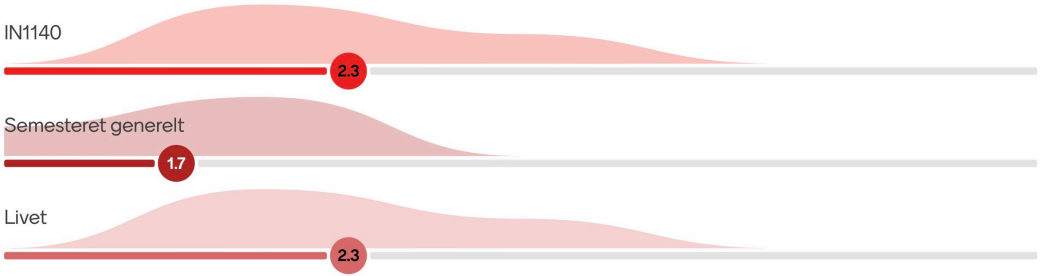
IN1140 Gruppetime #5

Oblig 1b!
(og litt glatting...)

Dagens gruppetime

- Mini menti
- Begreper innenfor sannsynlighet
- Glatting (+MLE)
- Oblig 1b

"Jeg har kontroll på..."



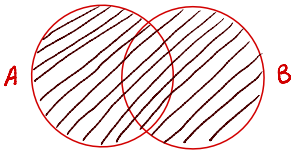
Viktige prinsipper

Med gjentakelse
Liste!

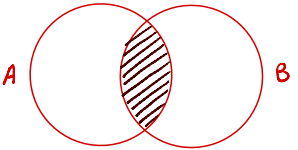
Uten gjentakelse
Set (mengde)!

- Tokens VS ordtyper
- Delmengde, Union og snitt
- Kardinalitet (størrelse)

$(A \cup B)$



$(A \cap B)$



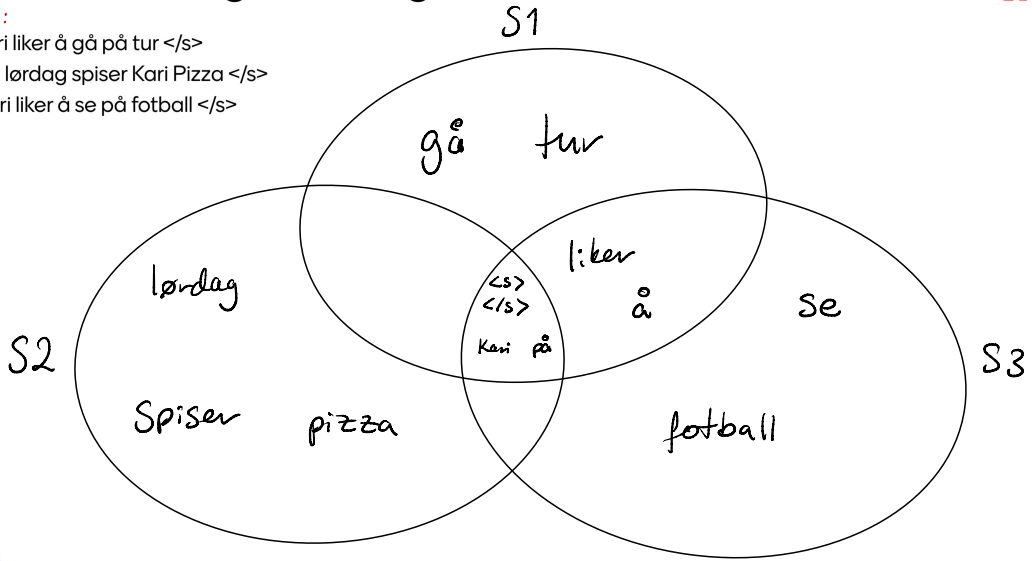
Markov, MLE og Glatting

Mini korpus:

S1: <s> Kari liker å gå på tur </s>

S2: <s> På lørdag spiser Kari Pizza </s>

S3: <s> Kari liker å se på fotball </s>



Kardinalitet:

