$L = \{ so, soo, sooo, \dots \}$, $|L| = \infty$

the regular expression (RE, re) that describes L is so+

↙

a meta-character, 1 or more repetitions.

A, B two tokens

pmi (A, B) = $\log_2 \left(\frac{p(AB)}{p(A)\,p(B)} \right)$ ← WordPiece

$p(AB) = \frac{\#\text{ times } AB \text{ appears in text } T}{\#\text{ tokens in } T}$

$p(A) = \frac{\#\text{ times } A \text{ appears in } T}{\#\text{ tokens in } T}$

$p(B) = \frac{\#\text{ times } B \text{ appears in } T}{\#\text{ tokens in } T}$ → N

$= \frac{\frac{\#AB}{N}}{\frac{\#A}{N}\,\frac{\#B}{N}} = N \cdot \frac{\#AB}{\#A \cdot \#B}$

=> ranking pairs AB based on pmi (A, B) <=> ranking using $\frac{\#AB}{\#A\,\#B}$.