# IN1140: Introduksjon til språkteknologi

## *Forelesning #13*

Samia Touileb

Universitetet i Oslo

21. november 2019

# 1. Regulære uttrykk (2 poeng)

Hvilket av alternativene kan ikke gjenkjennes med det følgende regulære uttrykket:

```
[1-9][0-9]*\s((cent(s)?)|(dollar(s)?
\s+([1-9][0-9]*\scent(s)?)))
```

Velg ett alternativ

1. 1 dollar 35 cents
2. 35 dollars
3. 99 cents
4. 99 dollars 1 cent

# 1. Regulære uttrykk (2 poeng)

Svar 1: 35 dollars

# 2. Tokenisering (10 poeng)

**Tokenisering** er en viktig oppgave i språkteknologiske systemer og går ut på å dele opp en tekst i løpende ord. Ta for deg følgende setning og besvar spørsmålene under.

*På låven sitter nissen med sin julegrøt , så god og søt , så god og søt .*

1. Vi skiller ofte mellom **tokens** og **typer** når vi skal regne på ordforekomster i en tekst. Hva er forskjellen mellom disse? Hvor mange tokens og typer består eksempelsetningen av?
2. En svært enkel tokeniserer vil splitte en tekst på mellomrom, samt skille ut all tegnsetting, slik det er gjort i eksempelsetningen over. En slik tilnærming vil imidlertid kunne føre til en del problemer. Gi minst to eksempler på ulik bruk av tegnsetting som en slik tokeniserer ikke vil behandle korrekt.

# 2. Tokenisernig (10 poeng)

**Løsningsforslag**:

1. Typer er antall unike ord i teksten, mens tokens er antall løpende ord (der like forekomster telles flere ganger). Teksten inneholder 13 typer og 18 tokens.
2. Fenomener som vil skape problemer for en slik tokeniserer:
   - URL'er: http://www.uio.no
   - forkortelser: f.eks.
   - apostrofer: I'll
   - visse bindestreker: Oslo-borgeren
   - tall: 10,26, 10:26

# 2. Tokenisernig (10 poeng)

**Løsningsforslag**:

1. Typer er antall unike ord i teksten, mens tokens er antall løpende ord (der like forekomster telles flere ganger). Teksten inneholder 13 typer og 18 tokens.
2. Fenomener som vil skape problemer for en slik tokeniserer:
   - URL'er: http://www.uio.no
   - forkortelser: f.eks.
   - apostrofer: I'll
   - visse bindestreker: Oslo-borgeren
   - tall: 10,26, 10:26

**Poengfordeling**:

5 poeng per deloppgave (1 og 2), videre

- deloppgave 1 (5 poeng), videre:
  - 3 poeng for riktig beskrivelse av forskjellen mellom typer og tokens
  - 2 poeng for riktig antall typer og tokens, trekker 1 poeng per feil
- deloppgave 2 (5 poeng), videre:
  - full pott dersom de har nevnt to eller flere riktige svaralternativer
  - trekker 2.5 per feil/manglende svaralternativ

# Ordklasser (7 poeng)

Her skal vi jobbe med følgende setning:

`Filmen har spennende vendinger som tar deg til flotte steder i Oslo og Wien`

Gitt ordklassene i Tabell 1 under, tildel ordklasser til alle ordene i setningen. Du må velge ett alternativ for hvert tilfelle.

| | |
|---|---|
| NOUN | Substantiv |
| VERB | Verb |
| ADJ | Adjektiv |
| PREP | Preposisjon |
| PRON | Pronomen |
| CONJ | Konjunksjon |
| ADV | Adverb |
| SUBJN | Subjunksjon |
| DET | Determinativ |

Table: 1

# Ordklasser (7 poeng)

Her skal vi jobbe med følgende setning:

```
Filmen / NOUN
har / VERB
spennende / ADJ
vendinger / NOUN
som / SUBJN eller PREP
tar / VERB
deg / PRON
til / PREP
flotte / ADJ
steder / NOUN
i / PREP
Oslo / NOUN
og / CONJ
Wien / NOUN
```

# Ordklassetagging (6 poeng)

Algoritmer for ordklassetagging faller under to hovedkategorier, hvilke? Forklar kort hva som kjennetegner disse og hva som skiller dem fra hverandre.

