Søketeknologi – Gruppe 1

trulshes@uio.no

Agenda

Første time

- Litt om meg
- Praktisk info
- Repetisjon
- Assignment A
- Ukas shoutout

Andre time

- Selvstendig jobbing/starte på assignment A

Truls Hestetræet

- trulshes@uio.no (trulshes på Mattermost)
- 5. året på prosa-master
- Tok faget i fjor (aldri vært gruppelærer)

Agenda

Første time

- Litt om meg
- Praktisk info
- Repetisjon
- Assignment A
- Ukas shoutout

Andre time

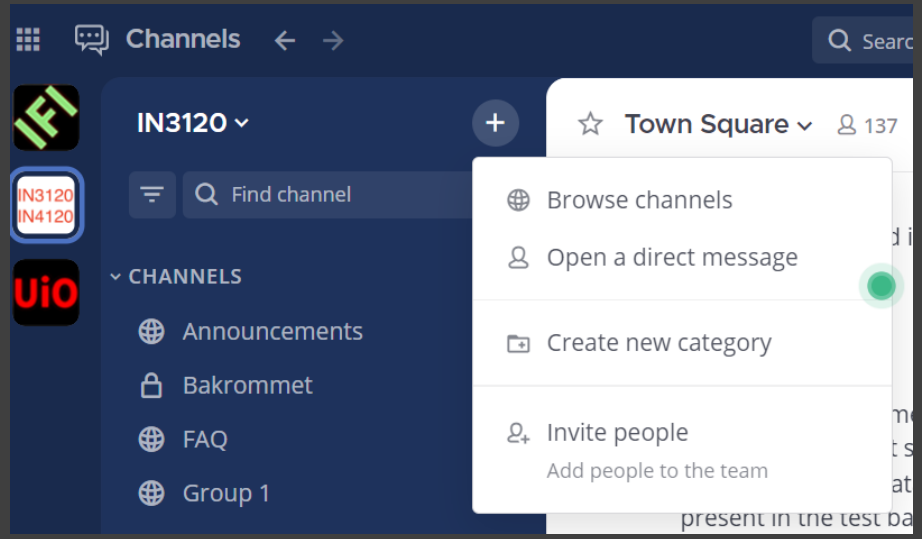
- Selvstendig jobbing/starte på assignment A

