

Søketechnology- webSearch

2022-11-09

(Siste ordinære gruppetime)

Websearch

Internett (perspektiv: crawling)

Søkesystemet ved UiO

Gjennomå oblig D

Oblig E-hjelp



Neste uke

- Science fair 10:15-14:00
- På Python
- Følgelig ikke noe gruppetime

Internett

- Nettverk med servere
- HTTP(S)
- Hypertext transfer protocol
- Stor graf

Crawling i praksis

- Du vil ikke crawle.
- /robots.txt
- .htaccess

Robots.txt

- En tekstfil
- "Hvis dette er deg får du ikke gå hit"
- E.g. <https://www.nrk.no/robots.txt>
- Skal holde bots ute. Honour-system.

← → ↻ Secure <https://www.buzzfeed.com/robots.txt>

```

User-agent: msnbot
Crawl-delay: 120
Disallow: /*.xml$
Disallow: /buzz/*.xml$
Disallow: /category/*.xml$
Disallow: /mobile/
Disallow: *?s=mobile
Disallow: *?s=lightbox
Disallow: /bfmp/
Disallow: /buzzfeed/
Disallow: /contest
Disallow: /contests
Disallow: /plugin/
Disallow: /embed/
Disallow: /_comments/

User-agent: *
Disallow: /buzz/*.xml$
Disallow: /category/*.xml$
Disallow: /mobile/
Disallow: *?s=lightbox
Disallow: /bfmp/
Disallow: /buzzfeed/
Disallow: /contest
Disallow: /contests
Disallow: /_ga/
Disallow: /static/
Disallow: /dashboard/
Disallow: /plugin/
Disallow: /api/
Disallow: /buzzfeed/api/
Disallow: /embed/
Disallow: /_comments/

User-agent: discobot
Disallow: /

User-agent: Slurp
Crawl-delay: 4
        
```

Buzzfeed.com wants msnbot to wait 120 msc before crawling each page and NOT crawl any of these URL strings.

AND

Buzzfeed.com wants all other user-agents (except for msnbot, discobot, and Slurp) to NOT crawl any of these URL strings

AND

Discobot should not crawl ANY URLs on buzzfeed.com.

AND

Slurp (Yahoo's user-agent) should wait 4 msc before crawling each page, but crawl all URLs on buzzfeed.com.

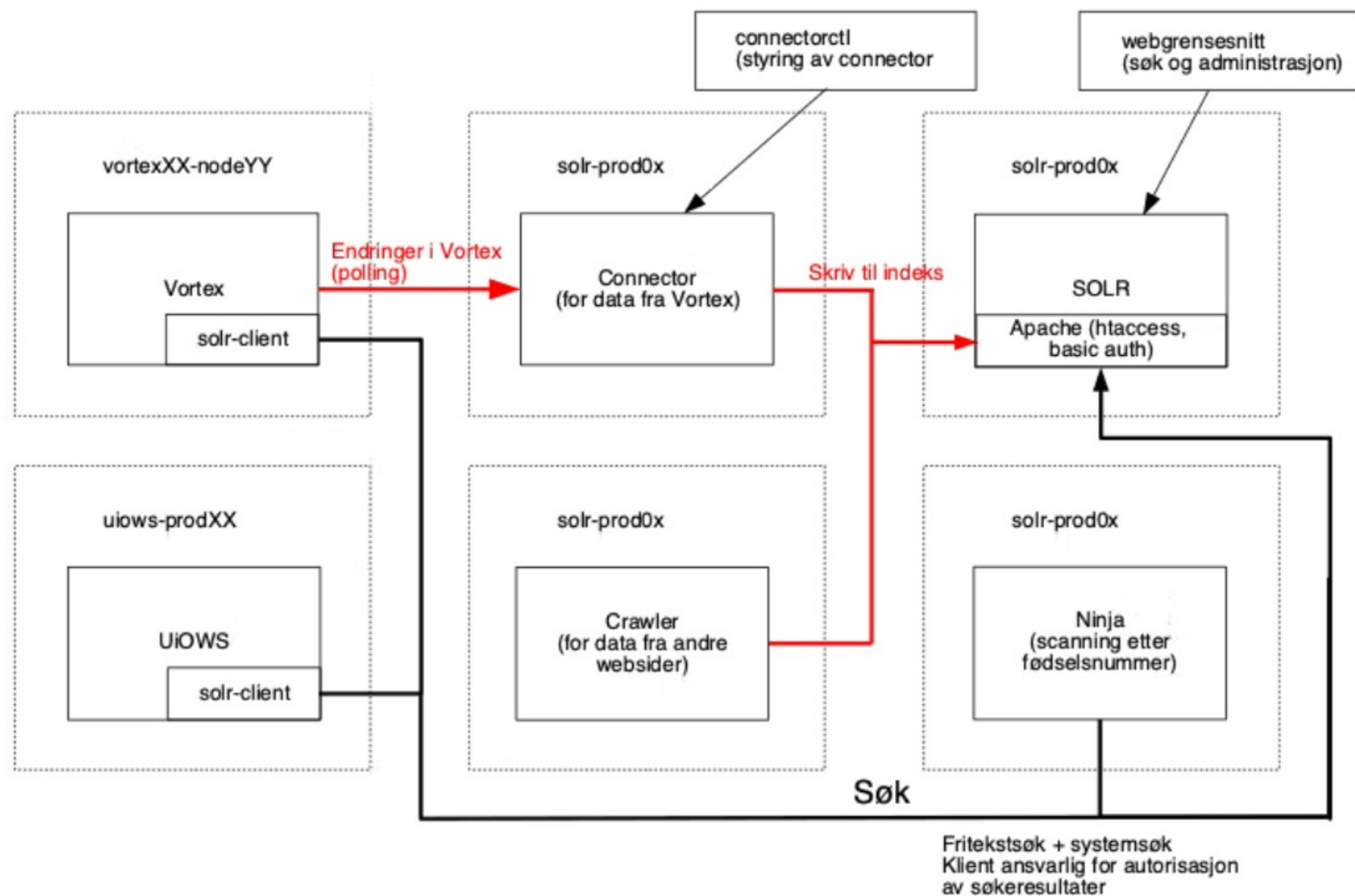
Eksempel på en robots, forklart.

Søk ved uio.no

- Docs: <https://www-int.uio.no/tjenester/it/web/sok/drift-utvikling/rutiner-og-dokumentasjon/rutiner.md#mcf>
- <https://www-int.uio.no/tjenester/it/web/sok/drift-utvikling/rutiner-og-dokumentasjon/solr-teknisk-oversikt.html>
- Robots: <https://www.uio.no/robots.txt>

Vortex

- UiOs egne CMS
- (CMS = Content management system)
- Utvikles av GPL ved USIT
- (Samme folk som dealer med søkesystemet)
- Brukes også av f.x. INN.no



Søk ved uio, akritektur



Søk v/UiO forts

- Solr, AND-queries
- <https://www.uio.no/tjenester/it/web/sok/hjelp/index.html>
- "UiOs søkemotor benytter "AND-søk" som standard, som betyr at alle søkeord må matche for å gi treff."
- <https://www.uio.no/tjenester/it/web/sok/hjelp/syntaks/>

Apache Solr

- FOSS søkesystem
- Skrevet i Java, 2004-2022
- <https://solr.apache.org/>
- https://en.wikipedia.org/wiki/Apache_Solr

Arbeid med oblig E

- Frist på fredag
- Naïve Bayes
- Siste oblig :)