**Løsningsforslag:**

- Regelbaserte taggere: Manuelt definerte regler for å tildele ord riktig tagg i en gitt kontekst. Eksempel: drikke er substantiv, og ikke verb, dersom det følger et adjektiv.
- Statistiske taggere: Bruker et (manuelt) ordklassetagget korpus ("treningskorpus") til å beregne en statistisk model for tagging

**Poengfordeling:**

- Maks 3 poeng. 1 poeng for å si regelbaserte taggere.Beskrivelsen (2 poeng): studenten må nevne regler som blir manuelt definert for å få fullpott. Trekke et poeng om de ikke sier eksplisitt at disse er definert manuelt.
- Maks 3 poeng. 1 poeng for å si statistiske taggere. Beskrivelsen (2 poeng): studentene må nevne et det brukes et ordklassetagget korpus, og at det beregnes en statistisk model basert på den. Trekke 1 poeng om kun 1 av de to elementene er nevnt.

# Kontekstfri grammatikk (10 poeng)

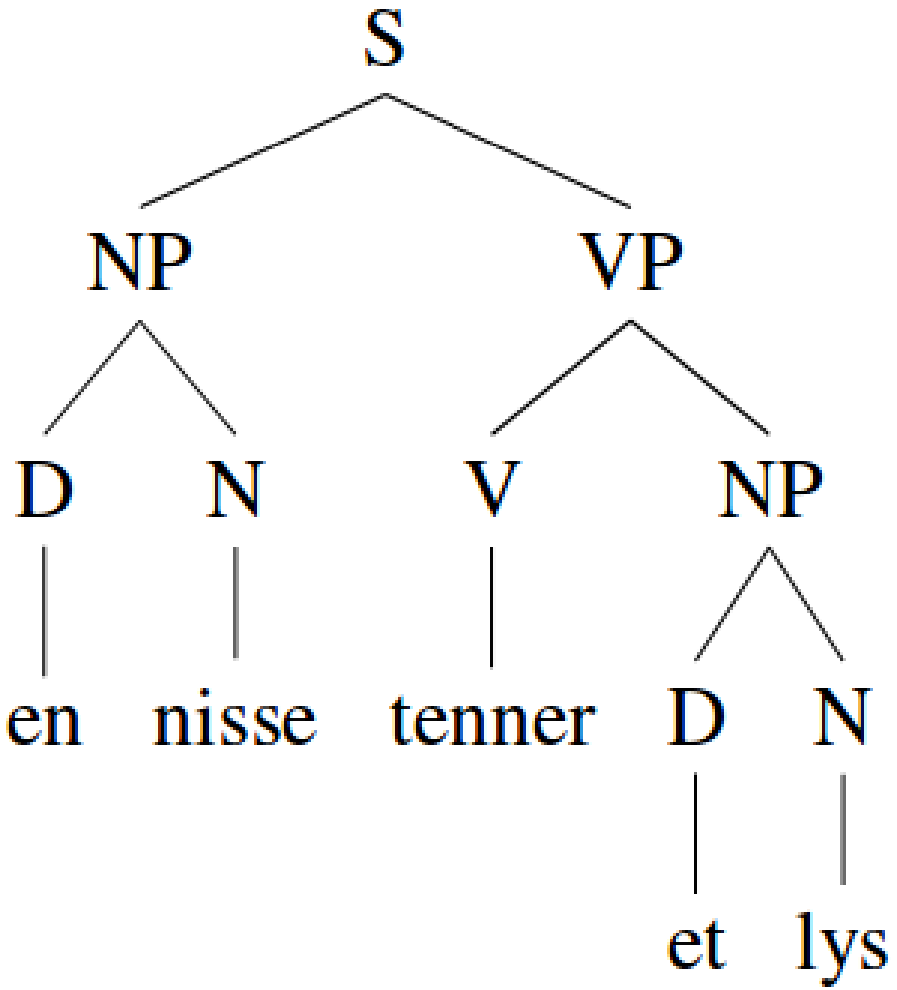
Ta utgangspunkt i vedlagte syntaktiske tre og besvar følgende spørsmål:

1. Angi frasestrukturreglene som tilsvarer det vedlagte treet. Du skal inkludere regler for både ikke-terminale og terminale noder.