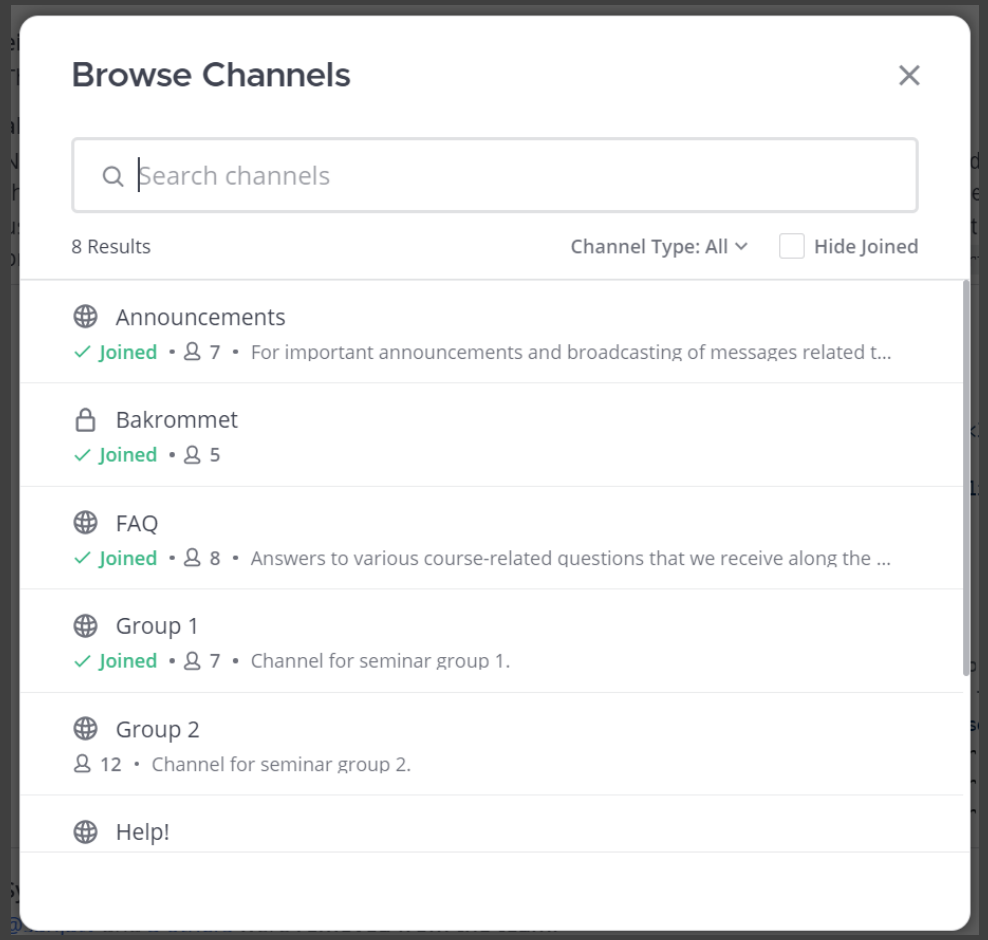
Github

- Alle ressurser for semesteret
  - Slides
  - Obliger
  - Gruppetime-slides
- Trenger ikke pulle etter obliger

Mattermost

- Emnesiden blir ikke brukt
- Aleksander kan gi oblig-tips
- Join Gruppe 1!





Eksamensoversikt

Agenda

Første time

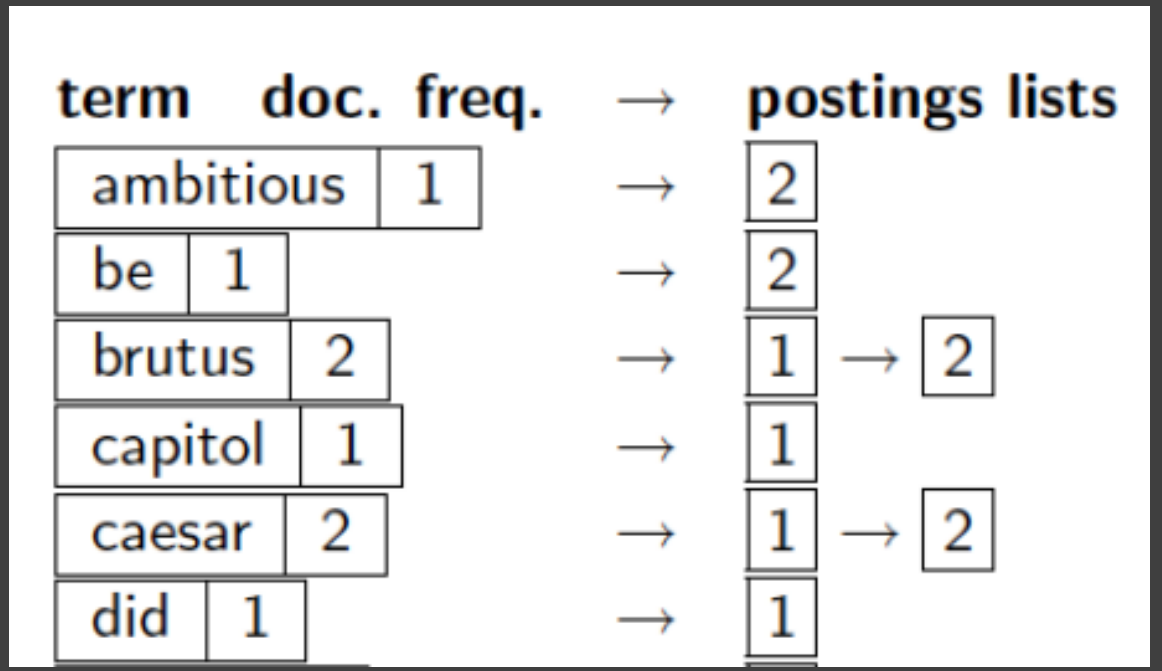
- Litt om meg
- Praktisk info
- Repetisjon
- Assignment A
- Ukas shoutout

Andre time

- Selvstendig jobbing/starte på assignment A

Postingliste

- Liste med postinger for en spesifikk term
- Posting: Metadata om en term i et dokument
  - Dokument-ID
  - Frekvens
  - Posisjon
  - ...



