IN1140 H2017 – Oblig 2a
Korpora og ordklassetagger

Innleveringsfrist 05.10 kl.23.59

Lever inn svarene dine i Devilry i filer som angir brukernavnet ditt, slik: `oblig2a_brukernavn.py`.
En perfekt besvarelse på denne oppgaven er verdt 100 poeng.

1 Tagg- og ordfrekvens (50 poeng)

1.1 Del 1 – (10 poeng)

1. Hva er ordklasser?
2. Hvilken ordklasser opererer vi med på norsk?
3. Vi skiller mellom såkalte innholdsord og funksjonsord. Hva skiller de to typene fra hverandre?
4. Forklar bøyning og avledning.

1.2 Del 2 – (20 poeng)

Ta for deg setningene under og tildel ordklasser til alle ordene i alle setninger. Du skal benytte deg av ordklassene i tabell 1.

DT	determinativ
NN	substantiv
JJ	adjektiv
VB	verb
CC	konjunksjon
PR	preposisjon
PO	pronomen
RB	adverb
SB	subjunksjon

Table 1: Ordklasser

- Politiet ber fortsatt om tips etter et stort våpentyveri i Oslo sentrum
- Laks og poteter til middag
- Lars sang vakkert!

1.3 Del 3 – (20 poeng)

I denne delen av oppgaven skal vi jobbe med ordklassetaggede data fra nyhetsdelen i Brown-korpuset. Du får tilgang til korpuset i programet ditt slik:

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

Variabelen `brown` er nå en liste med tupler, der det første elementet er ordformen og det andre elementet er taggen som er blitt tilordnet ordet. En liste av taggene i Brown og hva de betyr finner du på `http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM#bc6`

Merk til slutt at det her er viktig å passe på at du ikke skiller mellom store og små bokstaver, slik at for eksempel *The* og *the* telles sammen, ikke hver for seg. Det kan du gjøre med methoden `lower()` slik:

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

Ved hjelp av Python dictionaries og sortering, finn ut følgende:

1. Hva er det mest frekvente ordet i Brown?
2. Hva er den mest frekvente ordklassen i Brown?
3. Hva er den minst frekvente ordklassen?

Hint:

- Bruk `sorted()` metoden for sortering av ord og ordklasser: `sorted(`**<dict>**`.items(), key = lambda pair: pair[1], reverse = True)`
- For å finne ut hva er den *minst* frekvente ordklassen, bruker vi `min_pos =` **<sortert liste av ordklasser>** `[len(`**<sortert liste av ordklasser>**`)-1]`.

2 Flertydighet (50 poeng)

1. Hva mener vi når vi sier at det finnes flertydighet i språket? Gi et eksempel for norsk, og et for engelsk.
2. Hvor mange ord er flertydige i Brown-korpuset? Det vil si, hvor mange ord forekommer med mer enn én ordklassetagg?
3. Hvilket ord har flest tagger, og hvor mange distinkte tagger finner du?
4. Skriv en funksjon `freqs(w)` som tar et ord som argument og skriver ut hvor ofte ordet forekommer med hver av taggene. For eksempel forekommer *run* 20 ganger med NN, 11 ganger med NN, og 4 ganger med VBN.
5. Ved hjelp av funksjonen fra punkt 4, finn frekvenslisten for det mest flertydige ordet i Brown.

I de to første deloppgavene er et viktig poeng at du må passe på at for hvert ord telles hver distinkte tagg nøyaktig en gang; det vil si at for et ord med to forekomster med NP og tre med NN har vi listen `["NP", "NN"]` og ikke `["NP", "NN", "NN", "NN", "NP"]`.

Merk at oppgavene her skal løses *uten* bruk av de NLTKs funksjonalitet for frekvens `nltk.FreqDist` og `nltk.ConditionalFreqDist`.