2. Vis hvordan du kan utvide grammatikken fra deloppgave 1. til å gi en syntaktisk analyse for følgende setninger:
   - Et barn spiser en pepperkake
   - En nisse synger

3. I norsk har vi samsvarbøyning mellom determinativer og substantiv i substantivfrasen. Vis hvordan du kan utvide grammatikken slik at du utelukker ugrammatiske substantivfraser som de under fra grammatikken.
   - *et nisse
   - *en barn

# Kontekstfri grammatikk (10 poeng)



# Kontekstfri grammatikk (10 poeng)

**Løsningsforslag:**
Treet tilsvarer følgende regler (3 poeng):
S -> NP VP
NP -> D N
VP -> V NP
D -> en | et
N -> nisse | lys
V -> tenner

Vi utvider grammatikken med følgende regler (3 poeng):
VP -> V | V NP
N -> nisse | lys | barn | pepperkake
V -> tenner | synger | spiser

# Kontekstfri grammatikk (10 poeng)

**Løsningsforslag forts.:**
For å utelukke de ugrammatiske frasene modifiserer vi regelene for NP-frasen samt reglene for D og N til å ta høyde for kjønn (4 poeng):

NP -> Dm Nm | Dn Nn
Dm -> en
Dn -> et
Nm -> nisse | pepperkake
Nn -> lys | barn

# Kontekstfri grammatikk (10 poeng)

**Poengfordeling:**
3 deloppgaver: $3 + 3 + 4$

- 3 poeng: trekk 2 poeng for manglende regler, 1 poeng for slurvefeil i regler
- 3 poeng: trekk 1 p for feil/manglende terminalregler (leksikale regler) og 2 p for feil/manglende VP-regel
- 4 poeng: det viktigste her er at besvarelsen viser forståelse av at nye regler må introduseres for å skille mellom ulike kjønn for determinativ og substantiv samt at dette også må gjøres på NP-nivå. Trekker 2 poeng for feil/manglende NP-regel, 1 poeng for feil/manglende terminalregel

# Språkmodeller (10 poeng)

Ta for deg følgende formel for en språkmodell (en såkalt bigrammodell) og et lite tekstkorpus.

**Formel:**

$P(w_1 \dots w_k) = \prod_{i=1}^{k} P(w_i|w_{i-1})$

**Tekstkorpus:**

<s> Jeg liker indisk mat <\s>
<s> Gina liker kinesisk mat <\s>
<s> Thomas foretrekker kinesisk mat <\s>
<s> Jeg foretrekker indisk mat <\s>

# Språkmodeller (10 poeng)

1. Hvilke bigrammer forekommer i korpuset?

2. Hvordan beregner vi sannsynligheten for et ord gitt det foregående ordet fra et korpus?

3. Du skal nå bruke bigrammodellen samt tekstkorpuset til å beregne sannsynligheten for setningen <s> *Jeg foretrekker kinesisk mat* <\s>. Vis hvilke sannsynligheter du trenger samt hvordan disse beregnes fra korpuset. Du trenger ikke å regne ut den totale sannsynligheten for setningen.

