IN[34]120 Søketeknologi - Datastrukturer for strenger 1

2024-09-03 14:15 @ Chill

Gruppelærer: Oliver, oliverrj@ifi.uio.no

Agenda:

- Prefix search
- Suffix arrays
- Tries
- Oblighjelp

Lecture recap

Altså fra i går, 2024-09-02

Husker noen noe??

Ting som ble husket fra forelesningen. Gjerne stikkord:

Jeg var ikke der rip

jeg så den ikke

meg heller

Mye 3130 pensum

ikke suffixarrays, var ikke der

Tries, edit distance, Ukkonen. Aho-Corasick

Indeksering

jeg tweaker. kodebasen får meg til å tweake.

Språktek-begreper 101

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

Hva ER søk?

- Information retrieval (IR)
- Avgjøre om noe finnes
- Gjenfinne info
- Finne closest match

DET LIGGER LITT I NAVNET

Skillet mellom strukturerte og ustrukturerte data

Strukturerte data

- Definert format
- Ofte et eksisterende formål
- JSON, XML, UML(?)
- Trivielt anvendbar

Ustrukturerte data

- "Vi må lage vår egen struktur"
- Rå tekst
- Data fra ikke-foremålstjenelige kilder
- Må behandles for å kunne brukes til noe
- Parsing

Parsing m/venner

- → Tokenisering
- → Stemming / Lemmatisering
- → Stoppord

Tokenisering

- "token" = "ord", for det meste
- Basic: splitte på mellomrom
- Fancy: "United Kingdom" er 1 token

Lemmatisering

- Samle forskjellige former av ord til stamme
- "bok", "bøker" og "boka" blir alle "bok"
- Bevare semantikk

Stoppord

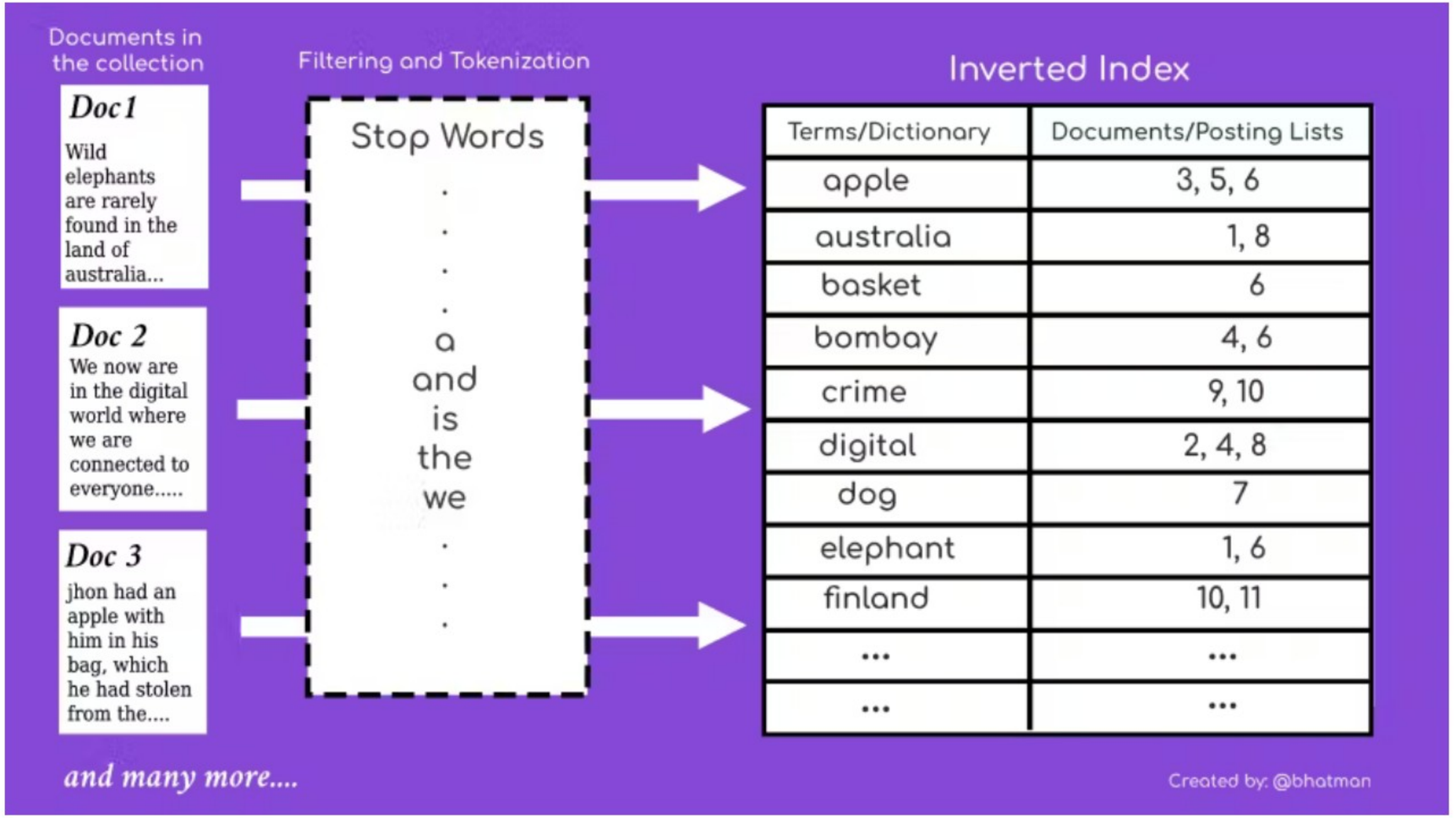
- "a", "the", "her"
- Betyr ikke noe
- Mange av dem -> dyrt å behandle
- Ignorer!

