IN[34]120 Søketeknologi - Introduction

2023-09-01 10:15 @ Chill

Gruppelærer: Oliver (Ruste Jahren), oliverrj@ifi.uio.no

(Join denne mentien!!)

- Praktisk/administrativ info
- Inverted indeces
- Posting lists
- Python refresher

Hvis tid: strenger/oblighjelp/pip-ting

Grl: Oliver Ruste Jahren

- → 5. år på ifi
- → MAPS
- → Grl i søketek i fjor òg 😎
- → Bsc språktek, msc prosa

Søketeknologiens kjerne

- → Språktek
- → Prosa
- → (Robotikk?)

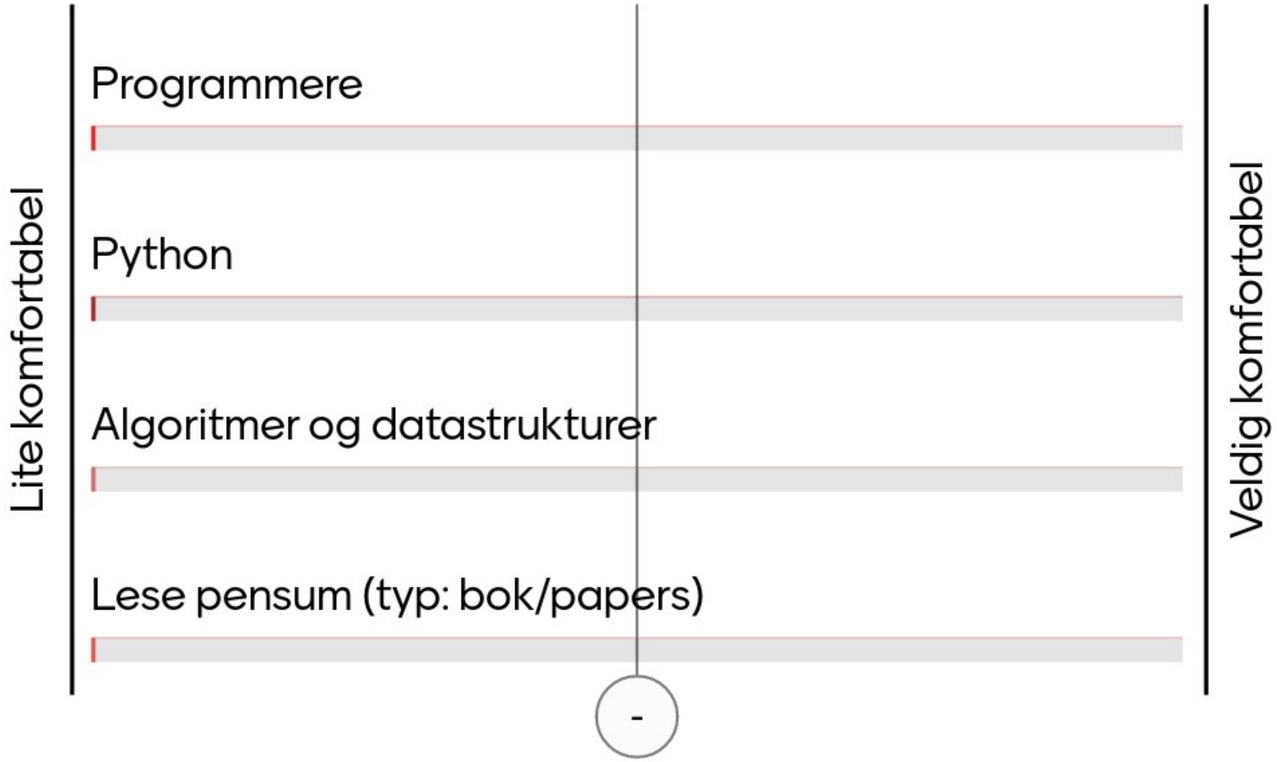
Emnets tematikk

- Data
- Algoritmer
- Datastrukturer
- Maskinlæring / classification
- (Og mer)