# Språkmodeller (10 poeng)

1. Bigrammene:
   <s> Jeg
   Jeg liker
   liker indisk
   indisk mat
   mat <\s>
   <s> Gina
   Gina liker
   iker kinesisk
   kinesisk mat
   <s> Thomas
   Thomas foretrekker
   foretrekker kinesisk
   Jeg foretrekker
   foretrekker indisk

# Språkmodeller (10 poeng)

1. $P(w_i|w_{i-1}) = \frac{C(w_{i-1},w_i)}{C(w_{i-1})}$

2. P(Jeg|<s>) * P(foretrekker|Jeg) * P(kinesisk|foretrekker) * P(mat|kinesisk) * P(<\s>|mat)

   $\frac{C(<s>,Jeg)}{C(<s>)} * \frac{C(Jeg,foretrekker)}{C(Jeg)} * \frac{C(foretrekker,kinesisk)}{C(foretrekker)} * \frac{C(kinesisk,mat)}{C(kinesisk)} * \frac{C(mat,<\s>)}{C(mat)}$

   2/4 * 1/2 * 1/2 * 2/2 * 4/4

**Poengfordeling:**

1. Maks 3 poeng. Det er totalt 14 bigrammer. 1 poeng trekkes om det mangler en eller fler bigrammer. 1 poeng trekkes om <s> og <\s> ikke tas hensyn til i bigrammene.

2. Maks 3 poeng. Her må studenten bevise at hun kan formelen. Formelen må være identisk til løsningsforslaget for å få fullpott. Om posisjon av $w_i$ og $w_{i-1}$ forveksles må det trekkes 2 poeng.

3. Maks 4 poeng. Her må studenten vise stegene for å beregne sannsynlighetene. Trekke 1 poeng om studenten ikke viser $P(w_i|w_{i+1})$ for hvert ord i setningen, og 1 poeng om studenten ikke viser $C(w_{i-1}, w_i)/C(w_{i-1})$ for hvert ord i setningen. Trekke 1 poeng om studenten ikke viser siste delen av løsningsforslaget (altså tellingen).

# Leksikale relasjoner (6 poeng)

## Leksikale relasjoner (6 poeng)

Hvilken semantisk relasjon holder mellom følgende ord-par:

| | Synonymi | Antonymi | Hyponymi | Hypernymi | Meronymi | Homonymi |
|---|---|---|---|---|---|---|
| sykkel - terrengsykkel | ○ | ○ | ○ | ○ | ○ | ○ |
| rask – hurtig | ○ | ○ | ○ | ○ | ○ | ○ |
| bryter (idrettsutøver) - bryter (lysbryter) | ○ | ○ | ○ | ○ | ○ | ○ |
| medlem - forening | ○ | ○ | ○ | ○ | ○ | ○ |
| gammel - ung | ○ | ○ | ○ | ○ | ○ | ○ |
| far - mann | ○ | ○ | ○ | ○ | ○ | ○ |

# Leksikale relasjoner (6 poeng)

## Leksikale relasjoner (6 poeng)

Hvilken semantisk relasjon holder mellom følgende ord-par:

| | Synonymi | Antonymi | Hyponymi | Hypernymi | Meronymi | Homonymi |
|---|---|---|---|---|---|---|
| sykkel - terrengsykkel | ○ | ○ | ○ | ● ✓ | ○ | ○ |
| rask – hurtig | ● ✓ | ○ | ○ | ○ | ○ | ○ |
| bryter (idrettsutøver) - bryter (lysbryter) | ○ | ○ | ○ | ○ | ○ | ● ✓ |
| medlem - forening | ○ | ○ | ○ | ○ | ● ✓ | ○ |
| gammel - ung | ○ | ● ✓ | ○ | ○ | ○ | ○ |
| far - mann | ○ | ○ | ● ✓ | ○ | ○ | ○ |

# Leksikale relasjoner - Husk!

- Homonymi: urelaterte betydninger av samme fonologiske ord (homograf/homofon)1. f.eks.lap “circuit of course” og lap “part of body when sitting down”.
- Hyponymi: inkluderingsrelasjon – et hyponym inkluderer betydningen til et mer generelt ord. f.eks. hund, katter hyponymer av dyr.
- Hypernymy: motsatt av hyponymi der det mer generelle ordet kalles hypernym. f.eks dyr er hypernym av hund, katter.
- Meronymi: relasjon mellom del og helhet. `X er del av Y`, `Y har X`. f.eks. et omslag er del av en bok, eller en bok har et omslag.