Rep: Inverted index

- Mapping: term -> posting list
- Som registeret i ei bok
- 1/2 Oblig A (2024-09-13)

Rep: Posting list

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

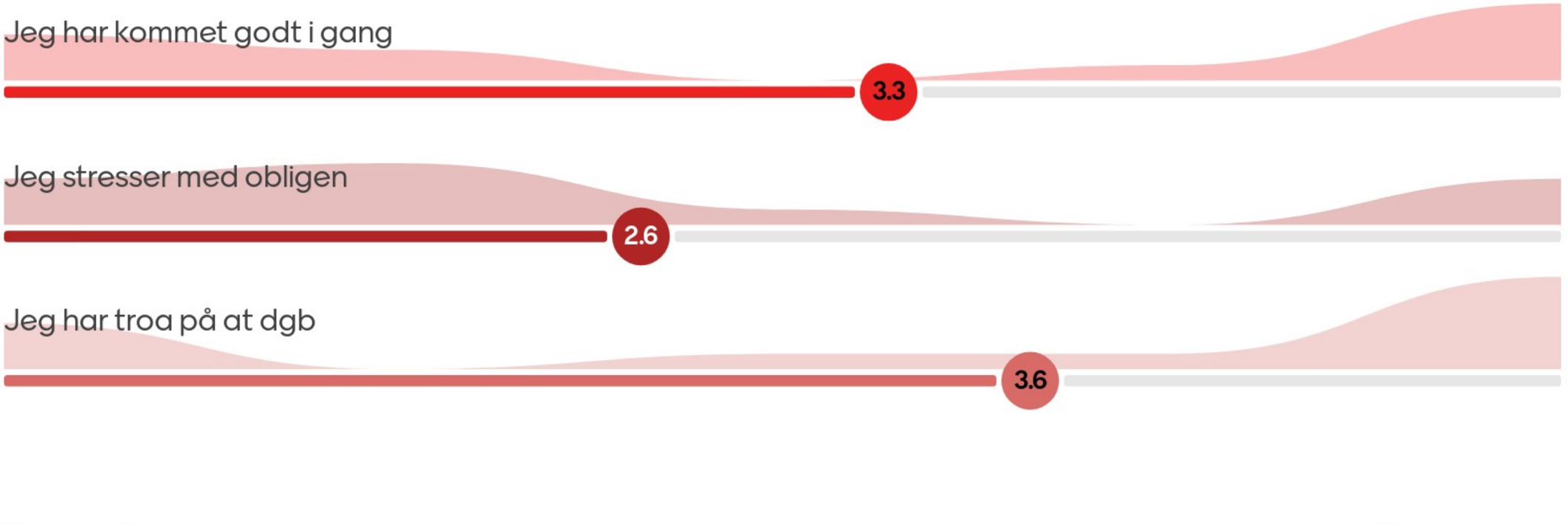


Visualisering: inverted index m/posting lists

Primitivt søk

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

Status for oblig A (frist om 1.5 uke)



