# INF1820, V2012 – Obligatorisk oppgave 3b

## Innlevering: 4/5

Registrer svarene dine i en fil som angir brukernavnet ditt slik:

`oblig3a_brukernavn.py`

Som vanlig skjer innlevering av oppgaven i Devilry. Se emnesiden for mer informasjon om reglement rundt innlevering, samt bruk av Devilry.

En perfekt løsning av denne oppgaven er verdt 100 poeng.

1. **Et grammatikkfragment for engelsk (40 poeng)**
   Engelsk har samsvarsbøyning mellom subjekt og verb: subjekter i tredje person entall krever en spesiell form av verbet (*she likes, it likes* versus *I like, they like*). Engelsk har også, som norsk, noe kasusmarkering for pronomen. For eksempel, *I, he, they* er i nominativ kasus, mens *me, him, them* er i akkusativt kasus. I denne oppgaven skal du skrive en kontekstfri grammatikk som aksepterer setningene i (a), men ikke setningene i (b):

   (a) Setninger grammatikken skal akseptere:
   - i. she likes him
   - ii. they like him
   - iii. they like the actress
   - iv. they like the actresses
   - v. I know her
   - vi. I know the actor whom she likes
   - vii. I know the actress who likes him
   - viii. the actress who likes him sings
   - ix. the actresses who like him sing

   (b) Setninger grammatikken *ikke* skal akseptere:
   - i. *she likes he
   - ii. *they likes him
   - iii. *me know her
   - iv. *I know the actress whom likes him
   - v. *the actress who likes him sing
   - vi. *the actress who like him sings
   - vii. *the actresses who likes him sing
   - viii. *the actresses who like him sings

   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.parse_cfg("""
```

```
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 deg at flere alternativer for samme kategori skal formuleres med disjunksjon `VP -> V NP | V NP PP`, dvs. at `VP` enten kan bestå av en `V NP` eller `V NP PP`. Merk deg videre at `RecursiveDescent`-parseren ikke håndterer venstre-rekursjon av typen `VP -> VP PP`, så du må formulere grammatikken din uten denne formen for rekursjon.

Du kan teste grammatikken din på en setning slik:

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

Parseren skriver da ut et tre i klammenotasjon:

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

Implementer grammatikkfragmentet for engelsk som beskrevet over, test grammatikken din på setningene i (a) og (b) ovenfor og skriv ut resultatet. Pass på at grammatikken din faktisk generer et tre for alle setningene i (a) og ikke genererer noe tre for setningene i (b).

2. **Manuell annotering av ordbetydning (25 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 `oblig3a_brukernavn.py`.

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

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

   (b) The departure of the Giants and the Dodgers to California **left** New York with only the Yankees .

   (c) After the coach listed all the boy s faults , Hartweger said , Coach before I **leave** here , you ll get to like me .

   (d) 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 .

   (e) 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 .

   (f) A second and also good practice is to shear off the tops , **leaving** an inch high stub with just a leaf or two on each branch .

   (g) No doubt some experiences vanish so completely as to **leave** no trace on the sleeper s mind .

(h) He is a widower , his three children are dead , he has no one **left** on earth ; also he is a drunk , and has lost his job on that account

(i) Piepsam tries to stop him by force , receives a push in the chest from Life , and is **left** standing in impotent and growing rage , while a crowd begins to gather .

(j) The audience **leaves** the play under a spell , It is the kind of spell which the exposure to spirit in its living active manifestation always evokes .

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

For hver av setningene i (a)-(j) skal du velge *en* betydning ("sense") fra WordNet for verbet *leave* i setningen og notere valget ditt. På "Use WordNet 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 Bayes-klassifiserer (35 poeng)** 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*.

(a) Bruk treningsdataene i `wsd_tren.txt` til å beregne sannsynligheten for betydningen REMOVING: beregn sannsynligheten $P(Removing)$ ved Maximum Likelihood Estimation (MLE) fra dataene. Bruk Python til å utføre beregningene dine.

(b) Ett av trekkene som forekommer i treningsfilen er *day*. Beregn sannsynligheten for dette trekket, gitt READING-betydningen, dvs $P(day|Reading)$. Bruk MLE over treningsdataene og bruk Python til å utføre beregningene dine.

(c) 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:

$$\hat{s} = \text{argmax}_{s \in S} P(s) \prod_{j=1}^{n} P(f_j|s)$$

Du kan lese mer om bruk av Naive Bayes for WSD-oppgaven i Jurafsky & Martin 20.2.