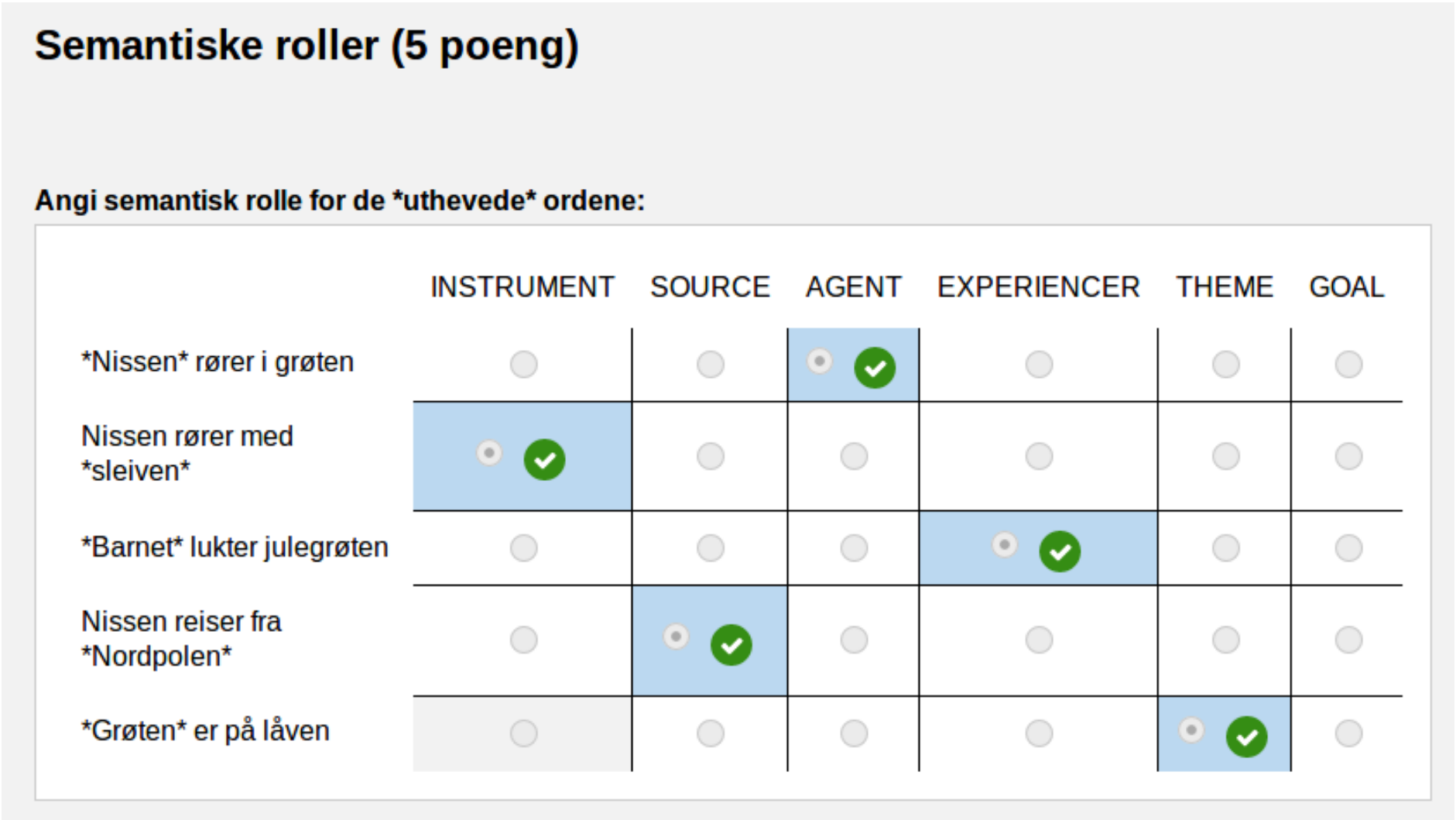
# Semantiske roller (5 poeng)

