IN3120/4120 2022-09-14

(postings, posting lists), strings, suffix arrays, tries, oblig

Agenda du jour

1. Lyn-recap av postings (utgikk forrige uke)
2. Suffix arrays
3. Tries
4. Introdusere oblig B
5. Et knippe oppgaver (på menti)
6. Oblig-workshop

Postings

- Et par: <document id, term frequency>
- "Dennne termen finnes y ganger i document x"
- Organiseres i postinglister (neste slide)

Posting lists

- Består av postings. Basically bare en vanlig liste.
- Må alltid være sortert (så vi kan ha effektive algoritmer)
- Lar oss gjøre AND/OR-queries ved boolean retrieval
- "Her er alle dokumentene i corpuset som inneholder dette settet med termer"

Oblig A

- Inverted index
- OR mellom 2 posting lists
- AND mellom 2 posting lists
- Frist denne fredagen
- Omfang: under 100 linjer

Suffix arrays

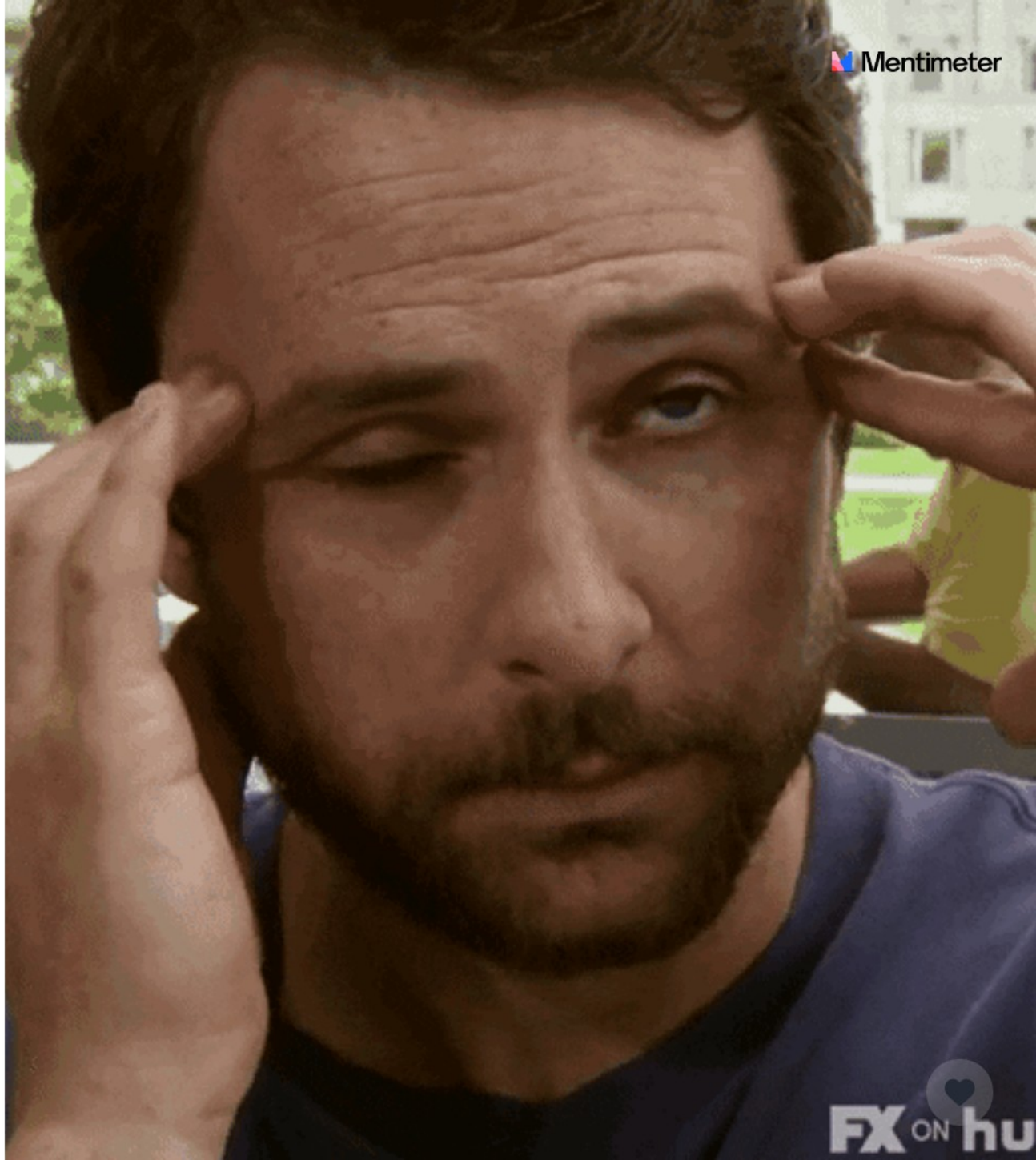
- Bullet 1
- Bullet 2
- Bullet 3

Tries

- Bygg et søketre fra termene i korpuset
