第5章
テキストにおける統計法則と指標

佐々木研究室
荒井悠有

・概要

テキストにおける主な法則と指標を紹介

-ジップの法則

-語彙の豊富さ指標（トークン比、K特性値）

-語の重みの指標TF-IDF

・ジップの法則

データの順位（ランク）と頻度の間には次の法則がある。

順位 × 頻度 ≒ 定数

この法則をジップの法則と呼ぶ。

→言語に限らず何らかの順位と頻度の関係に適用可能。

頻度、順位、定数をそれぞれf,r,cとすると

$$f_r \fallingdotseq \frac{c}{r} \qquad (r=1,2,3,\cdots,n)$$

でジップの法則は表される。
また、ジップの法則を次のように拡張した法則もある。

1. $f_r \fallingdotseq \frac{c}{r^{\,a}}$

2. $f_r \fallingdotseq \frac{c}{(b+r)^{\,a}}$

共に(r=1,2,3,・・・,n)でa,b,cはデータに依存する定数。
式2は**Zipf-Mandelbrot法則**と呼ばれている。

▪語彙の豊富さの指標

-総単語数を「延べ語数」→N
-単語の種類を「異なり語数」→V
-テキスト中でm回使用された語数をV(m,N)

→mとV(m,N)のデータを<u>頻度スペクトル</u>と呼ぶ。

N,V,V(m,N)は相互に次の関係を持っている。

$$V = \sum_{all\_m} V(m, N)$$

$$N = \sum_{all\_m} mV(m, N)$$

・トークン比

語彙の豊富さを示す最も簡単な指標。
延べ語数Nに対する異なり語数Vの比率(Type-Token-Ratio)

$$TTR = \frac{V}{N}$$

これを<u>タイプ・トークン比</u>と呼ぶ。

・K特性値

単語が用いられている回数（頻度スペクトル）を用いた指標。
単語の出現頻度はポアソン分布に従うと仮定。

$$K = 10^4 \frac{\sum_{all\_m} m^2 V(m, N) - N}{N^2}$$

Kの値が小さいほど語彙が豊富であることを示す。

・語の重み指標

語の重み指標としてTF-IDFがある。

-TF(Term Frequency)

テキストdにおける語tの頻度(tf)

-IDF(Inverted Document Frequency)

語tが検索対象の中のどれくらいのテキストに使用されているかに関する指標

$$IDF = \log\left(\frac{N}{df}\right)$$

→ $$TF - IDF = tf \times \log\left(\frac{N}{df}\right)$$

この値が大きいほど、その語の重要度が大きい。