## Semantiske roller (5 poeng)

**Angi semantisk rolle for de *uthevede* ordene:**

| | INSTRUMENT | SOURCE | AGENT | EXPERIENCER | THEME | GOAL |
|---|---|---|---|---|---|---|
| *Nissen* rører i grøten | ○ | ○ | ○ | ○ | ○ | ○ |
| Nissen rører med *sleiven* | ○ | ○ | ○ | ○ | ○ | ○ |
| *Barnet* lukter julegrøten | ○ | ○ | ○ | ○ | ○ | ○ |
| Nissen reiser fra *Nordpolen* | ○ | ○ | ○ | ○ | ○ | ○ |
| *Grøten* er på låven | ○ | ○ | ○ | ○ | ○ | ○ |

# Semantiske roller (5 poeng)



# Semantiske roller - Husk!

- `agent`: den som setter i gang en handling, i stand til å handle medviten og vilje ("volition").
- `patient`: entiteten som påvirkes av en handling, gjennomgår ofte en forandring.
- `theme`: entiteten som blir beveget av en handling eller hvis beliggenhet beskrives.
- `experiencer`: er bevisst på handlingen eller tilstanden, men er ikke i kontroll over den.
- `beneficiary`: entiteten som en handling utføres for.
- `instrument`: middelet som gjør at en handling kan utføres eller finner sted.
- `goal`: entiteten som noe beveger seg mot (bokst./fig.).
- `source`: entiteten som noe beveger seg fra (bokst./fig.).

# Bayes regel (8 poeng)

Vis hvordan Bayes regel blir utledet fra følgende betinget sannsylighet:

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

# Bayes regel (8 poeng)

Vis hvordan Bayes regel blir utledet fra følgende betinget sannsylighet:

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

**Løsningsforslag**

Betinget sannsynlighet: $P(A|B) = \frac{P(A,B)}{P(B)}$

Produktsetningen: $P(A,B) = P(A|B)P(B) = P(B|A)P(A)$

Bayes regel: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

**Poengfordeling:**

- Produktsetningen: 4 poeng om svaret er identisk til løsningsforslaget. Om kun en del av svaret er riktig, bør det gis 2 poeng.
- Bayes regel: 4 poeng om svaret er identisk til løsningsforslaget. Om kun en del av svaret er riktig, bør det gis 2 poeng.

# Setningssemantikk (8 poeng)

Hva vil det si at to setninger står i en entailment-relasjon? Illustrér svaret ditt med minst ett eksempel.

# Setningssemantikk (8 poeng)

Hva vil det si at to setninger står i en entailment-relasjon? Illustrér svaret ditt med minst ett eksempel.

**Løsningsforslag:**

En entailment-relasjon mellom to setninger innebærer at den ene setningen medfører den andre. Den er gitt ved lingvistisk informasjon (enten leksikal eller syntaktisk) og trenger ikke å avgjøres empirisk (ved å sjekke fakta i verden).

Eksempel leksikal entailment:

- Terroristen myrdet statsministeren
- Statsministeren er død

Eksempel syntaktisk entailment:

- Egypterne bygget pyramidene
- Pyramidene ble bygget av egypterne

Entailment kan også defineres logisk ved at: En setning p medfører ("entails") en annen setning q dersom det er slik at når p er sann er q sann og når q er usann så er p usann

**Poengfordeling:**

- Det viktigste her er at de har med at det er en relasjon mellom setninger der den ene medfører den andre og at den er gitt lingvistisk (dvs trenger ikke å sjekke i verden). Trekker 2 poeng dersom en av disse ikke er med.
- Trekker 3 poeng dersom det ikke er gitt noe eksempel og/eller eksempelet er feil (dvs ikke angir entailment).
- Trekker 1 poeng dersom den logiske definisjonen ikke er gitt.

