

R

共起スコア

logDice

■ まず、Dice 係数を出す

1. 「strong tea」の共起頻度
2. 「strong」の頻度
3. 「tea」の頻度

Dice 係数 = (共起頻度 × 2) ÷ (片方の頻度 + もう片方の頻度)

■ Dice 係数の底を 2 とする対数を取り、14 を足したものが、logDice

$\text{logDice} = 14 + \log_2(\text{Dice})$

■ 解釈

- ・ 最大値 14 (すべての場合に共起する)
- ・ スコア 0 : いずれかの語の 16,000 語あたりの共起回数が 1 以下
- ・ スコアがマイナスの場合 : 統計的に共起関係はないといえる。
- ・ スコアを比較して、
 - ・ 片方が 1 大きい場合、共起頻度は 2 倍
 - ・ 片方が 7 大きい場合、共起頻度は 100 倍
- ・ スコアはコーパスサイズに依存しない。(相対頻度なので)

■ R のスクリプト

- ・ 使い方
 - ・ このスクリプトを R で読み込んで、
 - ・ 後は必要な頻度数を入れる。
 - ・ 共起表現 xy 頻度、単語 x の頻度、単語 y の頻度

LD(xy, x, y)

■ Reference

Pavel Rychlák; (2008) A Lexicographer-Friendly Association Score, In Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2008, pp. 6–9, 2008.