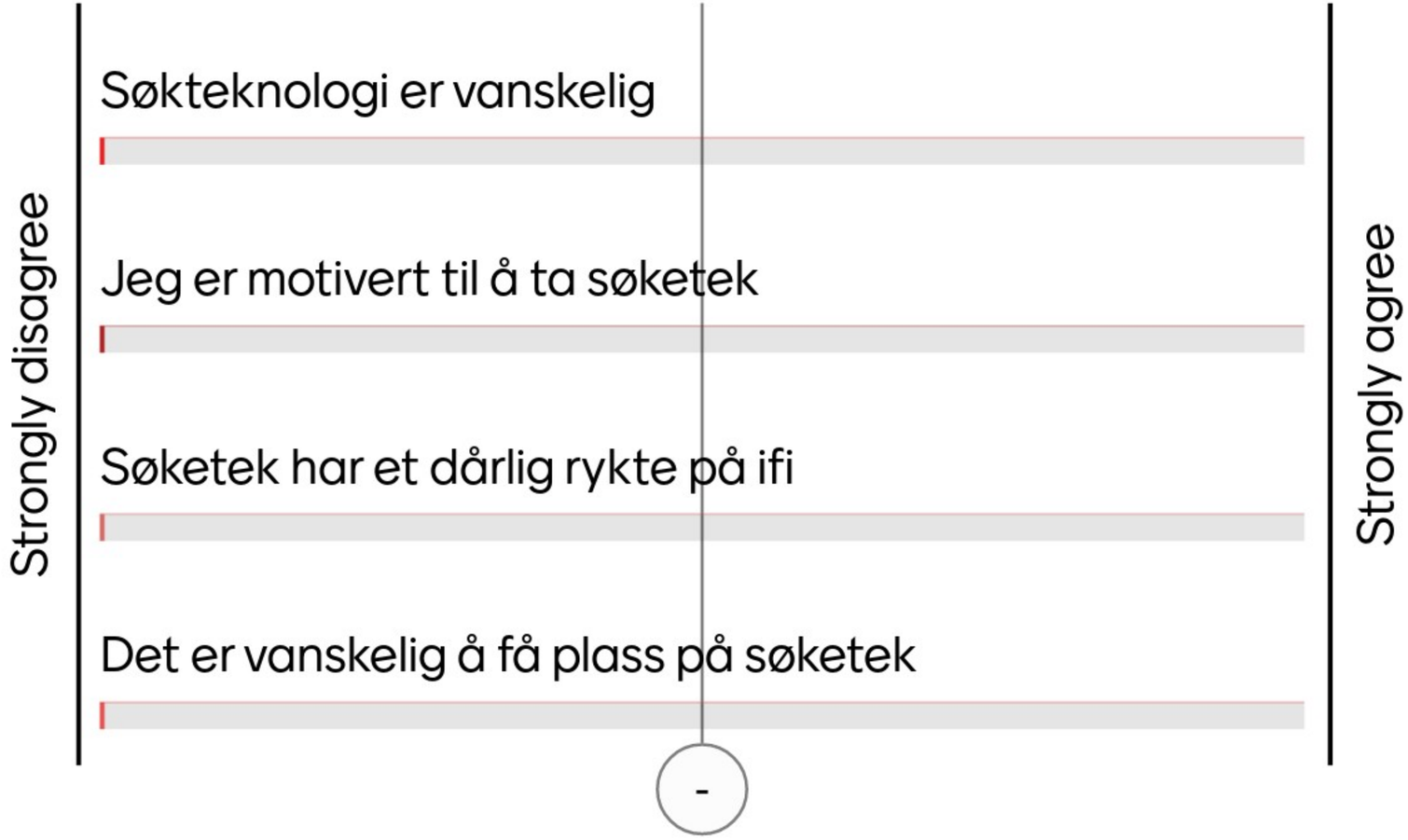
Hva slags forventninger har du til søketek?

There's no correct answer!

Hvor komfortabel er du med hver ting? (anonymt)



Hvor enig er du i hvert utsagn?



Assorterte stykker praktisk/administrativ info

(Viktig å vite)

Github

- Emnets kjerne
- Pensum
- Obliger
- Gruppetime-materiale etc. Alt annet enn opptak
- https://github.com/aohrn/in3120-2023

Mattermost

- (Open source Slack)
- Team
- Kanaler
- ((*Husk å vise hvordan man joiner kanaler*))
- Info

Gruppe 1-mattermost

- Vår egen kanal 😎
- Bli med
- Uklart formål (det blir bra)
- https://mattermost.uio.no/ifi-in3120/channels/group-1

</praktisk info>

(</x> betyr at x er ferdig, det er en SGML-greie)

Lecture recap

Altså fra forrige onsdag, 2023-08-24

Husker noen noe??

Ting som ble husket fra forelesningen.
Gjerne stikkord:

There's no correct answer!

Begreper

- Document
- Term
- Posting
- Query
- Boolean
- Retrieval
- "Boolean retrieval"

