INF5830, H2009, Obigatorisk innlevering 1

Innleveringsfrist 7.10

1 Oppgave: språkmodeller

Vi skal i denne oppgaven se mer på ngrammodeller og redskaper for dem.

a Det er to redskapsbokser ("toolkits") som er nevnt i J& M: SRILM og Cambridge-CMU. I denne oppgaven skal vi bruke Cambridge-CMU som er det minst omfattende. Finn det på nettet. Last det ned og installer det. Gjør deg kjent med det. Du bør også lese artikkelen, Clarkson & Rosenfeld, 1977, *Statistical Language-Modeling using the CMU-Cambridge Toolkit*, som gir en del mer enn dokumentasjonen.

b Vi skal bruke Brown-korpuset. Du finner en versjon på ∼jtl/nlp/browncorpus.txt. Den er uten tagger og med en setning per linje. Trekk ut hver 10. setning til testkorpus. Behold resten til treningskorpus. Du har fått tildelt et unikt nummer n på e-post. Testkorpuset ditt skal bestå av setning nummer n, $n + 10$, $n + 20$, osv., altså setning 7, 17, 27, ... hvis nummeret ditt er 7.

c Lag en språkmodell fra treningskorpuset og finn *preplexity* fra testkorpuset.

d Lag så språkmodeller for ulike ngrammer, fra $n = 1$ (unigram), $n = 2$ (bigram), osv. til $n = 6$. Finn perpleksiteten for de forskjellige modellene. Finn også perpleksiteten for treningskorpuset med de ulike modellene. Tolk resultatene. Velg den modellen du finner best og bruk den videre.

e Programvaren gir mulighet til å eksperimentere med forskjellige glattingsteknikker: absolutt, linjær, Good-Turing, Witten-Bell. Prøv dette. Oppsummer resultatene i en tabell. Hvilken teknikk gir best resultat. Hvilken teknikk er "default" for programmet.

f Hold deg fra nå av til Good-Turing og prøv ut forskjellige "discounting ranges". Sett resultatene i en tabell. Hvilke verdier gir best resultat?

Innleveringen skal bestå av en kort forskningsrapport (2–3 sider) der du under hvert av punktene (d–f) kort beskriver det du har gjort i eksperimentet, presenterer resultatene i en tabell og gir en tolkning av dem.

2 Oppgave: tagging

Vi skal her manuelt simulere en tagger. Vi har gitt et lite korpus med omtrent 10 ordformer og tre forskjellige tagger. Vi vil gjennom hele oppgaven anta at det ikke finnes flere tagger.

```
Jenta/NP sov/V.
Fiskeren/NP laget/V maten/NP.
Fiskeren/NP ga/V jenta/NP maten/NP.
Jenta/NP laget/V bygget/NP.
Fiskeren/NP ga/V jenta/NP maten/NP som/SBU fiskeren/NP laget/V.
Fiskeren/NP som/SBU ga/V jenta/NP maten/NP bygget/V laget/NP.
Maten/NP kastet/V jenta/NP som/SBU bygget/V huset/NP.
Huset/NP som/SBU jenta/NP ga/V fiskeren/NP bygget/V laget/NP.
Laget/NP kastet/V fiskeren/NP som/SBU bygget/V laget/NP.
```

a For taggene, regn ut unigram- og bigramsannsynligheter fra korpuset, f.eks. $P(\text{NP})$ og $P(\text{NP V})$. Regn ut de betingete sannsynlighetene for at en tag skal følge en annen, f.eks. $P(\text{NP} \mid \text{V})$.

b Regn ut betingete sannsynligheter for ordformer gitt tagger f.eks. $P(\text{laget} \mid \text{NP})$.

c Lag en (manuell) uglattet bigramtagger på grunnlag av dette og bruk den til å tagge

```
Jenta bygget laget.
```

Gjør dette ved å sette opp de alternative taggsekvensene, regn ut sannsynlighetene for dem og velg den beste. Husk å ta hensyn til setningsbegynnelse og setningsslutt.

d Hvordan vil taggeren tagge

```
Jenta som sov ga fiskeren maten.
```

Hva er problemet her? Hva må glattes, og hvordan vil du gjøre det?

e Hvordan vil taggeren tagge

```
Jenta kastet.
```

Hva er problemet her? Hva må glattes, og hvordan vil du gjøre det?

-slutt