Invertert indeks

- Består av ordbok og postinglister
- Ordboka mapper termer til postingliste-***referanser***

```
invertert_indeks = {
    "søketek": [1, 3],
    "informatikk": [1, 4],
    "indeks": [4]
}
```

```
invertert_indeks_med_frekvens = {
    "søketek": [(1, 4), (3, 42)],
    "informatikk": [(1, 1), (4, 2)],
    "indeks": [(4, 100)]
}
```

Stop words

- Ord som forekommer veldig ofte
- «the», «a», «to», «of»
- Dyrt å behandle, ofte lite betydning
- Kan filtreres bort

NASA astronauts Sunita Williams and Barry “Butch” Wilmore were supposed to be on the International Space Station for eight days. But as the Boeing Starliner capsule they were in was approaching the space station, the spacecraft’s thrusters started to fail. Since then, Boeing and NASA have struggled to figure out what went wrong. NASA decided last week that the astronauts should stay put for eight months until they could come back in a SpaceX capsule.

Boolean retrieval

- «Informatikk er gøy» → «Informatikk» AND «er» AND «gøy»
  - Postingsmerger i Assignment A
- Ulempe: Vil også matche «Informatikk er ikke gøy»

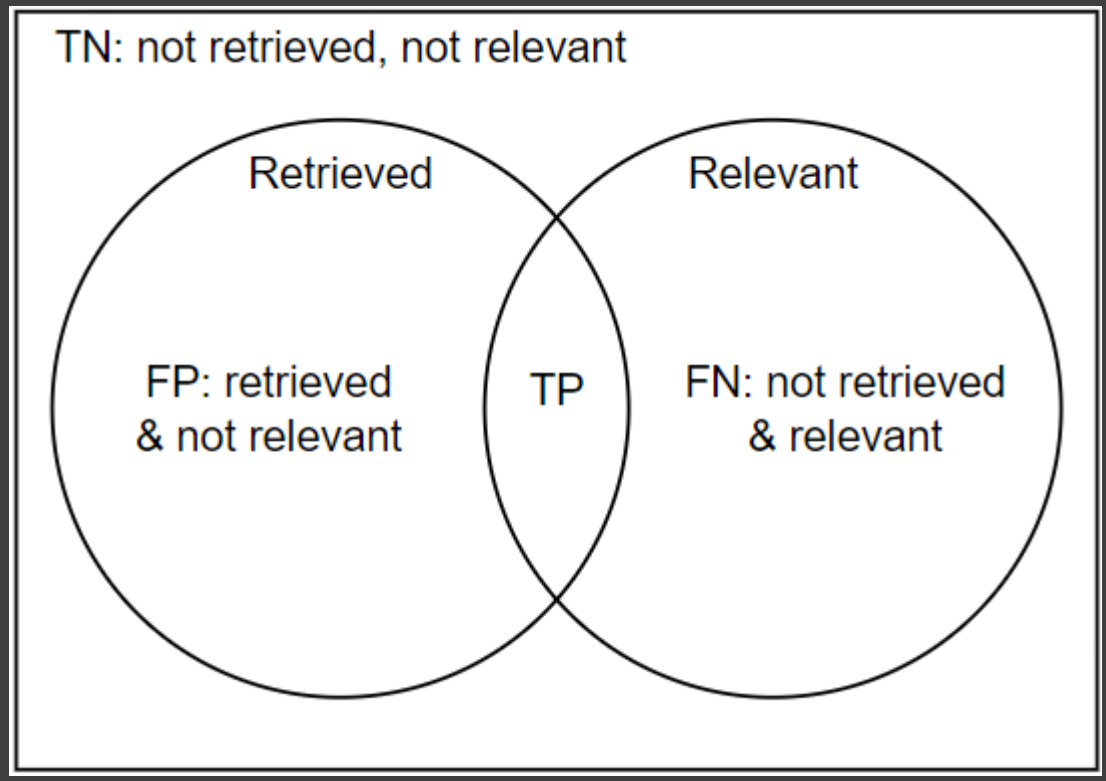
Precision & recall

Precision: Av dokumentene vi hentet, hvor mange er relevante?

