IN1140 H2021 – Gruppeoppgave 5
Syntaks og klassifisering

1 Strukturell flertydighet (eksamen V2017)

Anta følgende grammatikk for et fragment av norsk:

$$
\begin{aligned}
S &\rightarrow NP\ \ VP \\
NP &\rightarrow D\ N \\
NP &\rightarrow NP\ \ PP \\
VP &\rightarrow V \mid V\ NP \\
VP &\rightarrow VP\ \ PP \\
PP &\rightarrow P\ NP
\end{aligned}
\qquad
\begin{aligned}
D &\rightarrow \text{en} \mid \text{et} \\
N &\rightarrow \text{barn} \mid \text{baby} \mid \text{vindu} \mid \text{vugge} \\
V &\rightarrow \text{sover} \\
P &\rightarrow \text{i} \mid \text{ved}
\end{aligned}
$$

1. Hvor mange analyser tildeler grammatikken til setningene under?
   (a) en baby sover i en vugge ved et vindu
   (b) et barn i en vugge sover
2. Tegn treet eller trærne grammatikken tilordner setning (a).
3. Utvid grammatikken slik at den tillater vilkårlig mange adjektiver foran substantivet i substantivfrasen og godtar setninger som:
   (a) en liten søt baby sover i en stor vugge
4. Hvor mange analyser tilordner den utvidede grammatikken din til setningen «en rund liten søt baby sover»?

2 En grammatikk for norsk

NLTK inneholder flere forskjellige parsere som tildeler syntaktisk struktur til en setning automatisk, i henhold til en grammatikk. Her skal du bruke RecursiveDescent-parseren som står beskrevet i seksjon 8.3 i NLTK-boka. Du kan formulere grammatikken din direkte som en streng, slik:

```
grammar = nltk.CFG.fromstring("""
  S -> NP VP
  VP -> V NP | V NP PP
  PP -> P NP
  V -> "saw" | "ate" | "walked"
  NP -> "John" | "Mary" | "Bob" | Det N | Det N PP
  Det -> "a" | "an" | "the" | "my"
  N -> "man" | "dog" | "cat" | "telescope" | "park"
  P -> "in" | "on" | "by" | "with"
  """)
```

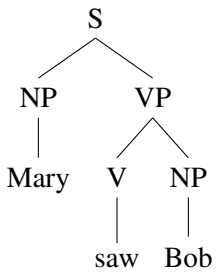
Merk at `RecursiveDescent`-parseren ikke håndterer venstre-rekursjon, av typen `VP -> VP PP`, så du må formulere grammatikken uten denne formen for rekursjon. Du kan teste grammatikken på en setning slik:

```
sent = "Mary saw Bob".split()
rd_parser = nltk.RecursiveDescentParser(grammar)
for tree in rd_parser.parse(sent):
    print(tree)
```

Parseren skriver da ut et tre i klammenotasjon:

```
(S (NP Mary) (VP (V saw) (NP Bob)))
```

Dette tilsvarer dette syntaktiske treet:



1. Du skal nå implementere en kontekstfri grammatikk med denne parseren som analyserer et fragment av norsk slik at setningene under gis riktig analyse:

   (a) `(S (NP Per) (VP (V gir) (NP (Det en) (N bok)) (PP (P til) (NP Kari))))`

   (b) `(S (NP Kari) (VP (V gir) (NP Per) (NP boka)))`

   (c) `(S (NP Ola) (VP (V sover)))`

   (d) `(S (NP Kari) (VP (V spiser)))`

   (e) `(S (NP Kari) (VP (V spiser) (NP middag)))`

   (f) `(S (NP Per) (VP (V finner) (NP boka)))`

   Vis at setningene i a-f gis korrekt analyse ved å parse dem med grammatikken og skrive ut analysen som beskrevet over.

2. Grammatikken slik den er implementert er imidlertid ikke tilfredsstillende, siden den for eksempel vil godta setninger som *Kari sover boka* og *Ola finner*. Verifiser dette ved å skrive ut analysene grammatikken din tildeler disse setningene.

3. Du skal nå skrive en ny og forbedret versjon av grammatikken slik at de grammatiske konstruksjonene i 1-6 tillates, men ugrammatiske konstruksjoner (som *Kari sover boka* og *Ola finner*) er utelukket. Skriv ut analysene den nye og forbedrede grammatikken tildeler de grammatiske setningene 1-6, og vis videre at de ugrammatiske setningene ikke tildeles noen analyse.

3 Naive Bayes-klassifisering

I oppgavene under skal vi bruke Naive bayes-formelen:

$$\hat{b} = argmax_{b \in B} \ P(b) \prod_{j=1}^{n} P(v_j|b)$$

Altså: Den estimerte kategorien for et dokument, er lik den kategorien som får høyest verdi når vi gjør følgende: Vi ganger sannsynligheten for kategorien med produktet av sannsynlighetene for alle ordene i dokumentet vi vil klassifisere. Vi beregner sannsynlighetene som følger:
Sannsynligheten for en kategori er lik antall ganger kategorien forekommer i korpuset, delt på antall kategorier totalt. Sannsynligheten for et ord er lik antall ganger ordet forekommer i en gitt kategori, delt på antall ord i den kategorien (tokens).

Hvis vi bruker legg-til-én-glatting, plusser vi på én over brøkstreken, og plusser på lengden av hele ordforrådet (types). Merk at et dokument kan være så kort eller langt vi vil, selv kun én setning.

3.1 Sannsynlighet for en setning i en kategori

Vi har et lite korpus med filmanmeldelser, og vil regne ut sannsynligheten for at en ny setning er del av en positiv anmeldelse eller en negativ anmeldelse. Under er sannsynlighetene for hvert ord gitt kategorien `positiv` eller `negativ`. Sannsynlighetene for ordene er allerede regnet ut. Vi antar at det er like mange positive klasser som negative, altså er sannsynligheten for hver av klassene lik.

ord	**POS**	**NEG**
Jeg	0.09	0.16
liker	0.07	0.06
alltid	0.29	0.06
utenlandske	0.04	0.15
filmer	0.08	0.11

Hvilken klasse vil Naive bayes gi til setningen 'Jeg liker alltid utenlandske filmer'?

3.2 Sannsynlighet for at et dokument er i en kategori

I denne oppgaven har vi noen dokumenter med ord, som hører til sjangerne `action` eller `komedie`. Gitt et nytt dokument D, bruk formelen under til å klassifisere det nye dokumentet.

$$\hat{b} = argmax_{b \in B} \ P(b) \prod_{j=1}^{n} P(v_j|b)$$

1. moro, par, kjærlighet, kjærlighet	**komedie**
2. rask, rasende, skyte	**action**
3. par, fly, rask, moro,moro	**komedie**
4. rasende, skyte, skyte, moro	**action**
5. fly, rask, skyte, kjærlighet	**action**

Dokumentet vi vil klassifisere inneholder følgende ord: **rask, par, skyte, fly**

Du må altså regne ut sannsynlighetene for de forskjellige ordene selv, men du trenger bare å regne ut for ordene i det nye dokumentet. **Bruk legg-til-én-glatting** (Laplace). Regn så ut hvilken verdi som er størst. Blir dokumentet klassifisert som `action` eller `komedie`?