SPRÅKMODELLER

En modell som beregner sannsynligheten for en sekvens av ord

N-GRAMMODELLER

Brukes for sannsynlighetsberegning

Minikorpus:

<s> Lise elsker å danse <\s>

<s> Ola hater å sykle <\s>

<s> Peter liker å danse <\s>

Unigrammer (n=1): <s>, <Lise>, <elsker>, <å>, <danse>, <\s>

Bigrammer (n=2): <<s>, Lise>, <Lise, elsker>, <elsker, å>, <å, danse>, <danse, <\s> >

BEREGNE SANNSYNLIGHET FOR EN SETNING

Betinget sannsynlighet- sannsynligheten for at noe skjer, *gitt* noe annet:

$$P(A|B) = \frac{P(A \cap B)}{\mathrm{P(B)}} \longrightarrow P(første\ ord| < s >) = \frac{P(første\ ord| < s >)}{P(< s >)}$$

For å beregne sannsynligheten for *resten* av setningen, ganger vi sannsynligheten for hvert bigram sammen:

$$\frac{P(første\ ord|< s >)}{P(<s>)} * \frac{P(neste\ ord|forrige\ ord)}{P(forrige\ ord)} * \frac{P(< \backslash s >|siste\ ord)}{P(siste\ ord)}$$

$$P(< s > \ Ola\ hater\ \text{å}\ danse\ < \backslash s >)$$

Minikorpus:
<s> Lise elsker å danse <\s>
<s> Ola hater å sykle <\s>
<s> Peter liker å danse <\s>

$$P(Ola| < s >) * P(hater|Ola) * P(\text{å}|hater) * P(danse|\text{å}) * P(< \backslash s > |danse)$$

$$= \frac{count(< s >, Ola)}{count(< s >)} * \frac{count(Ola, hater)}{count(Ola)} * \frac{count(hater, \text{å})}{count(hater)} * \frac{count(\text{å}, danse)}{count(\text{å})} * \frac{count(danse, < \backslash s >)}{count(danse)}$$

$$= \frac{1}{3} * \frac{1}{1} * \frac{1}{1} * \frac{2}{3} * \frac{2}{2}$$

= 0.33 ∗ 1 ∗ 1 ∗ 0.66 ∗ 1 = 0.217

= 21.7% sjanse for setningen

NLTK

- Et genialt verktøy for språktekere!
- Inneholder biblioteker for språkprossesering
  - .. Blant annet en *n*-gram-metode for oblig 2a
- Installering: https://www.nltk.org/install.html

- Litt rustne matteegenskaper? 🤨 → forkurs