$$\frac{True\ positive}{True\ postivie + False\ positive}$$

Recall: Av alle relevante dokumenter, hvor mange henta vi?

$$\frac{True\ positive}{True\ postivie + False\ negative}$$



Agenda

Første time

- Litt om meg
- Praktisk info
- Repetisjon
- Assignment A
- Ukas shoutout

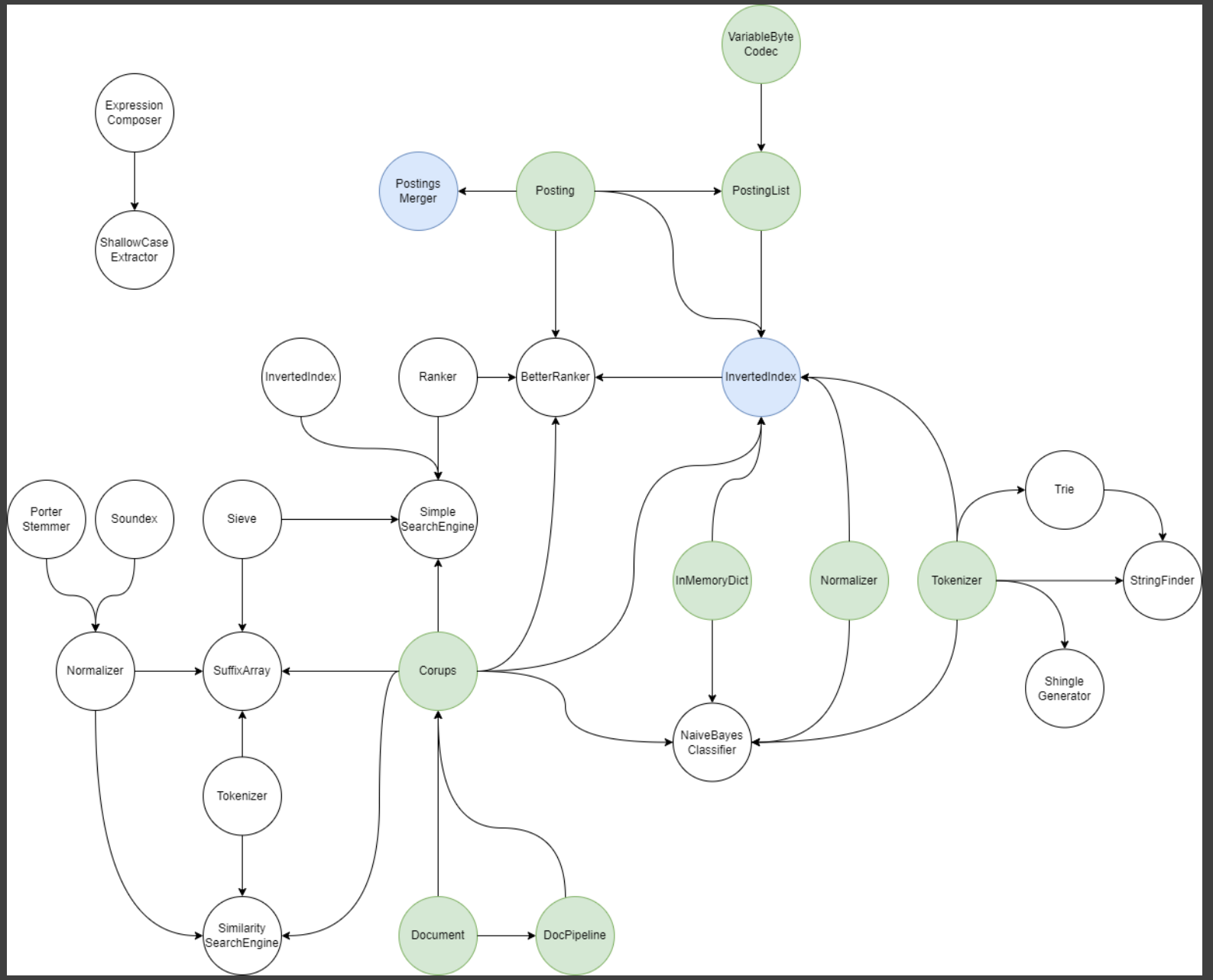
Andre time

- Selvstendig jobbing/starte på assignment A

Assignment A

Tips og triks

- Les teksten nøye
- Skjønne testene
- Få enkelttester og asserts til å kjøre


VariableByte Codec
Expression Composer
Postings Merger
Posting
PostingList
ShallowCase Extractor
InvertedIndex
Ranker
BetterRanker
InvertedIndex
Trie
Porter Stemmer