- Sjekk en query term opp mot treet (oblig 2) -> finn ut om det er noen matches
- Effektiv struktur, skalerer bra
- Kan også brukes for approximate matching, se paper [H. Shang, T. H. Merrett 1995]
- 10/10 anbefalt YT-video: O7_w001f58c

Oblig B

- Suffix arrays
- Trie search
- Harder than A (the hardest?)
- Prøv å begynne på obligen før neste seminar!!



Exercises

Stjålet fra ymse steder

Select Answer

Mentimeter

The correct answer is: Binary search on a sorted array of the dictionary

Leaderboard

Mentimeter

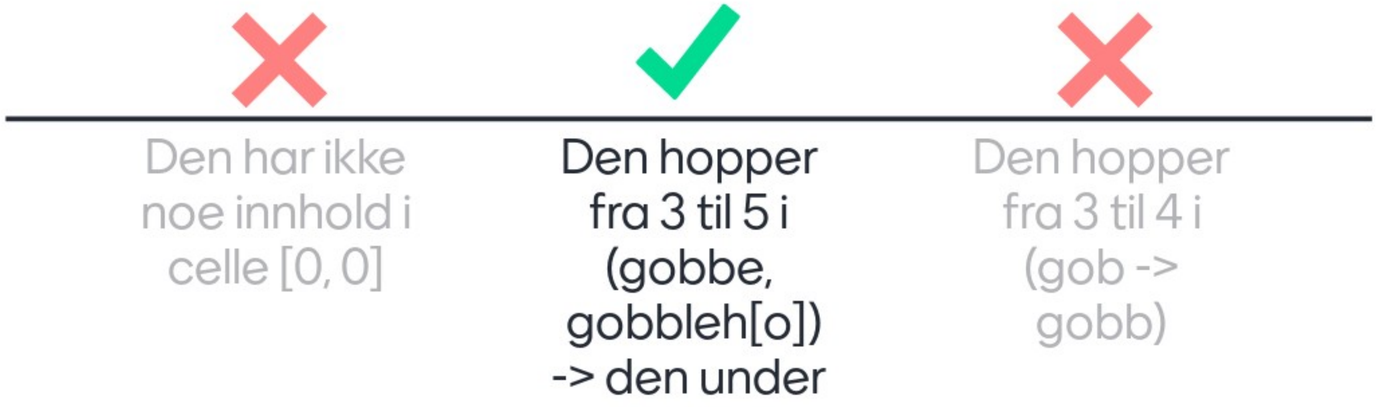
0 p  Kaptein Sabeltann

(e) [6%] If the requirement changes from exact prefix matching to allowing up to k edit errors when matching the prefixes, suggest a suitable data structure and describe the associated lookup algorithm. You can assume that the value of k is 0, 1 or 2 and determined as a function of the length of the prefix typed by the user.

~Les papers!~

Hva er feil med denne edit-tabellen hvis man har unit edit costs?

