INF1820, V2014 – Obligatorisk oppgave 2a
Korpora og ordklassetagger

1 Innlevering: 28/2

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

```
oblig2a_brukernavn.py
```

Innlevering av oppgaven skjer 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.

I denne oppgaven skal vi benytte oss av nyhetsdelen av Brown-korpuset i Natural Language Toolkit (NLTK), og særlig den delen av korpuset som inneholder informasjon om ordklasser. NLTK er tilgjengelig på IFI's servere. Dersom du ønsker å laste den ned til din egen maskin, kan du følge instruksene på
`http://www.nltk.org/data`
Du får tilgang til den ordklassetaggede delen av Brown-korpuset slik:

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

Dette gir deg en liste med tupler, dvs en liste med par. Første element i paret er ordet og andre element er ordklassen. Her er de første ti ordene og deres tagger:

```
>>> brown_news[:10]
[('The', 'AT'), ('Fulton', 'NP-TL'), ('County', 'NN-TL'),
('Grand', 'JJ-TL'), ('Jury', 'NN-TL'), ('said', 'VBD'),
('Friday', 'NR'), ('an', 'AT'), ('investigation', 'NN'),
('of', 'IN')]
```

For en liste over ordklassene i dette korpuset, med eksempler, kan du ta en titt på `http://www.scs.leeds.ac.uk/amalgam/tagsets/brown.html`

Pass på å mappe alle ord til små bokstaver i oppgavene under, slik at du ikke behandler *The* og *the* som to forskjellige ord. Du kan gjøre dette med metoden `lower()`, slik:

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

2 Ordfrekvens og taggfrekvens (30 poeng)

Hva er det mest frekvente ordet og hva er den mest frekvente ordklassetaggen i nyhetsdelen av Brown-korpuset? Hvor mange ord forekommer kun én gang (såkalte *hapax legomena*)? Skriv ut resultatet av beregningene dine. *Ikke* bruk datastrukturene `nltk.FreqDist` og `nltk.ConditionalFreqDist` fra NLTK for å løse denne oppgaven. Bruk Python dictionaries og den innebygde funksjonen `sorted()`.

3 Flertydige ord (30 poeng)

(a) Hvor mange ord er flertydige, dvs at de forekommer med minst to forskjellige ordklassetagger? *Ikke* bruk datastrukturene `nltk.FreqDist` og `nltk.ConditionalFreqDist` fra NLTK for å løse denne oppgaven, men bruk Python dictionaries.

Her vil vi kun ha distinkte tagger, dvs listen `["NP","NN"]` i stedet for `["NP","NN","NN","NN","NP"]`. Du kan enten passe på at du kun sparer tagger du ikke har sett før med et spesifikt ord, eller så kan du bruke mengder i stedet for lister. En mengde er som en liste bortsett fra at den ikke inneholder duplikater. Du legger et element til en mengde med `add()` i stedet for `append()` for lister:

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

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

(b) Hvilket ord har størst antall tagger og hvor mange distinkte tagger har det? Igjen skal du ikke bruk `nltk.FreqDist` og `nltk.ConditionalFreqDist`.

4 Hente ut eksempler for spesifikke ord/tagg-par (20 poeng)

Ta for deg ordet med størst antall distinkte tagger fra oppgaven over og skriv ut et eksempel for hver tagg, dvs en setning fra korpuset som inneholder det ord/tagg-paret.

I denne oppgaven skal du lese inn Brown-korpuset på en annen måte enn tidligere. `brown_news` inneholder kun en liste med ord/tagg-par, uten setningsgrenser, men i denne oppgaven trenger du derimot setningsgrenser. Du kan lese inn setningene i Brown-korpuset som følger:

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

`brown_sents` er da en liste med setninger, der hver setning igjen er en liste med ord/tagg-par. Slik ser første setning ut:

```
>>> brown_sents[0]
[('The', 'AT'), ('Fulton', 'NP-TL'), ('County', 'NN-TL'),
('Grand', 'JJ-TL'), ('Jury', 'NN-TL'), ('said', 'VBD'),
('Friday', 'NR'), ('an', 'AT'), ('investigation', 'NN'),
('of', 'IN'), ("Atlanta's", 'NP$'), ('recent', 'JJ'),
('primary', 'NN'), ('election', 'NN'), ('produced', 'VBD'),
(``, ``), ('no', 'AT'), ('evidence', 'NN'), ("''", "''"),
('that', 'CS'), ('any', 'DTI'), ('irregularities', 'NNS'),
('took', 'VBD'), ('place', 'NN'), ('.', '.')]
```

5 Fordeling av maskuline og feminine possessive pronomener (20 poeng)

Hvordan er fordelingen av maskuline versus feminine possessive artikler og pronomen i nyhetsdelen av Brown-korpuset? Igjen, ikke bruk `nltk.FreqDist` og `nltk.ConditionalFreqDist`, men bruk Python dictionaries.

For å løse denne oppgaven må du først hente ut en komplett liste av maskuline og feminine possessive pronomen og artikler. Du kan benytte deg av både selve ordet og ordklassen i definisjonen din: *her* forekommer som possessivt pronomen, som i *That is her house*, men ikke alltid (f.eks. *I saw her last Monday*). For å finne ut hvilke ordklasser som brukes for possessive pronomen/artikler kan du ta en titt på listen over ordklassene i Brown-korpuset (`http://www.scs.leeds.ac.uk/amalgam/tagsets/brown.html`).