# INF1820 V2015 — Oppgave 2a
# Korpora og tagger

## Innleveringsfrist, onsdag 18. mars

Lever inn svarene dine med Devilry (`https://devilry.ifi.uio.no`) i en fil som angir brukernavnet ditt, slik: `oblig2a_brukernavn.py`

En perfekt besvarelse på denne oppgaven er verdt 100 poeng.

I denne oppgaven skal vi se på nyhetsdelen av Brown-korpuset i NLTK, og særlig den ordklassetaggede delen av korpuset. Maskinene på IFI har de nødvendige pakkene installert, på din egen kan du følge instruksene på `http://nltk.org/data`.

Du får tilgang til det ordklassetaggede korpuset slik:

```
import nltk
brown_news = nltk.corpus.brown.tagged_words(categories="news")
```

Dette gir deg en liste med tupler, der første element i paret er ordformen og det andre elementet er taggen.

En liste over taggene i Brown og hva de betyr finner du på `http://www.scs.leeds.ac.uk/amalgam/tagsets/brown.html`.

Helt til slutt, pass på å mappe alle ordene til små bokstaver i oppgavene under slik at *The* og *the* behandles som samme ord. Det kan du gjøre med metoden `lower()`, slik:

```
>>> "The".lower()
'the'
```

## 1 Ordfrekvens og taggfrekvens (30 poeng)

Ved hjelp av Python dictionaries og den innebygde funksjonen `sorted()` (altså, *uten* å bruke `nltk.FreqDist` og `nltk.ConditionalFreqDist`), finn ut hva som er den mest frekvente ordklassetaggen i nyhetsdelen av Brown og hvor mange ord forekommer kun én gang. Skriv ut resultatene.

## 2 Flertydighet

I denne oppgaven skal vi se på ord som kan ha to eller fler forskjellige tagger. Det vil si at i stedet for listen `["NP", "NN", "NN", "NN", "NP"]` vil vi ha listen `["NP", "NN"]`. Da kan du enten passe på at du bare sparer på tagger du ikke allerede har set før med et spesifikt ord, eller du kan bruke mengder[1] i stedet for lister. Igjen, denne oppgaven skal løses uten bruk av `nltk.FreqDist` og `nltk.ConditionalFreqDist`.

1. Hvor mange ord er flertydige? Det vil si, hvor mange ord forekommer med mer enn én ordklassetagg?

2. Hvilket ord har størst antall tagger, og hvor mange distinkte tagger har det?

## 3 Finne spesifikke eksempler (20 poeng)

Her ser vi nærmere på det mest tvetydige ordet fra forrige oppgave: For hver mulig tagg ordet kan ha, skriv ut en setning der ordet forekommer med den taggen.

For å gjøre dette må vi bruke korpuset på en litt annen måte enn i de to første oppgavene, siden vi der så på korpuset uten setningsgrenser. For å laste inn korpuset med setningsgrenser bruker du:

```
brown_sents = nltk.corpus.brown.tagged_sents(categories="news")
```

Variabelen `brown_sents` er da en liste med setninger, der hver setning er en liste av ord-tagg par.

## 4 Fordelingen av maskuline kontra feminine possessive pronomener (20 poeng)

Uten å bruke `nltk.FreqDist` og `nltk.ConditionalFreqDist` (igjen), finn ut hvor mange maskuline pronomener det er i Brown i forhold til hvor mange feminine pronomener det er.

For å løse denne oppgaven må du ha en liste over alle de maskuline og feminine pronomenene i Brown. Her kan du benytte deg av både ordet og taggen; for eksempel forekommer *her* både som possessivt pronomen (*That*

[1] Se avsnitt A til slutt for en kort introduksjon til mengder.

*is her house*), men ikke alltid (*I saw her last Monday*). Ta utgangspunkt i listen på `http://www.scs.leeds.ac.uk/amalgam/tagsets/brown.html` for å finne ut hvilke tagger som brukes på possessive pronomener.

# A Mengder

Mengder i Python konstrueres ved hjelp av funksjonen `set()`, og elementer legges til med metoden `add()`. Vi kan da gjøre:

```
>>> mengde = set()
>>> mengde.add("NP")
>>> mengde.add("NN")
>>> mengde.add("NN")
>>> mengde.add("NN")
>>> mengde
set(["NP", "NN"])
```

Som du ser kan vi legge til det samme elementet så mange ganger vi vil, men det vil bare forekomme én gang i mengden.