Syntaks og klassifisering

Løsningsforslag

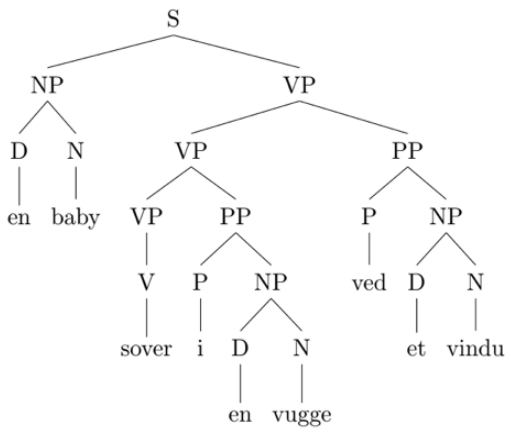
1 Strukturell flertydighet (eksamen V2017)

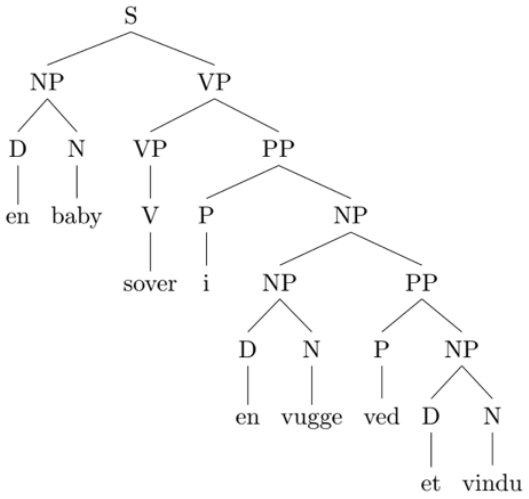
1.

a) Grammatikken tildeler 2 analyser til setningen “en baby sover i en vugge ved et vindu” (se trærne nedenfor).

b) Grammatikken tildeler 1 analyse til setningen “et barn i en vugge sover”.

2.





3. For at grammatikken skal tillate vilkårlig mange adjektiver må vi utvide den med følgende regler:

```
NP -> D N'
N' -> N
N' -> Adj N'
```

(Taggene `D` og `Det` er hyppig brukt om hverandre. De er begge forkortelser for “determinativ”.)

4. Den utvidede grammatikken tildeler 1 analyse til setningen “en rund liten søt baby sover”.

2 En grammatikk for norsk

I separat fil.

3 Naive Bayes-klassifisering

3.1 Sannsynlighet for en setning i en kategori

Setningen er "*Jeg liker alltid utenlandske filmer*"

Sannsynligheten for POS

$P(POS) = 1/2 = 0.5$

$P(setning|POS) = P(POS) * P(jeg|POS) * P(liker|POS) * P(alltid|POS) * P(utenlandske|POS) * P(filmer|POS)$

$= 0.5 * 0.09 * 0.07 * 0.29 * 0.04 * 0.08$

$P(setning|POS) = 0.0000029232$

Sannsynligheten for NEG

$P(NEG) = 1/2 = 0.5$

$P(setning|NEG) = P(NEG) * P(jeg|NEG) * P(liker|NEG) * P(alltid|NEG) * P(utenlandske|NEG) * P(filmer|NEG)$

$= 0.5 * 0.16 * 0.06 * 0.06 * 0.15 * 0.11$

$P(setning|NEG) = 0.000004752$

$P(setning|POS) < P(setning|NEG)$, så derfor vil Naive Bayes klassifisere setningen "Jeg liker alltid utenlandske filmer" som **negativ**.

3.2 Sannsynlighet for at et dokument er i en kategori

Dokument å klassifisere: *rask, par, skyte, fly*

Ordforråd |V| = 7 (gitt av antall unike ord, altså typer, i alle taggede dokumenter)
Antall tokens i *komedie*: 9
Antall tokens i *action*: 11

Komedie
Sannsynligheter for ord gitt *komedie*:
P(rask | komedie) = *(1 + 1) / (9 + 7) = 0,125*
P(par | komedie) = *(2 + 1) / (9 + 7) = 0,1875*
P(skyte | komedie) = *(0 + 1) / (9 + 7) = 0,0625*
P(fly | komedie) = *(1 + 1) / (9 + 7) = 0,125*

Prior-sannsynligheten for *komedie*:
P(komedie) = *2/5*

*(2 / 5) * 0,125 * 0,1875 * 0,0625 * 0,125 = 0.000073*

Action
Sannsynligheter for ord gitt *action*:
P(rask | action) = *(2 + 1) / (11 + 7) = 0,1667*
P(par | action) = *(0 + 1) / (11 + 7) = 0,0556*
P(skyte | action) = *(4 + 1) / (11 + 7) = 0,2778*
P(fly | action) = *(1 + 1) / (11 + 7) = 0,1111*

Prior-sannsynligheten for *action:*
P(action) = *3/5*

*(3 / 5) * 0,1667 * 0,0556 * 0,2778 * 0,1111 = 0.00017*

Dokumentet klassifiseres som **action** fordi *0.000073 < 0.00017*.