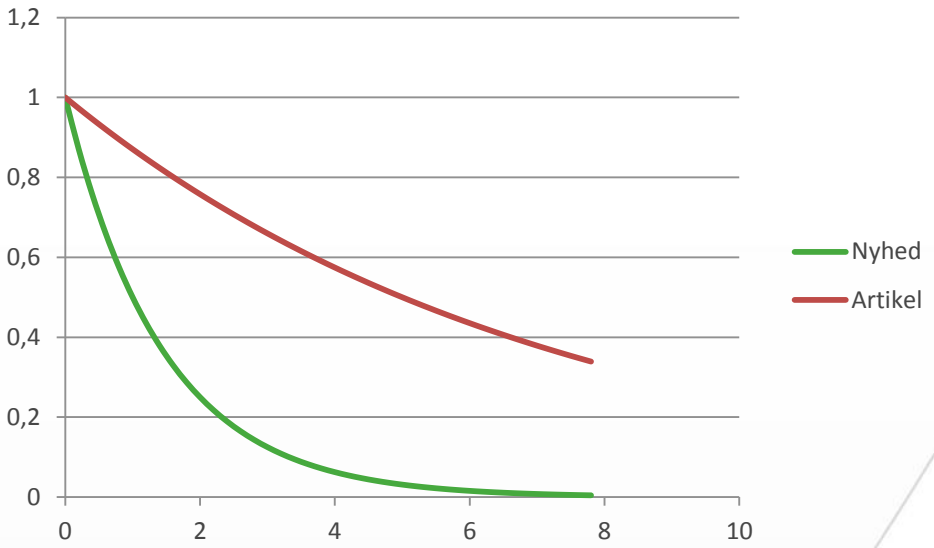
- Giv den gas!
- Et ord er vigtigere i titlen end i brødteksten
- Brug synonymer
- Brug tags
- Ret stavefjel
- En ny nyhed er bedre end en gammel…

SCORING

- ”Halveringstid”
  - En gammel artikel kan være bedre end en halvgammel nyhed
  - CustomScoreQuery

$$score' = score \cdot e^{-alder \cdot \lambda}$$

$$\lambda = \frac{ln2}{halveringstid}$$



DOCUMENT STORAGE?

- Lucene må aldrig bruges som primært lager
- Lucene er et godt lager
  - Spar udviklingstid på caching/performance-optimering
- Brug repository'ish pattern
  - POCO til Document
    - NumericField hvis du vil range eller sortere
  - Hydration (Document til POCO)

FACETTER

- Facetter hjælper brugeren med at afgrænse
- Custom Collector
- OpenBitSet

- Det er nemmest at cache
  - Sådan gør SOLR
  - Beregn facetter, når en ny Reader åbnes

OPENBITSET

- long[] wrapper
- Hver dokument er en bit
- 100.000 dokumenter fylder 12,2kB pr. facet
- Forenings- og fællesmængder
- Kardinalitet (antal 1-taller)

HIGHLIGHTING

- Field.TermVector.WITH_POSITIONS_OFFSETS!
- Highlighter
  - Kedelig ”out of the box”
  - Ganske tilpasselig
- FastVectorHighlighter
  - Meget hurtigere
  - Kræver c0re-hax for at levere rimelige resultater
  - Don’t go there med mindre du har store PDF’er

GOTCHAS

- Collation
  - Brug min fremragende port af ICUCollationKeyFilter
- String.StartWith gør mærkeligt på dansk
  - ”Østergaard” starter ikke med ”Østerga”
  - Thread.Current.CultureInfo = CultureInfo.InvariantCulture
- Highlighting virker ikke godt ”out of the box”

GOTCHAS

- Lav et felt All=”1” hvis du vil trække alt ud
- Søgninger på noget der ikke findes giver alt
- Hvis man ikke kan opgradere Lucene.NET kan man bruge en IKVM.NET-port fra Java

eksponent

MOR ELLER DATTER?

- En ren Lucene baseret løsning:
  - Kan deployes i shared hosting
  - Giver fuld kontrol
  - Kræver en indsats
- SOLR/Compass/Elasticsearch:
  - Er nemmere at gå til
  - Giver ikke fuld kontrol. Dog typisk nok.
  - Kræver en Java-server
  - Giver et mere komplekst deploymentscenarie

eksponent

HØR MERE!