"Adhere to the API"

Nøkkelen til å mestre prekoden. Les kommentarer og følg speccen

Strings (lecture 2)

Algorithms and data strucutres

Suffix arrays

- Data structure for search
- Find matching terms from suffixes
- Sorted lexigraphically - why?

"LOG ER KOSELIG" ~ØHRN, 2024-09-02

Hvorfor sorterer man suffixes alfabetisk?

Effektivitet

binærsøk

fordi det er så fint og hyggelig

binærsøk :)

binærsøk

vet ikke

Har lyst

å gå gjennom en million elementer er suppetreigt

1

8

"LOG ER KOSELIG" ~ØHRN, 2024-09-02

Hvorfor sorterer man suffixes alfabetisk?

Finne ord som ender med samme suffix

Spørsmål: hva er egentlig datastrukturen? Nøstet liste? Eller dict?

Suffix Array Example

Given String: banana

Suffixes		Sorted Suffixes
0 banana		5 a
1 anana	Sort the Suffixes	3 ana
2 nana	---------->	1 anana
3 ana	alphabetically	0 banana
4 na		4 na
5 a		2 nana

Suffix array: {5, 3, 1, 0, 4, 2}

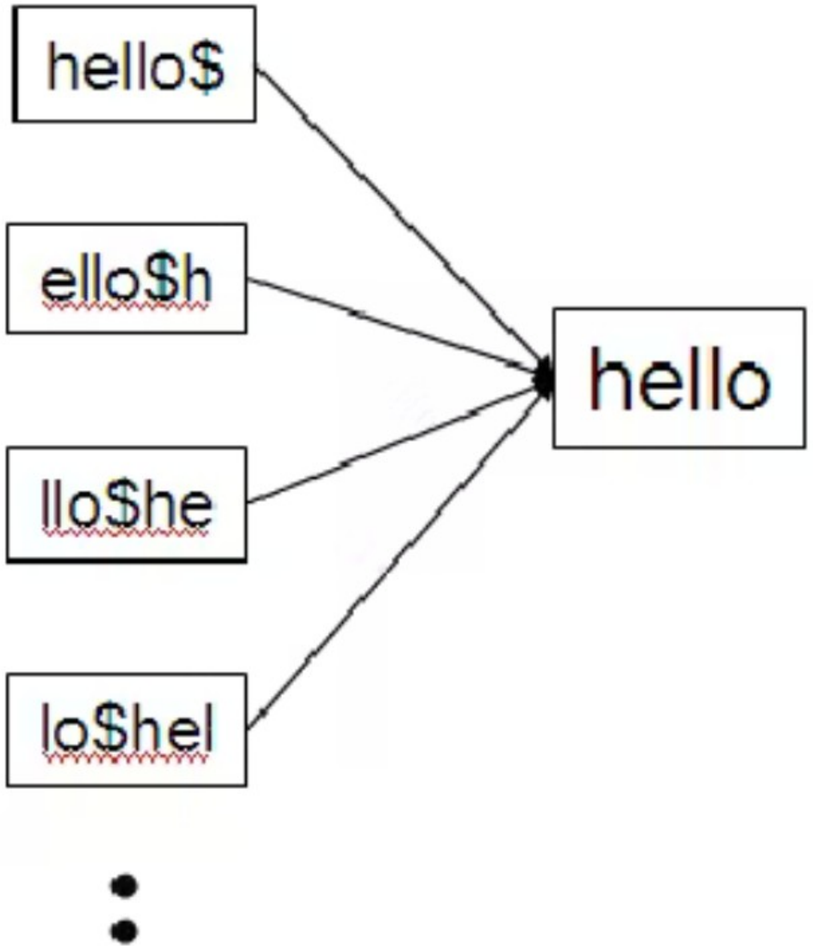
Visualisering av suffix et basic suffix array (Oblig B)