Soundex
Sieve
Simple SearchEngine
InMemoryDict
Normalizer
Tokenizer
StringFinder
Normalizer
SuffixArray
Corups
Shingle Generator
NaiveBayes Classifier
Tokenizer
Similarity SearchEngine
Document
DocPipeline

invertedindex.py

- Se på byggeklossene hver for seg
- Hvilke instansvariabler finnes?
- Les metode-signaturene

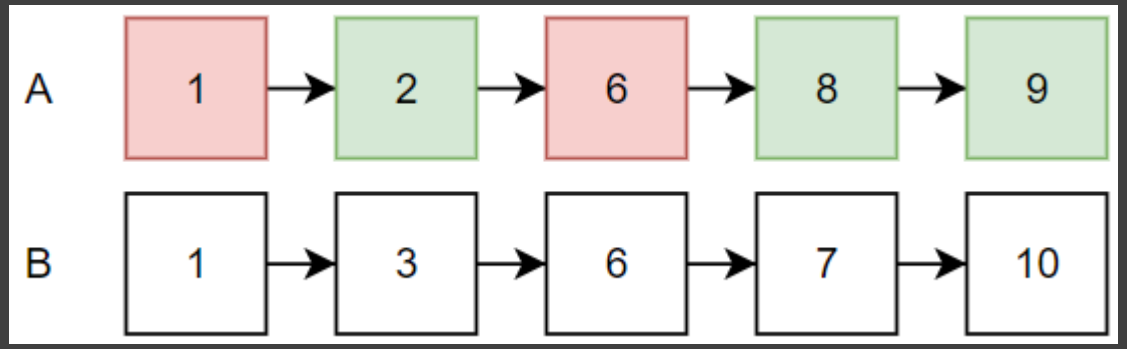
```
def _build_index(self, fields: Iterable[str], compressed: bool) -> None:
def _add_to_dictionary(self, term: str) -> int:
def _append_to_posting_list(self, term_id: int, document_id: int, term_frequency: int,
compressed: bool) -> None:
def get_terms(self, buffer: str) -> Iterator[str]:
```

