INF1820 V2017 – Oblig 3b
CFGer og semantikk

Innleveringsfrist: fredag 5 mai

Lever inn svarene dine i Devilry (`https://devilry.ifi.uio.no/`) i filer som angir brukernavnet ditt, slik: `oblig3b_brukernavn.py`. Pass på at filen din kan kjøres som et program; det skal ikke være en REPL-sesjon limt inn i en fil.

En perfekt besvarelse på denne oppgaven er verdt 100 poeng.

1 En grammatikk for norsk (50 poeng)

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. 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"
  """)
```

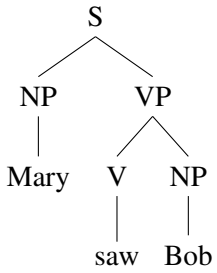
Merk at `RecursiveDescent`-parseren ikke håndterer venstre-rekursjon, av typen `VP -> VP PP`, så du må formulere grammatikken uten denne formen for rekursjon. Du kan teste grammatikken på en setning slik:

```
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 (D 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 1-6 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.

2 Manuell annotering av ordbetydning (20 poeng)

I denne oppgaven skal du gjøre en manuell annotering av ordbetydning og kommentere obervasjonene dine. Skriv svarene dine som utkommentert tekst i Python-filen din.

Setningene i (1)-(5) under er hentet fra SemCor-korpuset, et korpus som er annotert med ordbetydning, og alle inneholder verbet *leave*.

1. But questions with which committee members taunted bankers appearing as witnesses left little doubt that they will recommend passage of it .

2. The departure of the Giants and the Dodgers to California left New York with only the Yankees .

3. After the coach listed all the boy 's faults , Hartweger said , " Coach before I leave here , you 'll get to like me " .

4. R. H. S. Crossman , M.P. , writing in The Manchester Guardian , states that departures from West Berlin are now running at the rate not of 700 , but of 1700 a week , and applications to leave have risen to 1900 a week .

5. The house has been swept so clean that contemporary man has been left with no means , or at best with wholly inadequate means , for dealing with his experience of spirit .

Slå opp verbet *leave* i WordNet (bruk *Use Word-Net Online*: `http://wordnetweb.princeton.edu/perl/webwn`). Du skal ikke ta hensyn til betydningene for substantivet *leave*.

For hver av setningene i (1)-(5) skal du velge en betydning ("sense") fra WordNet for verbet *leave* i setningen og notere valget ditt. På *Use Word-Net online*-siden kan du klikke på Display Options og velge Show Sense Numbers for å få en nummerert oversikt over de forskjellige betydningene. Bruk disse nummerene i svaret ditt.

Videre skal du reflektere rundt arbeidet ditt og besvare følgende spørsmål:

- Hvilke setninger var det vanskelig å annotere og hvorfor?
- Hvilke par (eller grupperinger) av WordNet-betydninger var det vanskelig å skille fra hverandre og hvilke kriterier brukte du for å skille mellom dem?

3 Betydningsdisambiguering (WSD) med en Naive Bayesklassifiserer (30 poeng)

Begynn med å gjøre deg kjent med Naive Bayes-klassifisering, dersom du ikke allerede er det (se lysark fra forelesning og/eller les Jurafsky & Martin 20.2.)

Filen `wsd_tren.txt` inneholder (fiktive) data annotert med ordbetydning for lemmaet *skim*. Hver linje inneholder en liste med trekk og en kategori. Elementene i hver linje er adskilt med mellomrom. Første element i hver linje er kategorien og de andre elementene er trekk. Første linje ser slik ut:

```
Reading book day novel
```

Dette betyr at betydningskategorien for denne instansen er `Reading` og inneholder trekkene `book`, `day` og `novel`.

1. Bruk treningsdataene i `wsd_tren.txt` til å beregne sannsynligheten for de ulike betydningene i dataene. Hvor sannsynlig er betydningen `Removing`? Bruk Python til å utføre beregningene dine.
2. Ett av trekkene som forekommer i treningsfilen er `day`. Beregn sannsynligheten for dette trekket, gitt `Reading`-betydningen, dvs $P(day|Reading)$. Bruk Python til å utføre beregningene dine.
3. Filen `wsd_test.txt` inneholder en testinstans på samme format som treningsdataene, bortsett fra at kategorien er ukjent:

   ```
   ?  paper surface towards
   ```

   Bruk Naive Bayes-formelen for å beregne den mest sannsynlige betydningen for denne testinstansen. Bruk Python til å utføre beregningene dine. Husk at i Naive Bayes er den mest sannsynlige betydningen $\hat{s}$ gitt ved:

$$\hat{s} = \underset{s \in S}{\operatorname{argmax}} P(s) \prod_{i=0}^{n} P(f_i|s) \tag{1}$$