NB: oblig B-1: Token boundries

- → Ett suffix for hvert token
- → Ikke ett suffix for hvert tegn
- → Prekoden sin tokeniser har en metode ranges()

Permuterm indeces

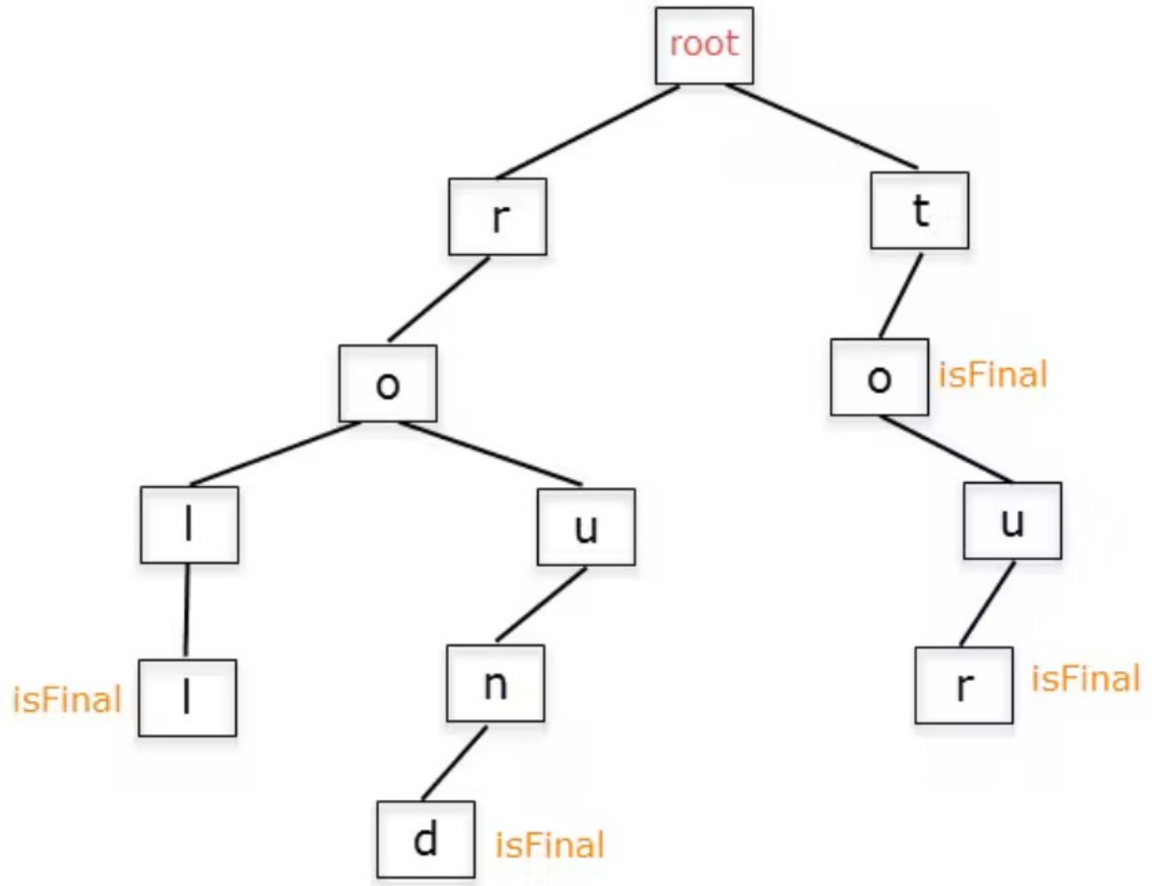
- "permutasjoner av typene"-index
- Støtter wildcard queries, e.g. /.ake/
- lsh litt samme use case som suffix arrays
- Lagrer typene fra korpuset "rotert"



Permuterm-index-struktur for termen "hello"

Tries

- Data structure - prefix tree
- Finn ut om en streng inngår i et korpus (raskt)
- Oblig B
- Mer om dette neste uke