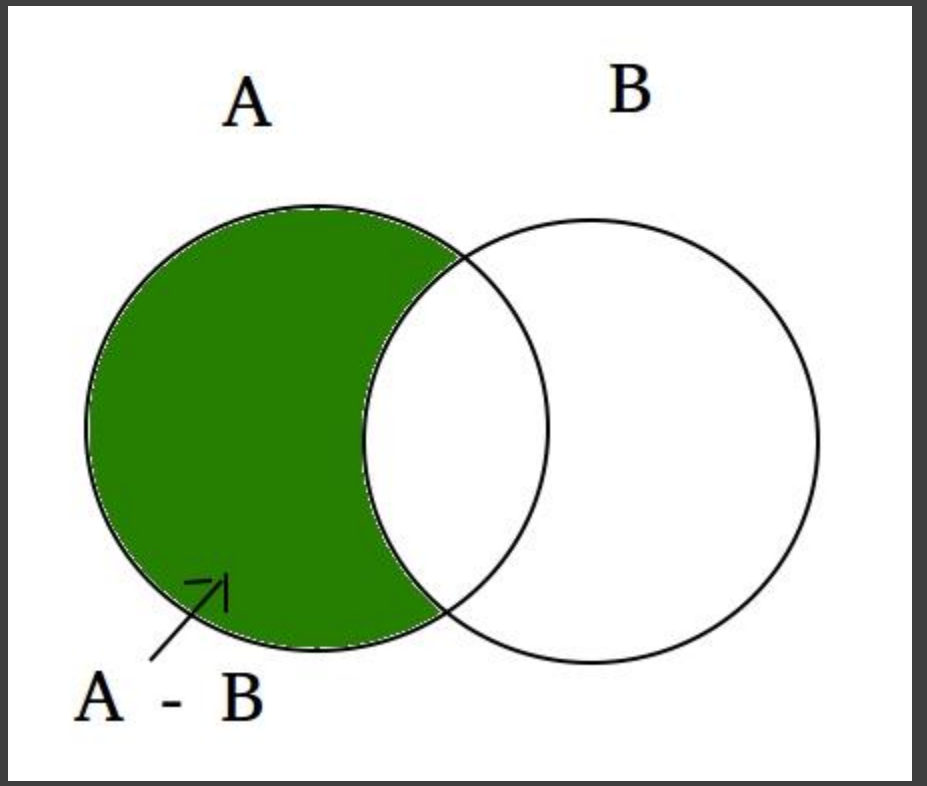
postingsmerger.py

- Gitt to postingslister, finn snitt, union eller difference
- Vi bruker iterator
  - current = next(postinglist, None)
  - yield current
- Bruk prekoden fra forelesningene
  - Union og difference er veldig likt snitt

Difference

$$A \backslash \mathrm{B} = 2, 8, 9$$





Agenda

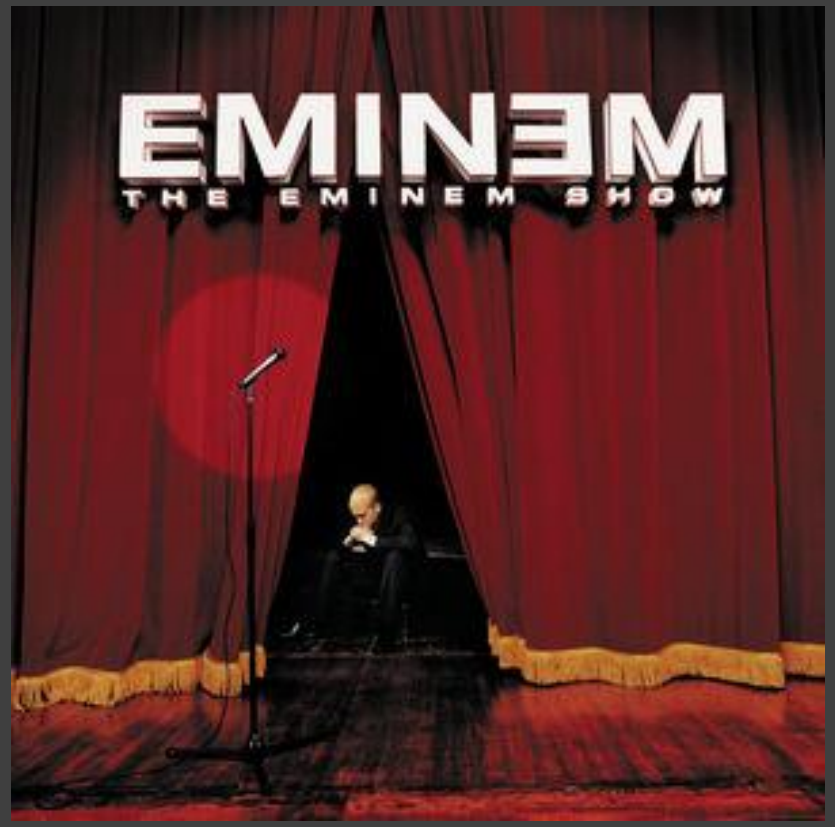
Første time

- Litt om meg
- Praktisk info
- Repetisjon
- Assignment A
- Ukas shoutout

Andre time

- Selvstendig jobbing/starte på assignment A

Ukas shoutout



Agenda

Første time

- Litt om meg
- Praktisk info
- Repetisjon
- Assignment A
- Ukas shoutout

Andre time

- Selvstendig jobbing/starte på assignment A