INF1820 V2014 — Oppgave 2b
Sannsynlighet og tagging

Innlevering: 14/3

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

```
oblig2b_brukernavn.py
```

En god del av besvarelsen din skal angis som ren tekst, og ikke Python-kode. Vennligst skriv besvarelsen din i Python-filen allikevel, som utkommentert tekst. Vi ønsker altså høyst én fil per student, og igjen er det viktig at skriptet ditt faktisk kjører. 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 Tilfeldig samsvar i annotering (15 poeng)

Astrid og Bjarne jobber med korpusannotering av ordklasseinformasjon i et datalingvistisk forskningsprosjekt. Desverre bryr de seg ikke om å lese teksten, men annoterer tilfeldig:

- I 30% av tilfellene tildeler Astrid taggen `DT`
- I 50% av tilfellene tildeler Astrid taggen `NN`
- I 20% av tilfellene tildeler Astrid taggen `VB`

Bjarne følger samme tilfeldige strategi (30% determinativer, 50% substantiver, og 20% verb), men sjekker ikke hva Astrid gjør.

Hva er sannsynligheten for at Astrid og Bjarne er enige om en tagg for et ord?

Du kan godt bruke Python til å løse oppgaven, men du trenger ikke. Det holder å vise utregningene dine. Begrunn svaret med et par setninger.

2 Zipfiansk distribusjon av ord i en tekst (35 poeng)

Dersom frekvensene til ord i et korpus har en Zipfiansk distribusjon, gjelder følgende:

- dersom du rangerer alle ordene etter frekvens
- vil det andre ordet (det andre mest frekvente ordet) opptre ca halvparten så ofte som det første ordet, det tredje ordet en tredjedel så ofte som det første, osv.
- Dvs at ords frekvens er omvendt proporsjonal ("inverse proportional") med dets rang

- Dette kan vi uttrykke som følger (r=rang, f=frekvens, k=en konstant):

$$f = \frac{k}{r}$$

Vi husker at i Brown-korpuset hadde *the* rang 1 og en frekvens på omtrent 70000, og ordet med rang 2 var *of* med frekvens på omtrent 36000, så dette gir oss altså nesten nøyaktig det Zipfs lov forutsier ($70000 \times 1 \approx 36000 \times 2$)

Det hypotetiske Ifi-korpuset inneholder 1 million ord. Det mest frekvente ordet i korpuset (*the*) forekommer 62702 ganger. Det forekommer 56000 forskjellige ord (**typer**, ikke tokens) i korpuset. Distribusjonen av frekvenser i korpuset er fullstendig Zipfiansk, dvs at den nøyaktig følger formelen ovenfor. Bruk Python til å beregne frekvensen for hvert av de 56000 ordene i korpuset basert på den Zipfianske formelen og frekvensen for *the*.

For å løse denne oppgaven trenger du litt ekstra info om Python og desimaltall. Det er slik at hvis begge tallene i en deleoperasjon (`/`) er heltall, vil også resultatet være heltall (det største heltallet som er mindre enn svaret med desimaler, for å være helt presis), og `2/3` vil returnere `0`. For å få desimaltall må heltall konverteres til flyttall med funksjonen `round()`: `float(2)/float(3)` returnerer `0.66666`.

Til slutt må vi runde av tall, siden uttrykket vi bruker til å beregne frekvens sjelden vil returnere heltall. Bruk den innebygde funksjonen `round()` til dette: `round(float(2)/float(3))` returnerer `1.0`.

Hvor mange ord i korpuset har frekvens 1? Skriv ut resultatet.

3 Flertydighet og ordklassetagger (15 poeng)

Velg ut 3 flertydige overskrifter fra denne nettsiden:
`http://www.fun-with-words.com/ambiguous_headlines.html`.
For hver av overskriftene du har valgt skal du gjøre følgende:

1. formulere flertydigheten med egne ord
2. tagge overskriften manuelt med ordklassetagger, bruk ord/tagg-format, slik:
   `Book/V that/DT flight/NN`
3. angi hvorvidt kunnskap om ordklasser entydiggjør overskriften eller ikke

Bruk følgende taggsett:

DT	determinativ (*the*, *a*)
NN	fellesnavn (*table*, *dog*)
NNP	egennavn (*Astrid*, *Oslo*)
JJ	adjektiv (*red*, *serious*)
V	verb (*dance*, *love*)
CC	konjunksjon (*and*, *or*)
P	preposisjon (*of*, *on*)
RB	adverb (*slowly*, *tomorrow*)

Dersom du er usikker, kan du konsultere manualen som ble skrevet for annoteringen av Penn Treebank:
`ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz`

4 Tagging med regulære uttrykk (35 poeng)

I denne oppgaven skal du lage en ordklassetagger med regulære uttrykk.

Taggeren med regulære uttrykk i NLTK-boken sjekker kun noen få uttrykk, så her fins det rom for forbedring. Det burde være mulig å forbedre denne ved å teste for flere prefikser og suffikser. For eksempel kan vi tagge ord som ender på *-able* som adjektiv. Definer en tagger ved hjelp av `nltk.RegexpTagger`, der du formulerer *minst* 10 mønstre i tillegg til de som er nevnt i boken. Dokumenter alle reglene du skriver, samt gi et eksempel på et ord som dekkes inn av regelen.

Ikke glem å håndtere stor/liten bokstav. Husk også at individuelle ord kan utgjøre et regulært uttrykk, så du kan eksempelvis tagge alle forekomster av *the* som determinativ.

Du kan utvikle taggeren din ved å teste den på "adventure"-kategorien fra Brown-korpuset underveis. Når du er ferdig med regelskrivingen, skal du teste taggerens nøyaktighet (s.k. *accuracy*) på "fiction"-kategorien fra Brown-korpuset, samt rapportere resultatene. Du må **ikke** forandre på reglene etter at du har gjort dette. (NB! Poengene du får for denne oppgaven vil ikke være basert på nøyaktigheten til taggeren og det er helt vanlig at denne synker noe når man beveger seg fra treningskorpus til testkorpus.)

Til slutt skal programmet ditt lese inn filen `test_setninger.txt` som ligger ute på emnesiden, tagge den og skrive ut resultatet. Kopier inn output'en i filen din og diskuter minst 3 av feilene taggeren gjør, samt gi forslag til hvordan den kan forbedres.