IN[34]120 Søketeknologi - Introduction

2024-08-27 14:15 @ Chill

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

Join denne mentien!!

Agenda:
- Praktisk info
- Inverted indeces
- Posting lists
- Avansert Python

Hvis tid: oblighjelp/pip-ting

Grl: Oliver Ruste Jahren

- 6. år på ifi (wow, ikke flex)
- Sitter i Fagutvalget (FUI)
- Grl i søketek i fjor òg 😎 (og før det 😎)
- BSC språktek, MSC prosa

Søketeknologiens kjerne

- Språktek.
- Prosa.
- (Emnet holdes av MLS/LTG (språktek))

Emnets tematikk

- Data science
- Algoritmer
- Datastrukturer
- Maskinlæring / classification
- Komprimering
- (Og (mye) mer)

Hva slags forventninger har du til søketek?

Lære fancy strengbehandling

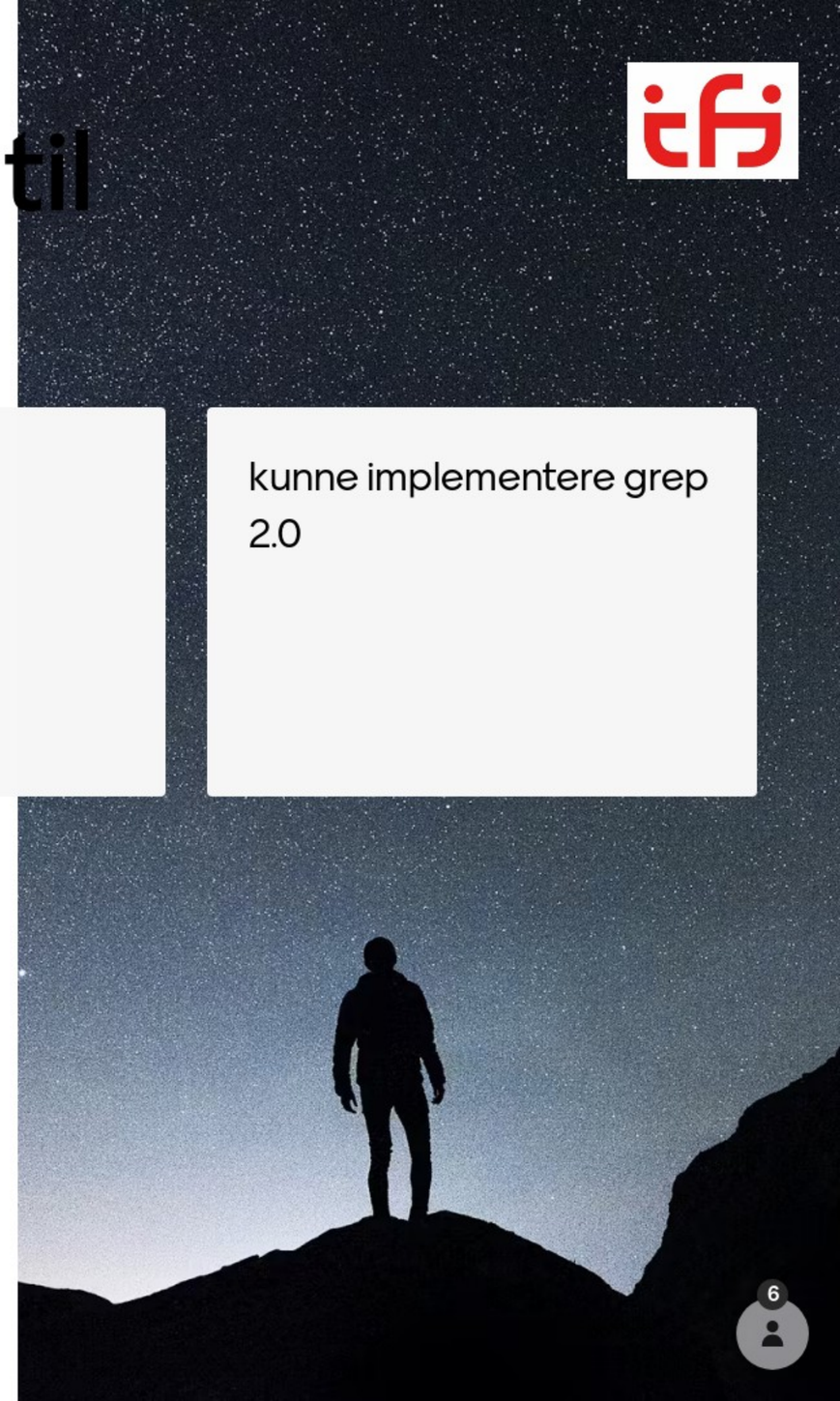
Kaos

søke litt tek

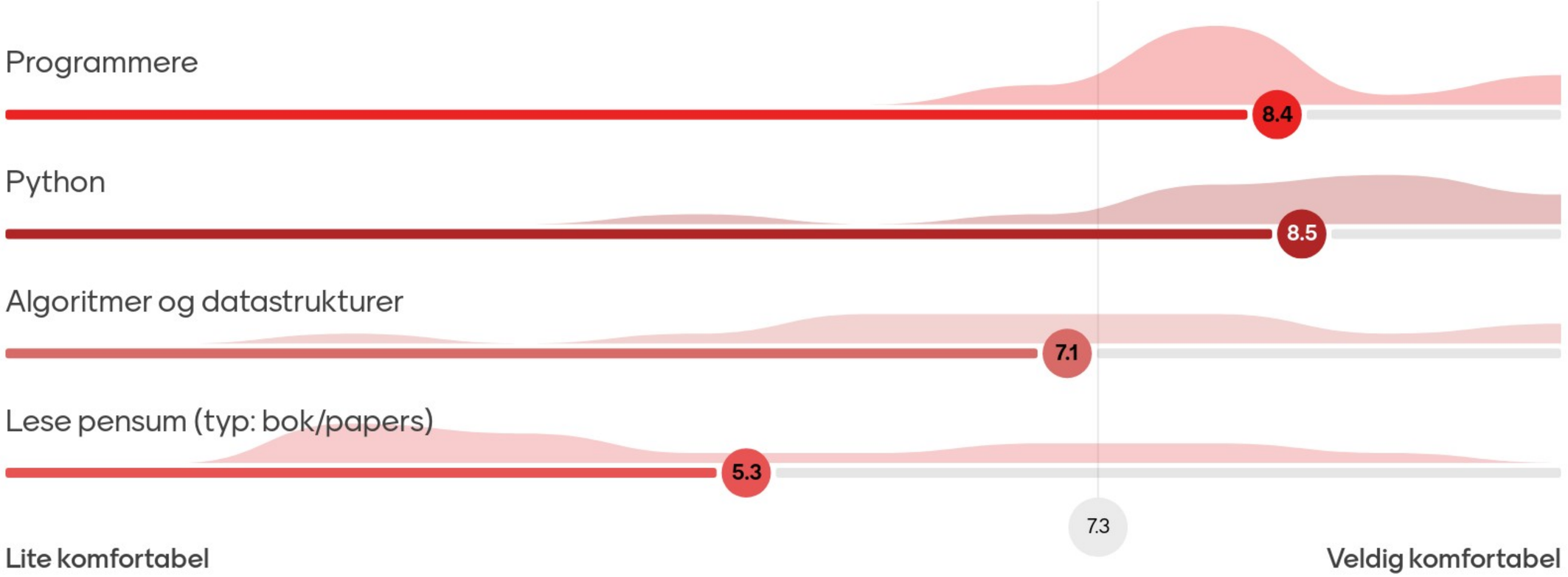
kunne implementere grep 2.0

Teksøk

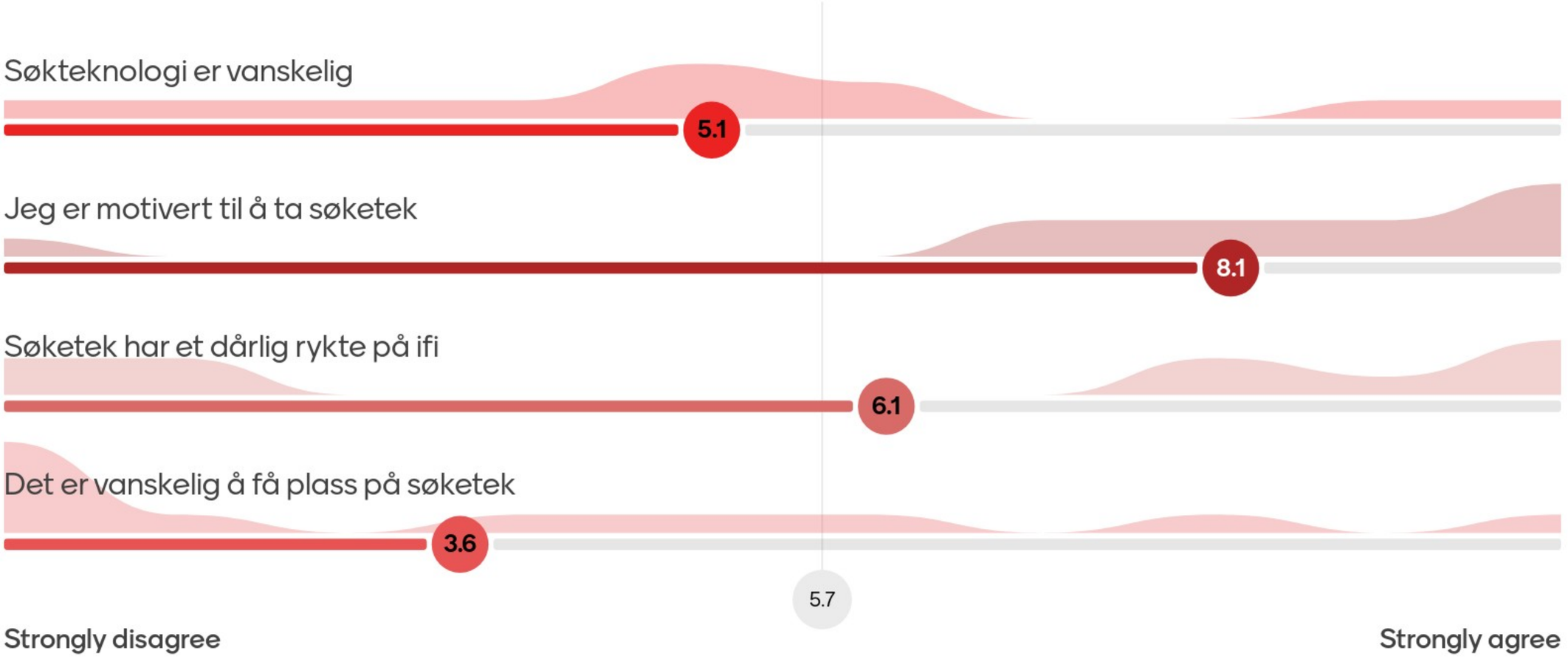
Søking



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 (utover boka/forelesningene)
- Obliger
- Gruppetime-materiale etc. Alt annet enn opptak
- https://github.com/aohrn/in3120-2024

Mattermost

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

Gruppe 2-mattermost

- Vår egen kanal 😎
- Bli med
- Uklart formål (det blir bra)
- https://mattermost.uio.no/ifi-in3120/channels/group-2

</X> BETYR AT X ER FERDIG, DET ER EN SGML-GREIE

</praktisk info>

Lecture recap

De 2 første, 2024-08-19 + 2024-08-26

Husker noen noe??

Ting som ble husket fra forelesningen. Gjerne stikkord:

9 responses

Begreper

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

Inverted index

- Mapping: term -> posting list
- Som registeret i ei bok
- 1/2 Oblig A

Posting list

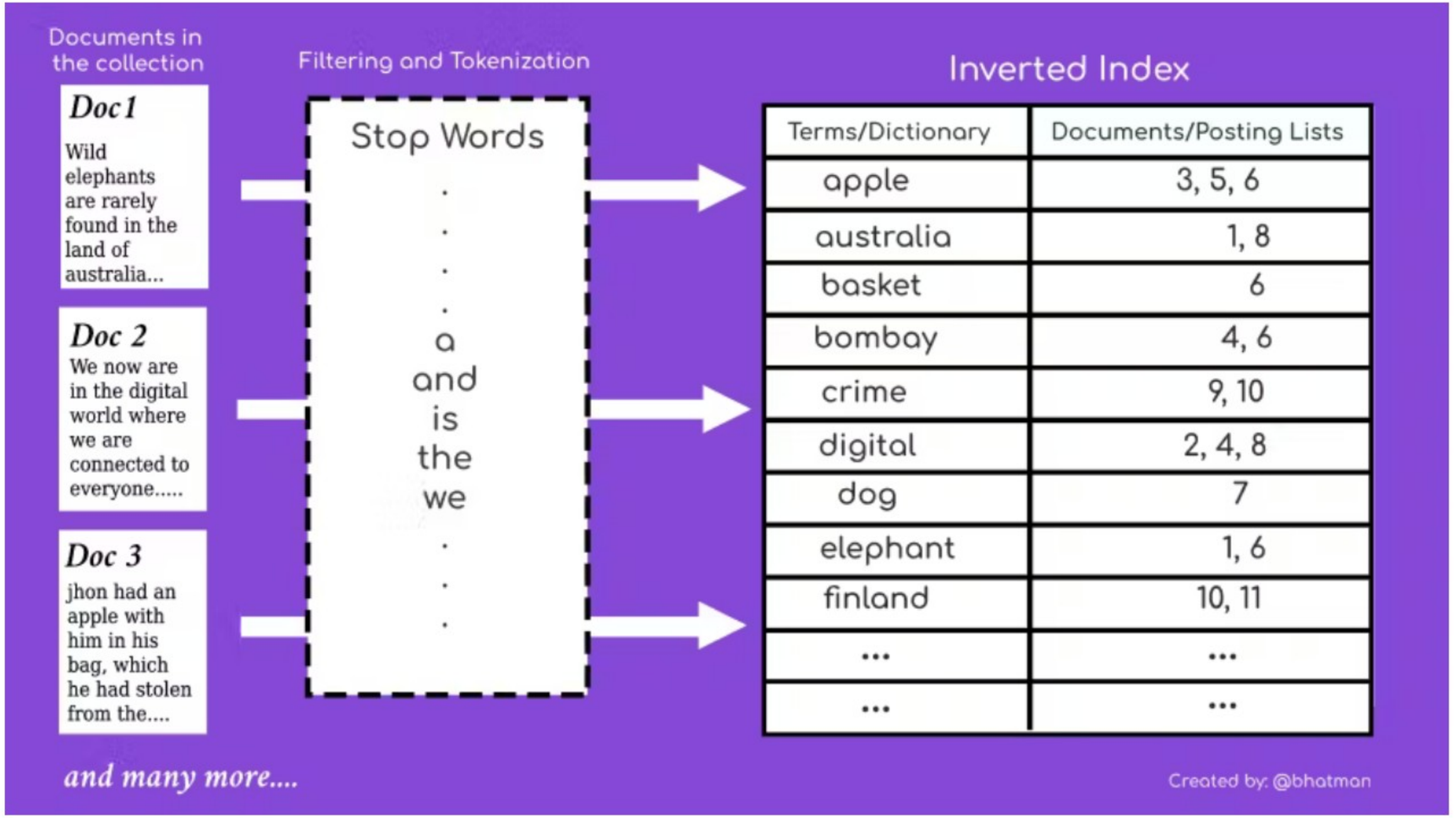
- En mengde dokumenter
- Alle dokumentene inneholder minst én gitt term
- "her er alle dokumentene med 'Edge'"

Posting lists forts.

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

NB: Stoppord

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



Visualisering: inverted index m/posting lists

Primitivt søk

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

Hvorfor må posting lists være sortert?

for effektiv merging ✗	for å finne riktig term/doc ✗	for merging ✗
ENklere å merge ✗	For effektiv sammenligning av to lister for å finne hvilke dokumenter query finnes ✗	For å kunne bruke dem effektivt ✗
må kunne inkrementere minste ✗	binary search ✗	merge ✗

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

Operasjoner på posting lists

- Union (AND, det som er i begge)
- Intersection (OR, det som er i en av listene)
- Nytt i år: Difference (ANDNOT, det som er i x og **ikke** i y)
- 2/2 Oblig A