# Named Entity Recognition (NER) (10 poeng)

Den vanligste måten å løse NER oppgaven er ved ord-for-ord klassifisering, såkalt **BIO-klassifisering**, der ordene representeres ved trekk (features).

1. Forklar kort hva **BIO** står for.

2. Gi eksempler på 4 typer trekk som kan brukes for å løse denne oppgaven.

# Named Entity Recognition (NER) (10 poeng)

1. BIO-klassifisering: taggen indikerer om ordet befinner seg i begynnelsen (B), innenfor (I) eller utenfor (O) et egennavn, samt indikerer kategori.
2. Trekk kan være:
   - ordform (tokenisering): of, George, Washington, led
   - lemma: of, George, Washington, lead
   - shape: lower, capital, capital, lower
   - affikser: of, rge, ton, ead
   - ordklasse: IN, NNP, NNP, VBD
   - chunk-kategori: PP, NP,
   - navneliste: 0, 1, 1, 0

**Poengfordeling:**

1. Maks 6 poeng. 2 poeng for hver riktig beskrivelse av hva B, I, og O står for.
2. Maks 4 poeng, 1 poeng per riktig svar. Trekker 1 poeng for hvert feil/manglende svar.

# Ekstra Naive Bayes eksempel

| | Cat | Document |
|---|---|---|
| Training | − | just plain boring |
| | − | entirely predictable and lacks energy |
| | − | no surprises and very few laughs |
| | + | very powerful |
| | + | the most fun film of the summer |
| Test | ? | predictable with no fun |

# Ekstra Naive Bayes eksempel

- Begynne med å regne ut prior-sannsynligheten for klassene $P(-)$ og $P(+)$.

# Ekstra Naive Bayes eksempel

- Begynne med å regne ut prior-sannsynligheten for klassene $P(-)$ og $P(+)$.
- Husk at vi beregner $P(c)$ slik:

$$P(c) = \frac{N_c}{N_{doc}}$$

$N_c$ = antall dokumenter i treningsdataen som er i klassen c
$N_{doc}$ = total antall dokumenter

- $P(-)$ og $P(+)$ er da:

# Ekstra Naive Bayes eksempel

- Begynne med å regne ut prior-sannsynligheten for klassene $P(-)$ og $P(+)$.
- Husk at vi beregner $P(c)$ slik:

$$P(c) = \frac{N_c}{N_{doc}}$$

$N_c$ = antall dokumenter i treningsdataen som er i klassen c
$N_{doc}$ = total antall dokumenter

- $P(-)$ og $P(+)$ er da:

$$P(-) = \frac{3}{5} \qquad P(+) = \frac{2}{5}$$

# Eksempel – NB for sentiment klassifisering forts.

- Vi beregner sannsynlighetene for hvert ord i testsetningen, altså: “predictable”, “no”, “fun”. Vi bruker smoothing/glatting. Vi tar ikke hensyn til “with” da denne ikke finnes i vårt vokabular (treningsdataen).

- Vi må da beregne sannsynligheten for individuelle trekk $P(f_j|c)$ altså :

$$P(w_i|c) = \frac{count(w_i, c) + 1}{(\sum_{w \in V} (count(w, c)) + |V|}$$

# Eksempel – NB for sentiment klassifisering forts.

$$P(\text{predictable}|-) = \frac{1+1}{14+20}$$

$$P(\text{no}|-) = \frac{1+1}{14+20}$$

$$P(\text{fun}|-) = \frac{0+1}{14+20}$$

$$P(\text{predictable}|+) = \frac{0+1}{9+20}$$

$$P(\text{no}|+) = \frac{0+1}{9+20}$$

$$P(\text{fun}|+) = \frac{1+1}{9+20}$$

# Eksempel – NB for sentiment klassifisering forts.

$$P(-)P(S|-) = \frac{3}{5} \times \frac{\overbrace{2 \times 2 \times 1}^{4}}{34^3} = 0,000061062 = 6.10 \times 10^{-5}$$

$$P(+)P(S|+) = \frac{2}{5} \times \frac{\overbrace{1 \times 1 \times 2}^{2}}{29^3} = 0,000032802 = 3.28 \times 10^{-5}$$

$$P(-)P(S|-) > P(+)P(S|+)$$

$\Longrightarrow$ Den blir derfor klassifisert som *negativ*.

