22 July 2024 12:28

-> Porter's Algorithm:

1.1 :

- sses -> ss (caresses -> caress)
- ies -> i (ponies -> poni)
- ss -> ss (caress -> caress)
- s -> ϕ (cats -> cat)

1.2 :

- (*v*)ing -> ϕ (walking -> walk, king -> king)
- (*v*)ed -> ϕ (played -> play)

2 :

- ational -> ate (relational -> relate)
- izer -> ize (digitizer -> digitize)
- ator -> ate (operator -> operate)

3 :

- al -> ϕ (revival -> reviv)
- able -> ϕ (adjustable -> adjust)
- ate -> ϕ (activate -> activ)

-> Language Models: Probabilistic Model

→ N-gram LM: (Context Sensitive Spelling Correction)

→ the office is about fifteen minuets from my house.

minutes = minuets 1700-1800

→ p(fifteen minutes) > p(fifteen minuets)

→ Sentence Plausibility:

→ karma → jefe

→ p(high winds) > p(large winds)

→ Probabilistic Lang Models:

$P(W) = P(W_1, W_2, W_3 \ldots W_n)$ ~ Joint

$P(W_4 \mid W_1, W_2, W_3)$ ~ Related

→ Joint Probability:

P(about, fifteen, minutes, from)

P(about, fifteen, minutes, from)

→ Basic : Rely on the chain rule of probability.

→ Conditional Prob : $P(B|A) = \frac{P(A,B)}{P(A)}$

Where $P(A,B) = P(A) \cdot P(B|A)$

→ Generalized Rule :

$$P(x_1, x_2 \cdots x_n) = P(x_1) \cdot P(x_2|x_1) \cdot P(x_3|x_1,x_2) \cdots \cdots P(x_n|x_1 \cdots x_{n-1})$$

$$\therefore P(w_1,w_2 \cdots w_n) = \prod_i P(w_i|w_1,w_2 \cdots w_{i-1})$$

→ P(about fifteen minutes from) =

P(about) * P(fifteen|about) * ....

→ Markov Assumption :

P(office | about fifteen min from) ≈ P(office | from)

→ Markov Assumption : Only k previous words (k^{th} order)

$$P(w_1,w_2 \cdots w_n) \approx \prod_i P(w_i|w_{i-k} \cdots w_{i-1})$$

→ Maximum Likelihood Estimation:

$$P(w_i \mid w_{i-1}) = \frac{\text{count}(w_{i-1}, w_i)}{\text{count}(w_{i-1})}$$

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$