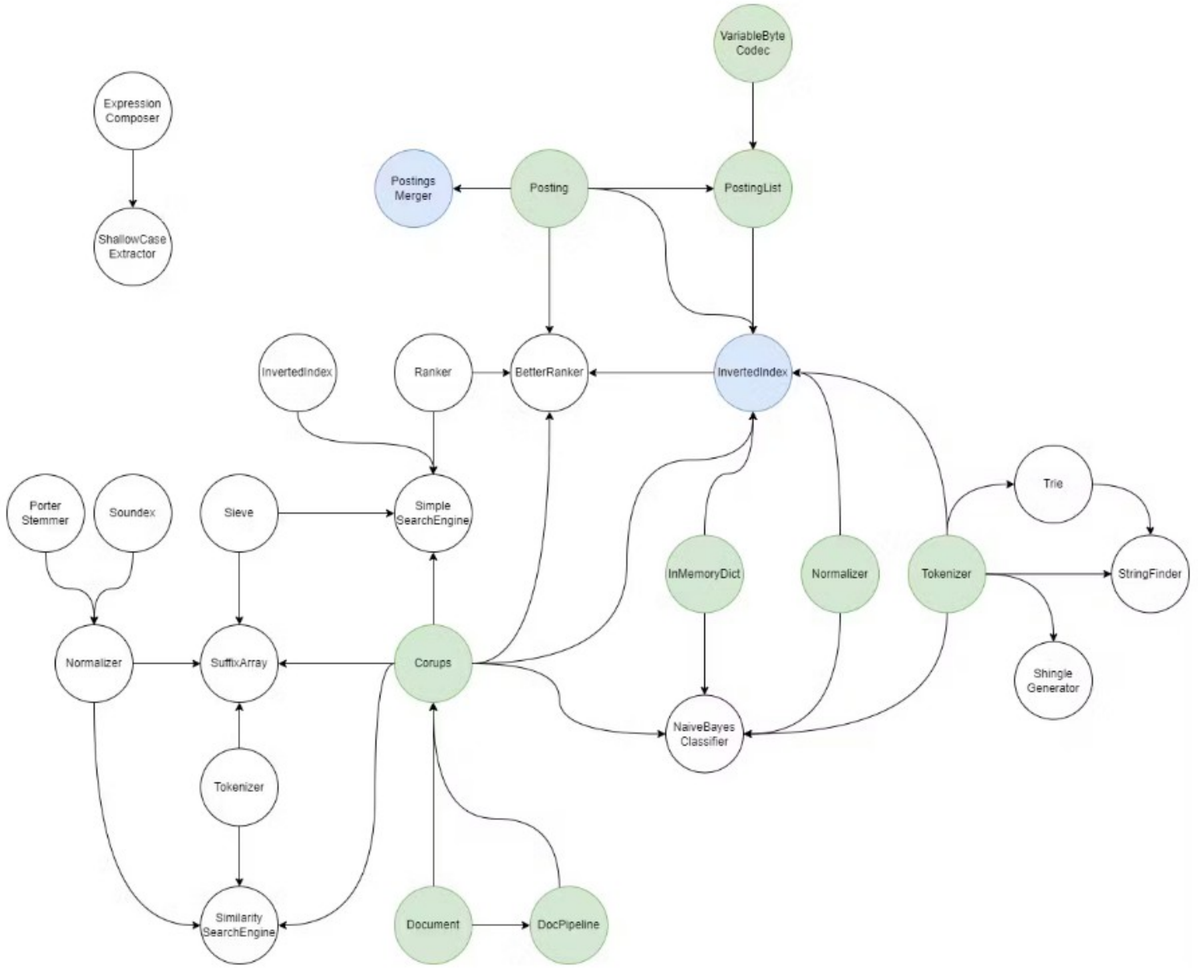
KAN LØSES ALLEREDE NÅ

Oblig A

- Frist: 2024-09-13
- Inverted index
- Postings-merger: union
- Postings-merger: intersection
- Postings-merger: difference
- PM: Konstant minnebruk!
- Generators.

Prekoden

- Det er *omfattende* prekode
- Må bli komfortabel med den
- Finnes noen ressurser fra ifjor, mulig de kommer ut på nytt



Klassene som var relevante for oblig A i 2023

OBLIGTIPS

Generatorer i Python

- Prekoden er litt sær - liker ikke return
- return < yield
- Se https://www.uio.no/studier/emner/matnat/ifi/IN3120/h20/material-for-group-sessions/advanced_python.pdf

Resten av tiden: Klon repoet, se på prekoden, kom igang med oblig A

Spørsmål utenfor gruppetimen? Mattermost: @oliverrj

Pause til kvart over

15:00-15:15

Still gjerne spørsmål i pausen