特徴語抽出に使用した統計的指標の式
Statistical Measures Used for Identifying Specified Vocabularies

内山将夫 (Masao Utiyama)　中條清美 (Kiyomi Chujo)

頻度： $Freq=a$

自己相互情報量： $MI = \log(an/((a+b)(a+c)))$

ダイス係数： $2a/((a+b)+(a+c))$

コサイン： $Co\sin e = a/\sqrt{(a+b)(a+c)}$

補完類似度： $CSM = (ad-bc)/\sqrt{(a+c)(b+d)}$

McNemar： $McNemar = \dfrac{(b-c)^2}{b+c}$

対数尤度比： $LLR_0 = a\log(an/((a+b)(a+c))) + b\log(bn/((a+b)(b+d))) + c\log(cn/((c+d)(a+c))) + d\log(dn/((c+d)(b+d)))$

カイ二乗値： $Chi2_0 = (n(ad-bc)^2)/((a+b)(c+d)(a+c)(b+d))$

イエーツの補正公式： $Yates_0 = n(|ad-bc|-n/2)^2/((a+b)(c+d)(a+c)(b+d))$

上記３指標の補正： $LLR = sign(ad-bc)\times LLR_0$， $Chi2 = sign(ad-bc)\times Chi2_0$， $Yates = sign(ad-bc)\times Yates_0$

$$sign(z)=\begin{cases}+1 & if\ z>0\\ -1 & otherwize\end{cases}$$