Oppgave 5, Syntaks og klassifisering (løsningsforslag)

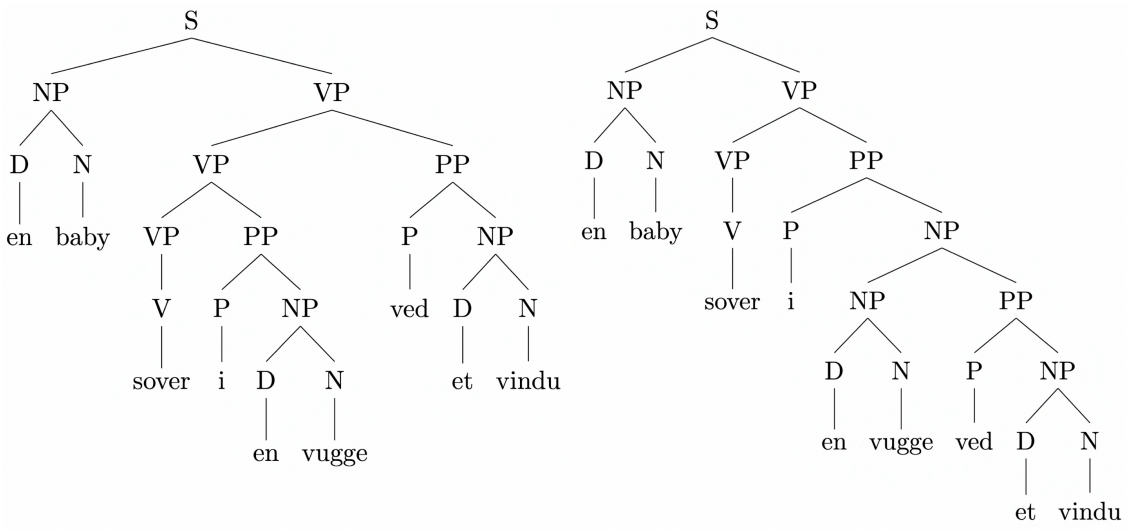
October 27, 2022

1 Strukturell flertydighet

1.

Grammatikken tildeler to analyser til setning **(a)** (se trær under) og én analyse til setning **(b)**.

2.



3.

For at grammatikken skal tillate vilkårlig mange adjektiver, kan vi utvide den med følgende regler:

NP $\rightarrow$ *D N'*

N' $\rightarrow$ *N*

N' $\rightarrow$ *Adj N'*

4.

Den utvidede grammatikken tildeler én analyse til setningen "en rund liten søt baby sover".

2 En gammatikk for norsk

1.

[2]:
```
import nltk

versjon1 = nltk.CFG.fromstring("""
        S -> NP VP

        NP -> "Per" | "Kari" | "Ola" | "middag" | "boka"
        NP -> D N

        D -> "en"
        N -> "bok"

        PP -> P NP
        P -> "til"

        VP -> V
        VP -> V NP
        VP -> V NP PP
        VP -> V NP NP

        V -> "gir"
        V -> "sover"
        V -> "spiser"
        V -> "finner"
    """)

setn = ["Per gir en bok til Kari",
        "Kari gir Per boka",
        "Ola sover",
        "Kari spiser",
        "Kari spiser middag",
        "Per finner boka"]

parser1 = nltk.RecursiveDescentParser(versjon1)

print("2.1 - Parser for grammatiske setninger:")
for s in setn:
    print("-", s)
    for parse in parser1.parse(s.split()):
        print("  ", parse)
```

```
2.1 - Parser for grammatiske setninger:
- Per gir en bok til Kari
   (S (NP Per) (VP (V gir) (NP (D en) (N bok)) (PP (P til) (NP Kari))))
- Kari gir Per boka
```

```
   (S (NP Kari) (VP (V gir) (NP Per) (NP boka)))
- Ola sover
   (S (NP Ola) (VP (V sover)))
- Kari spiser
   (S (NP Kari) (VP (V spiser)))
- Kari spiser middag
   (S (NP Kari) (VP (V spiser) (NP middag)))
- Per finner boka
   (S (NP Per) (VP (V finner) (NP boka)))
```

2.

[3]:
```
feil = ["Kari sover boka",
        "Ola finner"]

print("\n2.2 - Parser for ugrammatiske setninger:")
for s in feil:
    print("-", s)
    for parse in parser1.parse(s.split()):
        print("  ", parse)
```

```
2.2 - Parser for ugrammatiske setninger:
- Kari sover boka
   (S (NP Kari) (VP (V sover) (NP boka)))
- Ola finner
   (S (NP Ola) (VP (V finner)))
```

3.

