språkmodeller

October 9, 2023

1 Språkmodeller

1.0.1 En modell som beregner sannsynligheten for en sekvens av ord

2 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>)
```

3 Beregne sannsynlighet for en setning

Vi bruker *betinget sannsynlighet* → sannsynligheten for at noe skjer, *gitt* noe annet

$$P(B|A) \quad = \quad \frac{Count(A,B)}{Count(B)}$$

Sannsynlighet for det første bigrammet:

$$P(frste_ord| < s >) = \frac{Count(< s >, frste_ord)}{Count(< s >)}$$

For å beregne sannsynligheten for resten av setningen, ganger vi sannsynligheten for hvert bigram sammen Gitt setningen, $s = [< s > w_1 w_2 w_3 < \backslash s >]$:

$$P(s) = P(w_1| < s >)\ P(w_2|w_1)\ P(w_3|w_2)\ P(< \backslash s > |w_3)$$

$$P(s) = \frac{Count(< s >, w_1}{Count(< s >)} \frac{Count(w_1, w_2}{Count(w_1)} \frac{Count(w_2, w_3}{Count(w_3)} \frac{Count(w_3, < \backslash s >}{Count(< \backslash s >)}$$

3.1 Eksempel

Minikorpus:

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

Hva er sannsynligheten for setningen, "*Ola hater å danse*"?

Minikorpus

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

P(<s> *Ola hater å danse* <\s>)

= P(*Ola* | <s>) · P(*hater* | *Ola*) · P(*å* | *hater*) · P(*danse* | *å*) · P(<\s> | *danse*)

Unigram counts	
<s>	3
Ola	1
hater	1
å	3
danse	2

Bigram counts	
<s>, *Ola*	1
Ola, hater	1
hater, å	1
å, danse	2
danse, <\s>	2

$$\frac{Count(< s >, Ola)}{Count(< s >)} = \frac{1}{3}$$

$$\frac{Count(Ola, hater)}{Count(Ola)} = \frac{1}{1}$$

$$\frac{Count(hater, å)}{Count(hater)} = \frac{1}{1}$$

$$\frac{Count(å, danse)}{Count(å)} = \frac{2}{3}$$

$$\frac{Count(danse, < \backslash s >)}{Count(danse)} = \frac{2}{2}$$

$$P(< s > Ola\ hater\ å\ danse\ < \backslash s >) = \frac{1}{3} \cdot \frac{1}{1} \cdot \frac{1}{1} \cdot \frac{2}{3} \cdot \frac{2}{2} = 0.217$$

4 NLTK

Verktøy for språkprosessering

Installering: https://www.nltk.org/install.html