IN1140 H2017 – Oblig 2b
Språkmodeller og ordklassetagging

Innleveringsfrist, søndag 22.10 kl.23.59

Lever inn svarene dine i Devilry (`https://devilry.ifi.uio.no/`) i filer som angir brukernavnet ditt, slik: `oblig2b_brukernavn.py`.
En perfekt besvarelse på denne oppgaven er verdt 100 poeng.

1 Språkmodeller – (60 poeng)

1.1 Del 1 – (20 poeng)

1. I sannsynlighetsteori, hva er:
   (a) Utfallsrommet
   (b) Betinget sannsynlighet
   (c) Hendelse
   (d) Uavhengige hendelser
   (e) Uniform distribusjon
   (f) Felles sannsynlighet
2. Hva er sannsynligheten for alle mulige utfall?
3. Hva er et n-gram?
4. Skriv Bayes regelen, og gi et eksempel på hva vi kan beregne med den.
5. Forklar kort hva HMM-tagging er.

1.2 Del 2 – (40 poeng)

I Tabell 1 finner du en formel for en språkmodell (en såkalt bigrammodell) samt et lite tekstkorpus. I denne oppgaven skal du ta utgangspunkt i denne formelen og i tekstkorspuset for å besvare følgende spørsmål:

1. Hvilke bigrammer forekommer i korpuset?
2. Hvordan beregner vi sannsynligheten for et ord gitt det foregående ordet ($P(wi|w_{i-1})$) fra et korpus?
3. Du skal nå bruke bigrammodellen samt tekstkorpuset til å beregne sannsynligheten for setningen `<s> Katherine spiller piano <\s>`. Vis hvilke sannsynligheter du trenger samt hvordan disse beregnes fra korpuset.

Formel:

$$P(w_1 \dots w_k) = \prod_{i=1}^{k} P(w_i|w_{i-1})$$

Tekskorpus:

```
<s> Truls spiller piano <\s>
<s> Katherine spiller ikke piano <\s>
<s> Ludovico spiller piano <\s>
```

Table 1: Formel og tekstkorpus.

2 Ordklassetagging med regulære uttrykk (40 poeng)

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

Taggeren med regulære uttrykk i NLTK-boka (del 5.4.2, under overskriften *The Regular Expression Tagger*) sjekker kun noen få uttrykk, så her er det masse rom for forbedring: for eksempel kan vi tagge alle ord på *-able* som adjektiv.

Definer en tagger ved hjelp av `nltk.RegexpTagger` der du har minst 10 uttrykk i tillegg til de som er nevnt i boka. Dokumenter alle reglene dine, og gi minst ett eksempel (for hvert uttrykk) på ord som dekkes.

Husk å håndtere liten og stor bokstav, og at enkeltord også kan brukes i de regulære uttrykkene, slik at f.eks. *the* alltid vil tagges som bestemmer.

Bruk `adventure`-kategorien i Brown mens du utvikler taggeren din ved å teste nøyaktigheten til taggeren (med `evaluate`-metoden, se NLTK-boka). Når du er fornøyd med reglene dine, test taggeren på kategorien `fiction` i Brown, og rapporter resultatene. Det er viktig at du ikke endrer reglene dine etter dette.

Merk: Poengene du får på denne oppgaven er ikke basert på nøyaktigheten til taggeren i den endelige evalueringen på `fiction`-kategorien. Det er helt vanlig at nøyaktigheten til en tagger synker når man tester den på et annet korpus enn den det ble utviklet for.

Til slutt skal programmet ditt lese inn filen `TestSetninger.txt` som er lagt ut sammen med denne obligen, tagge den, og skrive ut resultatet. Kopier outputen inn i filen din og diskuter minst 3 av feilene taggeren gjør, og kom med forslag til hvordan den kan forbedres.