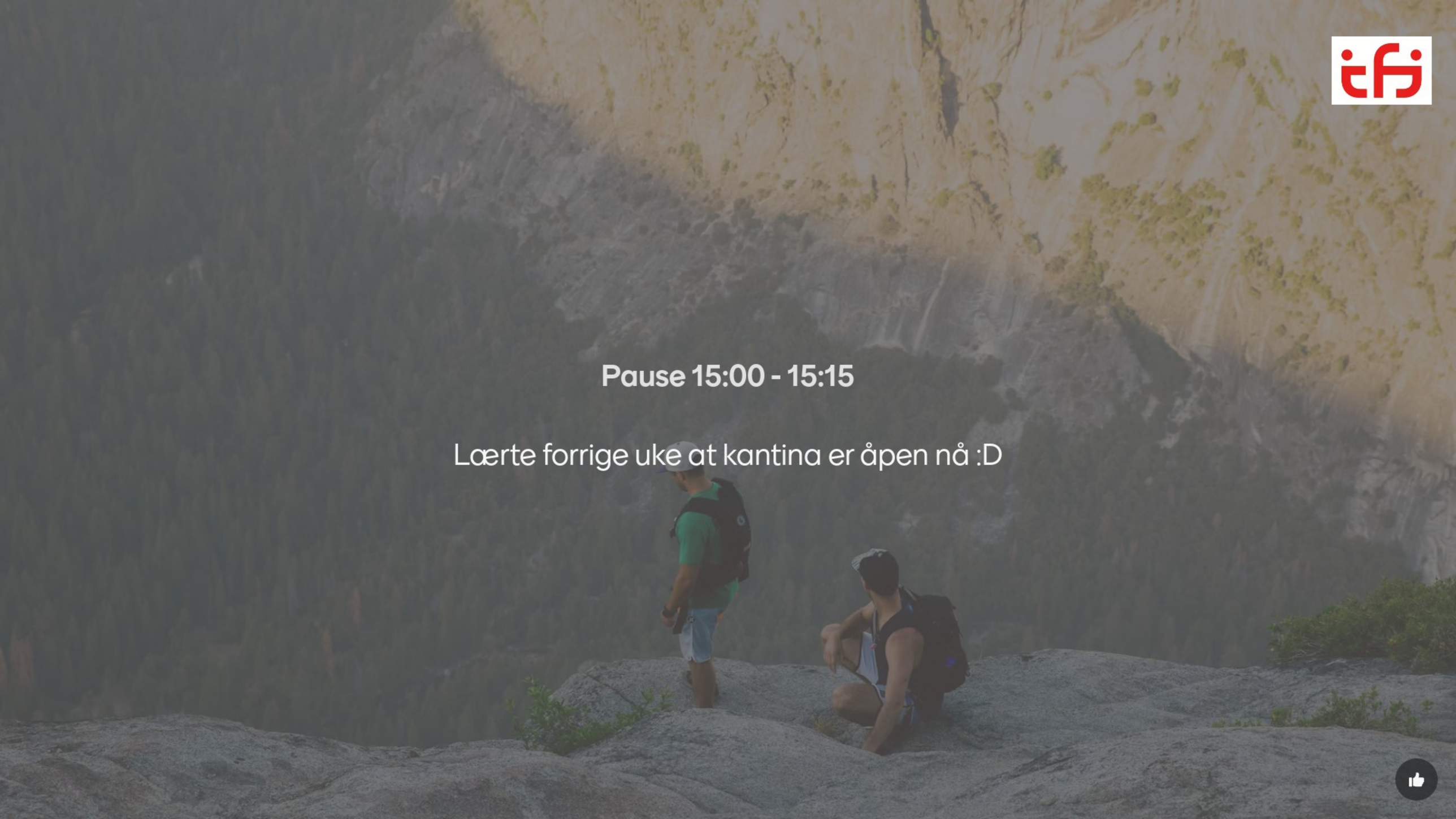
Visualisering av trie

Spørsmål? 

Resten av tiden (til 16): Oblighjelp

Neste gang: strengealgoritmer og oblighjelp

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


Pause 15:00 - 15:15
Lærte forrige uke at kantina er åpen nå :D