Inverted index

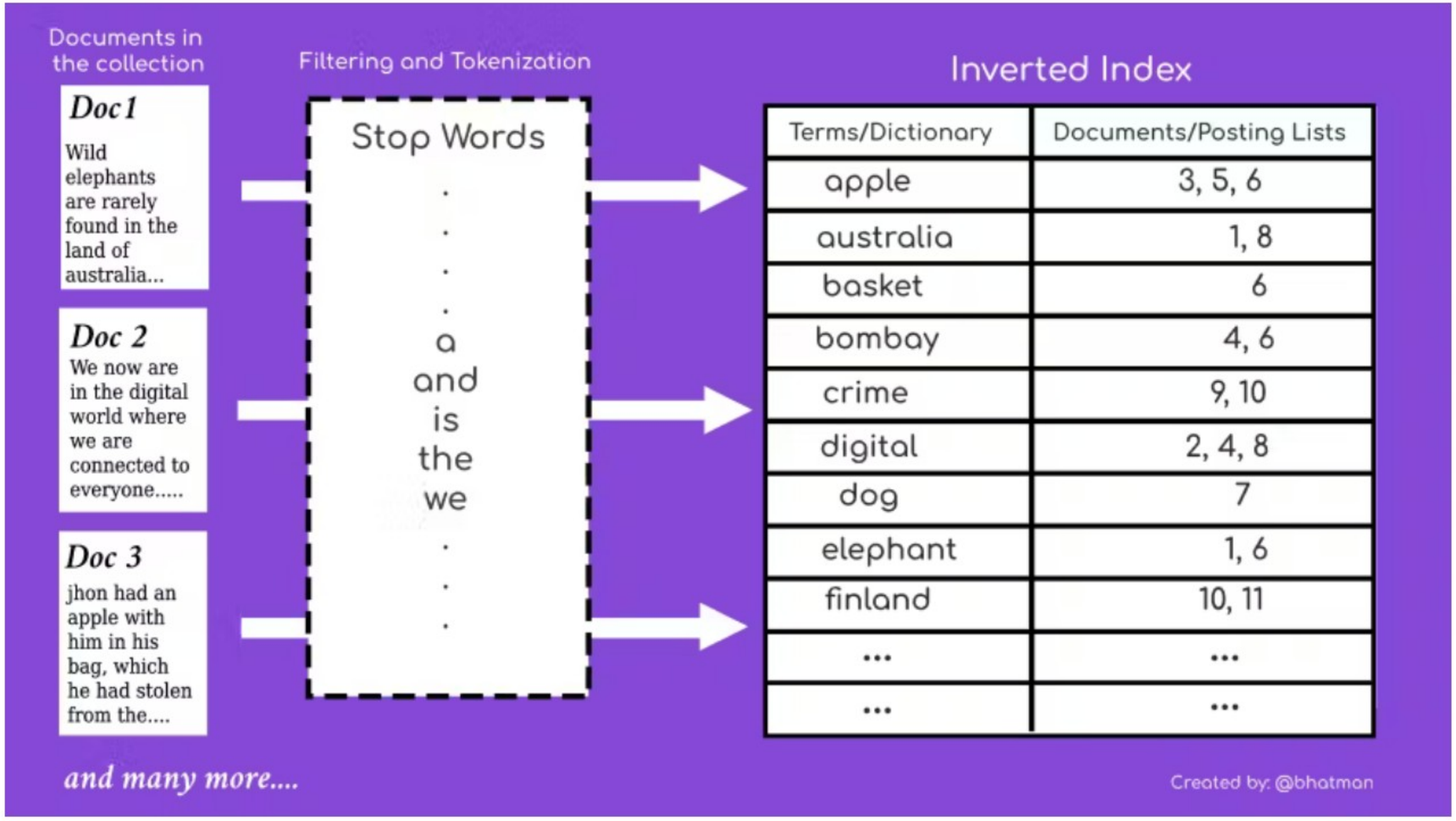
- → Mapping: term -> posting list
- → Som registeret i ei bok
- → 1/2 Oblig A (2023-09-15 (2 uker til))

Posting list

- En mengde dokumenter
- Alle dokumentene inneholder minst ett ord som er likt
- "her er alle dokumentene med 'Zeus'"

Posting lists forts.

- → Effektivitet: OOP?
- → Optimalisering: tall
- → 1 - 4 - 6 - 9
- → NB: Må være sortert



Visualisering: inverted index m/posting lists

NB: Stoppord

- → "a", "the", "her"
- → Betyr ikke noe
- → Mange av dem -> dyrt å behandle
- → Ignorer!

Primitivt søk

- → Boolsk relevan -> ingen ranking
- → Ingen tolerans (må ha 100% lik stavemåte)

Hvorfor må posting lists være sortert?

The correct answer is: Slik at man kan gjøre effektive operasjoner på dem

Operasjoner på posting lists

- → Union (det som er i begge)
- → Intersection (det som kun er i den ene)
- → 2/2 Oblig A

Oblig A

- Inverted index
- Postings-merger: union
- Postings-merger: intersection
- PM: Konstant minne
- Generators
- (2023-09-15)

Generators in Python

- Alternative to returning
- Se https://www.uio.no/studier/emner/matnat/ifi/IN3120/h20/material-for-group-sessions/advanced_python.pdf

Maps generalforsamling ("genfors")

- → 'Matematikk, Algoritmer og Programmering for Studenter'
- → Beste studentforening
- → Genfors i dag kl 16:15(?) @ C (i 3. etg)
- → Kom og bli med!!

Spørsmål? Mattermost, mail, brevdue: oliverrj@ifi.uio.no