$|S1| = 8$

$|S2| = 7$

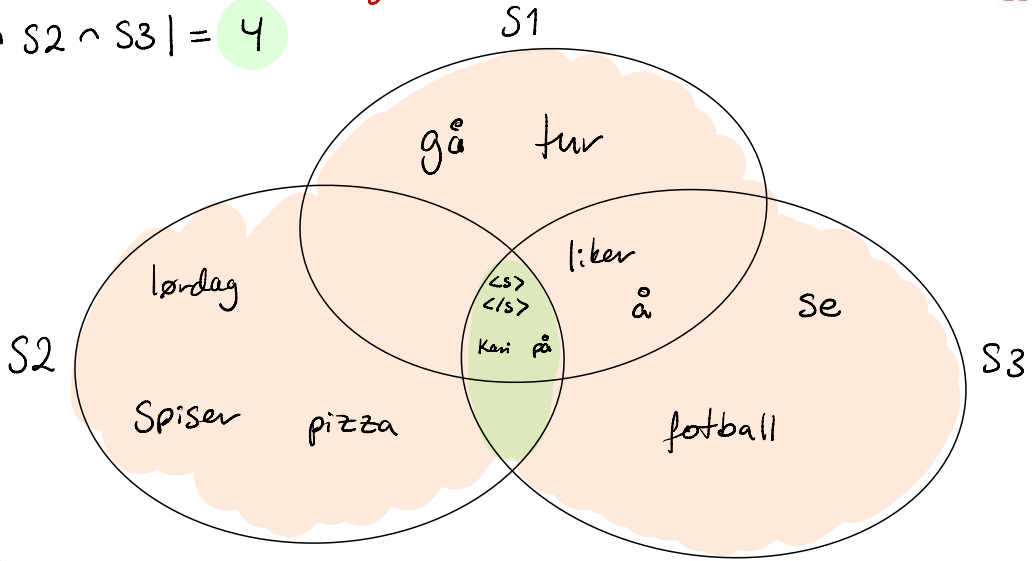
$|S3| = 8$

$= 23$ = antall tokens! (i korpuset)



$|S1 \cup S2 \cup S3| = 13$ = antall ordtyper i korpuset!
Også kjent som vokabularet, V

$|S1 \cap S2 \cap S3| = 4$



Markov, MLE og Glatting

S1: <s> Kari liker å gå på tur </s>
S2: <s> På lørdag spiser Kari Pizza </s>
S3: <s> Kari liker å se på fotball </s>

- Vi ønsker å finne sannsynligheten til ett ord i sammenheng av foregående ord.
  Markovantagelsen gjør det mulig å bare se på n-antall foregående ord → ets. bigram og trigram.

Bigrammer i korpuset:

⟨⟨s⟩, Kari⟩, ⟨Kari, liker⟩, ⟨liker, å⟩, ⟨å, gå⟩, ⟨gå, på⟩, ⟨på, tur⟩, ⟨tur, ⟨/s⟩⟩,
⟨⟨s⟩, På⟩, ⟨på, lørdag⟩, ⟨lørdag, spiser⟩, ⟨spiser, Kari⟩, ⟨Kari, pizza⟩, ⟨pizza, ⟨/s⟩⟩,
⟨⟨s⟩, Kari⟩, ⟨Kari, liker⟩, ⟨liker, å⟩, ⟨å, se⟩, ⟨se, på⟩, ⟨på, fotball⟩, ⟨fotball, ⟨/s⟩⟩

Markov, MLE og Glatting

S1: <s> Kari liker å gå på tur </s>
S2: <s> På lørdag spiser Kari Pizza </s>
S3: <s> Kari liker å se på fotball </s>

- Utregning av sannsynligheten for bigrammer
  - ↳ Maximum likelihood Estimation (MLE)

Sannsynlighet for "Kari liker pizza?" ⇒ ⟨⟨s⟩, Kari⟩, ⟨Kari, liker⟩, ⟨liker, pizza⟩, ⟨pizza, ⟨/s⟩⟩

"Kari, gitt setningsstart"

$$P = \frac{\text{Count}(\text{Kari}, \langle s \rangle)}{\text{Count}(\text{Kari})} \cdot \frac{\text{Count}(\text{liker}, \text{Kari})}{\text{Count}(\text{liker})} \cdot \frac{\text{Count}(\text{pizza}, \text{liker})}{\text{Count}(\text{pizza})} \cdot \frac{\text{Count}(\langle /s \rangle, \text{pizza})}{\text{Count}(\langle /s \rangle)}$$

$$= \frac{2}{3} \cdot \frac{2}{2} \cdot \frac{0}{1} \cdot \frac{1}{3}$$

↳ forklaring →

Markov, MLE og Glatting

S1: <s> Kari liker å gå på tur </s>

S2: <s> På lørdag spiser Kari Pizza </s>

S3: <s> Kari liker å se på fotball </s>

Bigrammer i korpuset:

⟨⟨s⟩, Kari⟩, ⟨Kari, liker⟩, ⟨liker, å⟩, ⟨å, gå⟩, ⟨gå, på⟩, ⟨på, tur⟩, ⟨tur, ⟨/s⟩⟩,

⟨⟨s⟩, På⟩, ⟨på, lørdag⟩, ⟨lørdag, spiser⟩, ⟨spiser, Kari⟩, ⟨Kari, pizza⟩, ⟨pizza, ⟨/s⟩⟩,

⟨⟨s⟩, Kari⟩, ⟨Kari, liker⟩, ⟨liker, å⟩, ⟨å, se⟩, ⟨se, på⟩, ⟨på, fotball⟩, ⟨fotball, ⟨/s⟩⟩