Mentimeter

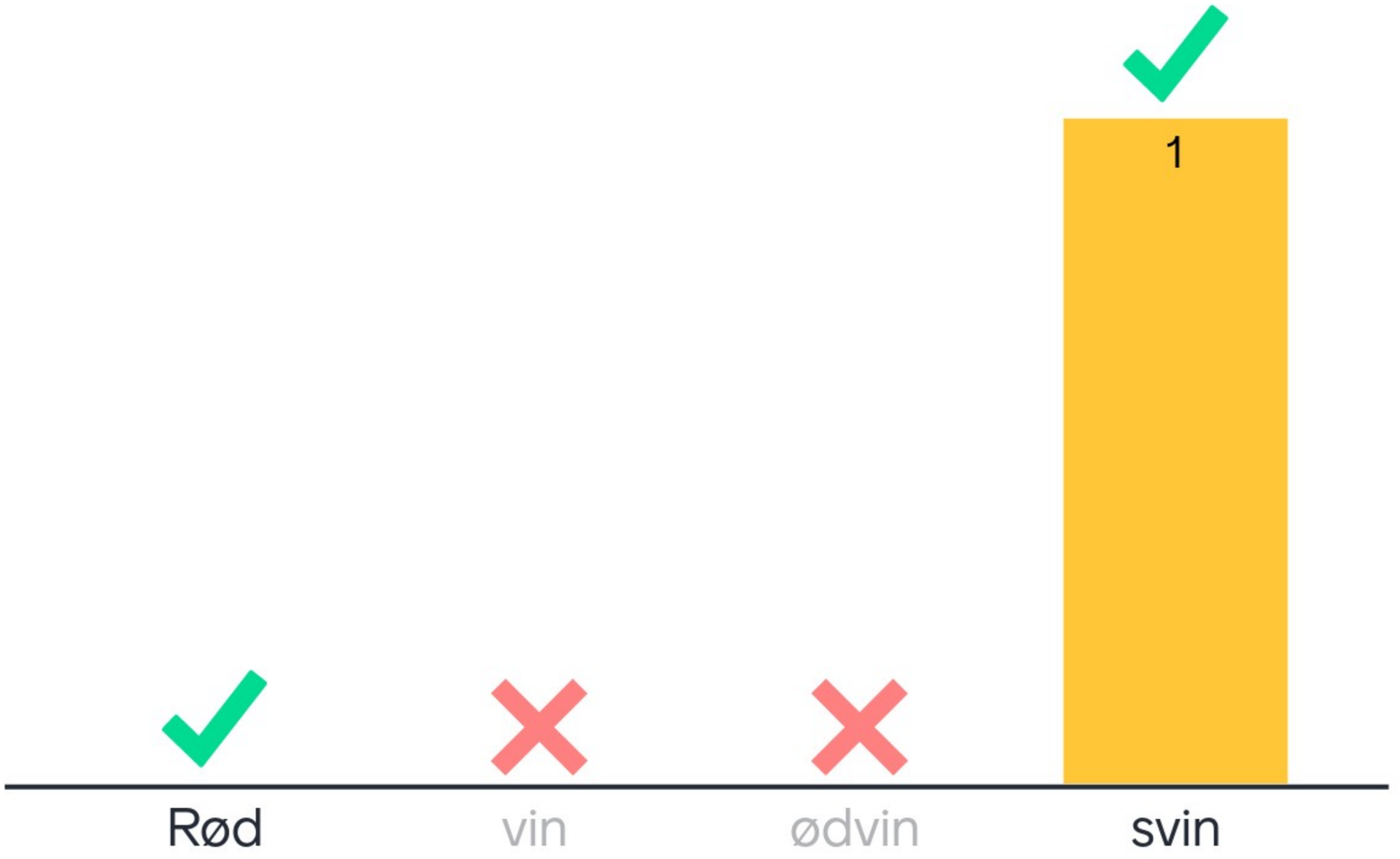


Leaderboard

Mentimeter

0 p  Kaptein Sabeltann

Hva er ikke et suffix av "rødvin"?



Leaderboard

Mentimeter

700 p

Kaptein Sabeltann

fastest

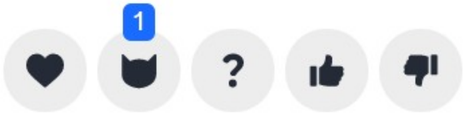
1

Oblig A common fails + hints

- Bruk Mattermost! Mange gode tips i kanalene + spørre om hjelp
- Man trenger ikke mer enn ~10 linjer for å bygge indexen
- ALDRI ta list(p1) -> obligen stryker selv om testene funker
- Postingsmerger må skalere lineært med lengden på postinglistene

Jobbe med oblig. Enten A eller B

Alle burde være ferdig med A innen i kveld. Still spørsmål så du kommer i mål B)!



Si hva som helst:

Mentimeter

hva som helst:

1