Problemet er at VP-regelen ikke holder styr på hva slags argumenter verbene krever. Vi deler verbene inn i fire kategorier: intransitive verb som ikke tar objekt (f.eks. “sover”), strengt transitive verb som tar objekt (f.eks. “finner”), ditransitive verb som tar to objekter (f.eks. “gir”), og valgfritt transitive verb som *kan* ta objekt (f.eks. “spiser”). De forskjellige VP-konstruksjonene krever så de riktige typene verb til hver konstruksjon slik av vi utelukker de ugrammatiske variantene.

[4]:
```
versjon2 = nltk.CFG.fromstring("""
        S -> NP VP

        NP -> "Per" | "Kari" | "Ola" | "middag" | "boka"
        NP -> D N

        D -> "en"
        N -> "bok"

        PP -> P NP
        P -> "til"

        VP -> Vintrans | Vtrans-opt
```

```
        VP -> Vtrans NP
        VP -> Vditrans NP PP
        VP -> Vditrans NP NP

        Vtrans -> Vtrans-opt | Vtrans-strict
        Vditrans -> "gir"
        Vintrans -> "sover"
        Vtrans-opt -> "spiser"
        Vtrans-strict -> "finner"
    """)

parser2 = nltk.RecursiveDescentParser(versjon2)

print("\n2.3 - Parser for grammatiske setninger:")
for s in setn:
    print("-", s)
    for parse in parser2.parse(s.split()):
        print("  ", parse)

print("\n2.3 - Parser for ugrammatiske setninger:")
for s in feil:
    print("-", s)
    for parse in parser2.parse(s.split()):
        print("  ", parse)
```

```
2.3 - Parser for grammatiske setninger:
- Per gir en bok til Kari
   (S
  (NP Per)
  (VP (Vditrans gir) (NP (D en) (N bok)) (PP (P til) (NP Kari))))
- Kari gir Per boka
   (S (NP Kari) (VP (Vditrans gir) (NP Per) (NP boka)))
- Ola sover
   (S (NP Ola) (VP (Vintrans sover)))
- Kari spiser
   (S (NP Kari) (VP (Vtrans-opt spiser)))
- Kari spiser middag
   (S (NP Kari) (VP (Vtrans (Vtrans-opt spiser)) (NP middag)))
- Per finner boka
   (S (NP Per) (VP (Vtrans (Vtrans-strict finner)) (NP boka)))

2.3 - Parser for ugrammatiske setninger:
- Kari sover boka
- Ola finner
```

Som vi ser, skrives det ikke ut noen analyser til de ugrammatiske setningene.

3 Naive Bayes-klassifisering

3.1 Sannsynlighet for en setning i en kategori

Setninga er *"Jeg liker alltid utenlandske filmer"*

Sannsynligheten for POS

Prior-sannsynligheten for POS $P(POS) = \frac{1}{2}$.

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

$P(setning|POS) = 0.5 \cdot 0.09 \cdot 0.07 \cdot 0.29 \cdot 0.04 \cdot 0.08 = 0.0000029232$

Sannsynligheten for NEG

Prior-sannsynligheten for NEG $P(NEG) = \frac{1}{2}$.

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

$P(setning|NEG) = 0.5 \cdot 0.16 \cdot 0.06 \cdot 0.06 \cdot 0.15 \cdot 0.11 = 0.000004752$

Sammenlikning

$P(setning|POS) < P(setning|NEG)$, 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 ordtyper i de taggede dokumentene.

Antall token i *komedie*: 9

Antall token i *action*: 11

Komedie

Sannsynligheten for hvert ord gitt *komedie*:

$P(rask|komedie) = \frac{1+1}{9+7}$

$P(par|komedie) = \frac{2+1}{9+7}$

$P(skyte|komedie) = \frac{0+1}{9+7}$

$P(fly|komedie) = \frac{1+1}{9+7}$

Prior-sannsynligheten for *komedie* $P(komedie) = \frac{2}{5}$.

$P(dokument|komedie) = \frac{2}{5} \cdot \frac{1+1}{9+7} \cdot \frac{2+1}{9+7} \cdot \frac{0+1}{9+7} \cdot \frac{1+1}{9+7} = 0.000073$

Action

Sannsynligheten for hvert ord gitt *action*:

$P(rask|action) = \frac{2+1}{11+7}$

$P(par|action) = \frac{0+1}{11+7}$

$P(skyte|action) = \frac{4+1}{11+7}$

$P(fly|action) = \frac{1+1}{11+7}$

Prior-sannsynligheten for *action* $P(action) = \frac{3}{5}$.

$P(dokument|action) = \frac{3}{5} \cdot \frac{2+1}{11+7} \cdot \frac{0+1}{11+7} \cdot \frac{4+1}{11+7} \cdot \frac{1+1}{11+7} = 0.00017$

Sammenlikning

Dokumentet klassifiseres som ***action*** fordi $P(dokument|komedie) < P(dokument|action)$.