Sannsynlighet for "Kari liker pizza?" ⇒ ⟨⟨s⟩, Kari⟩, ⟨Kari, liker⟩, ⟨liker, pizza⟩, ⟨pizza, ⟨/s⟩⟩

$$P = \frac{\text{Count}(\text{Kari}, \langle s\rangle)}{\text{Count}(\text{Kari})} \cdot \frac{\text{Count}(\text{liker}, \text{Kari})}{\text{Count}(\text{liker})} \cdot \frac{\text{Count}(\text{pizza}, \text{liker})}{\text{Count}(\text{pizza})} \cdot \frac{\text{Count}(\langle /s\rangle, \text{pizza})}{\text{Count}(\langle /s\rangle)}$$

$$= \frac{2}{3} \cdot \frac{2}{2} \cdot \frac{0}{1} \cdot \frac{1}{3}$$

Markov, MLE og Glatting

S1: <s> Kari liker å gå på tur </s>
S2: <s> På lørdag spiser Kari Pizza </s>
S3: <s> Kari liker å se på fotball </s>

Sannsynlighet for "Kari liker pizza?"

Gjør at sannsynligheten for setningen blir lik 0

Betyr det at setningen "Kari liker pizza" er umulig?

$$P = \frac{\text{Count}(\text{Kari}, \langle s \rangle)}{\text{Count}(\text{Kari})} \cdot \frac{\text{Count}(\text{liker}, \text{Kari})}{\text{Count}(\text{liker})} \cdot \frac{\text{Count}(\text{pizza}, \text{liker})}{\text{Count}(\text{pizza})} \cdot \frac{\text{Count}(\langle /s \rangle, \text{pizza})}{\text{Count}(\langle /s \rangle)}$$

$$= \frac{2}{3} \cdot \frac{2}{2} \cdot \boxed{\frac{0}{1}} \cdot \frac{1}{3}$$

Markov, MLE og Glatting

S1: <s> Kari liker å gå på tur </s>
S2: <s> På lørdag spiser Kari Pizza </s>
S3: <s> Kari liker å se på fotball </s>

Nei! Vi kan bruke add-one smoothing (Laplace) :)

Vi bruker vokabularet, V, som er antall likt som ordtyper i korpuset

Sannsynlighet for "Kari liker pizza?" ⇒ ⟨⟨s⟩, Kari⟩, ⟨Kari, liker⟩, ⟨liker, pizza⟩, ⟨pizza, ⟨/s⟩⟩

$$P = \frac{\text{Count}(\text{Kari}, \langle s \rangle)}{\text{Count}(\text{Kari})} \cdot \frac{\text{Count}(\text{liker}, \text{Kari})}{\text{Count}(\text{liker})} \cdot \frac{\text{Count}(\text{pizza}, \text{liker})}{\text{Count}(\text{pizza})} \cdot \frac{\text{Count}(\langle /s \rangle, \text{pizza})}{\text{Count}(\langle /s \rangle)}$$

$$= \frac{2+1}{3+V} \cdot \frac{2+1}{2+V} \cdot \frac{0+1}{1+V} \cdot \frac{1+1}{3+V}$$

Markov, MLE og Glatting

S1: <s> Kari liker å gå på tur </s>
S2: <s> På lørdag spiser Kari Pizza </s>
S3: <s> Kari liker å se på fotball </s>

Sannsynlighet for "Kari liker pizza?" $\Rightarrow \langle\langle s\rangle, \text{Kari}\rangle, \langle \text{Kari}, \text{liker}\rangle, \langle \text{liker}, \text{pizza}\rangle, \langle \text{pizza}, \langle/s\rangle\rangle$

$$P = \frac{\text{Count}(\text{Kari}, \langle s\rangle)}{\text{Count}(\text{Kari})} \cdot \frac{\text{Count}(\text{liker}, \text{Kari})}{\text{Count}(\text{liker})} \cdot \frac{\text{Count}(\text{pizza}, \text{liker})}{\text{Count}(\text{pizza})} \cdot \frac{\text{Count}(\langle/s\rangle, \text{pizza})}{\text{Count}(\langle/s\rangle)}$$

$$= \frac{2+1}{3+13} \cdot \frac{2+1}{2+13} \cdot \frac{0+1}{1+13} \cdot \frac{1+1}{3+13} = 3{,}35 \cdot 10^{4}$$

Markov, MLE og Glatting

S1: <s> Kari liker å gå på tur </s>
S2: <s> På lørdag spiser Kari Pizza </s>
S3: <s> Kari liker å se på fotball </s>

Sannsynlighet for "Kari liker pizza?"

Litt mer om glatting

Dersom et ord i setningen vi ønsker å finne sannsynligheten for ikke eksisterer i korpuset, for eksempel "Kari liker ananas på pizza", vil vi få et null under brøkstreken.
Da vil ikke bare $P=0$, men vi vil få regne feil som kan skape store problemer i programmet vårt.
Da er det veldig praktisk, og viktig, med glatting.