# Eksempel – NB for sentiment klassifisering forts.

- Hva om vi ikke bruker smoothing?

# Eksempel – NB for sentiment klassifisering forts.

- Hva om vi ikke bruker smoothing?
- Vi beregner sannsynlighetene for hvert ord i testsetningen, altså: “predictable”, “no”, “fun”. Vi tar ikke hensyn til “with” da denne ikke finnes i vårt vokabular (treningsdataen).
- Vi må da beregne sannsynligheten for individuelle trekk $P(f_j|c)$ altså :

$$P(w_i|c) = \frac{count(w_i, c)}{\sum_{w \in V} count(w, c)}$$

# Eksempel – NB for sentiment klassifisering forts.

$$P(\text{predictable}|-) = \frac{1}{14}$$

$$P(\text{predictable}|+) = \frac{0}{9}$$

$$P(\text{no}|-) = \frac{1}{14}$$

$$P(\text{no}|+) = \frac{0}{9}$$

$$P(\text{fun}|-) = \frac{0}{14}$$

$$P(\text{fun}|+) = \frac{1}{9}$$

# Eksempel – NB for sentiment klassifisering forts.

$$P(-)P(S|-) = \frac{3}{5} \times \frac{\overbrace{1 \times 1 \times 0}^{0}}{14^3} = 0$$

$$P(+)P(S|+) = \frac{2}{5} \times \frac{\overbrace{0 \times 0 \times 1}^{0}}{9^3} = 0$$

Problematisk!

# Ekstra Naive Bayes eksempel 2

| | Cat | Document |
|---|---|---|
| Training | $-$ | just plain boring |
| | $-$ | entirely predictable and lacks energy |
| | $-$ | no surprises and almost no fun |
| | $+$ | very powerful |
| | $+$ | the most fun film of the summer |
| Test | ? | predictable with no fun |

# Eksempel – NB for sentiment klassifisering forts.

- Hva om vi ikke bruker smoothing?

# Eksempel – NB for sentiment klassifisering forts.

- Hva om vi ikke bruker smoothing?
- Vi beregner sannsynlighetene for hvert ord i testsetningen, altså: “predictable”, “no”, “fun”. Vi tar ikke hensyn til “with” da denne ikke finnes i vårt vokabular (treningsdataen).
- Vi må da beregne sannsynligheten for individuelle trekk $P(f_j|c)$ altså :

$$P(w_i|c) = \frac{count(w_i, c)}{\sum_{w \in V} (count(w, c)}$$

# Eksempel – NB for sentiment klassifisering forts.

$$P(\text{predictable}|-) = \frac{1}{14}$$

$$P(\text{predictable}|+) = \frac{0}{9}$$

$$P(\text{no}|-) = \frac{2}{14}$$

$$P(\text{no}|+) = \frac{0}{9}$$

$$P(\text{fun}|-) = \frac{1}{14}$$

$$P(\text{fun}|+) = \frac{1}{9}$$

# Eksempel – NB for sentiment klassifisering forts.

$$P(-)P(S|-) = \frac{3}{5} \times \frac{\overbrace{1 \times 2 \times 1}^{2}}{14^3} = 0,000437318 = 4.37 \times 10^{-4}$$

$$P(+)P(S|+) = \frac{2}{5} \times \frac{\overbrace{0 \times 0 \times 1}^{0}}{9^3} = 0$$

$$P(-)P(S|-) > P(+)P(S|+)$$

$\Longrightarrow$ Den blir derfor klassifisert som *negativ*.