# TF-IDF 演算法

**假設有 D 篇文章，每一篇文章經過斷字處理所有文章的詞彙共有 T 個**

$$\begin{array}{cc} & \begin{array}{cccc} 文件_1 & 文件_2 & & 文件_D \end{array} \\ \begin{array}{c} 詞彙_1 \\ 詞彙_2 \\ \vdots \\ 詞彙_T \end{array} & \begin{bmatrix} n_{1,1} & n_{1,2} & \cdots & n_{1,D} \\ n_{2,1} & n_{2,2} & \cdots & n_{2,D} \\ \vdots & \vdots & \vdots & \vdots \\ n_{T,1} & n_{T,2} & \cdots & n_{T,D} \end{bmatrix} \end{array} \rightarrow \begin{array}{cc} & \begin{array}{cccc} 文件_1 & 文件_2 & & 文件_D \end{array} \\ \begin{array}{c} 詞彙_1 \\ 詞彙_2 \\ \vdots \\ 詞彙_T \end{array} & \begin{bmatrix} tf_{1,1} & tf_{1,2} & \cdots & tf_{1,D} \\ tf_{2,1} & tf_{2,2} & \cdots & f_{2,D} \\ \vdots & \vdots & \vdots & \vdots \\ tf_{T,1} & tf_{T,2} & \cdots & tf_{T,D} \end{bmatrix} \end{array} \tag{1}$$

| **詞彙** | 文件$_1$ | 文件$_2$ | 文件$_3$ | 文件$_4$ |
|---|---|---|---|---|
| 詞彙$_1$ =滷 | $n_{1,1} = 1$ | | | |
| 詞彙$_2$ =肉 | $n_{2,1} = 2$ | 1 | 2 | $n_{2,4} = 8$ |
| 詞彙$_3$ =飯 | $n_{3,1} = 2$ | 2 | 1 | $n_{3,4} = 3$ |
| 詞彙$_4$ =不 | $n_{4,1} = 6$ | 5 | 6 | $n_{4,4} = 4$ |
| 詞彙$_5$ =符合 | $n_{5,1} = 1$ | 1 | | |
| 詞彙$_6$ =期待 | $n_{6,1} = 1$ | | | |
| 詞彙$_7$ =偏 | $n_{7,1} = 1$ | | 1 | $n_{7,4} = 1$ |
| 詞彙$_8$ =乾 | $n_{8,1} = 1$ | | 2 | $n_{8,4} = 1$ |
| 詞彙$_9$ =其他 | $n_{9,1} = 1$ | 2 | 4 | |
| 詞彙$_{10}$ =還不錯 | $n_{10,1} = 1$ | | 2 | $n_{10,4} = 1$ |
| 詞彙$_{11}$ =個人 | $n_{11,1} = 3$ | 1 | | |
| 詞彙$_{12}$ =享用 | $n_{12,1} = 1$ | 1 | | |
| | $\sum_{k=1}^{12} n_{k,1} = 21$ | $\sum_{k=1}^{12} n_{k,2} = 13$ | $\sum_{k=1}^{12} n_{k,3} = 18$ | $\sum_{k=1}^{12} n_{k,4} = 18$ |

$$tf_{t,d} = \frac{n_{t,d}}{\sum_{k=1}^{T} n_{k,d}} \tag{2}$$

$$tf_{1,1} = \frac{詞彙_1在文件_1出現的次數}{文件_1中所有詞出現次數的總和} \tag{3}$$

| **詞彙** | 文件$_1$ | 文件$_2$ | 文件$_3$ | 文件$_4$ |
|---|---|---|---|---|
| 詞彙$_1$ = 滷 | $tf_{1,1} = \frac{1}{21} = 0.0476$ | | | |
| 詞彙$_2$ = 肉 | $tf_{2,1} = \frac{2}{21} = 0.0952$ | | | $tf_{1,4} = \frac{8}{18} = 0.4444$ |
| 詞彙$_3$ = 飯 | $tf_{3,1} = \frac{2}{21} = 0.0952$ | | | $tf_{2,4} = \frac{3}{18} = 0.1666$ |
| 詞彙$_4$ = 不 | $tf_{4,1} = \frac{6}{21} = 0.2857$ | | | $tf_{3,4} = \frac{4}{18} = 0.2222$ |
| 詞彙$_5$ = 符合 | $tf_{5,1} = \frac{1}{21} = 0.0476$ | | | |
| 詞彙$_6$ = 期待 | $tf_{6,1} = \frac{1}{21} = 0.0476$ | | | |
| 詞彙$_7$ = 偏 | $tf_{7,1} = \frac{1}{21} = 0.0476$ | | | $tf_{7,4} = \frac{1}{18} = 0.0555$ |
| 詞彙$_8$ = 乾 | $tf_{8,1} = \frac{1}{21} = 0.0476$ | | | $tf_{7,4} = \frac{1}{18} = 0.0555$ |
| 詞彙$_9$ = 其他 | $tf_{9,1} = \frac{1}{21} = 0.0476$ | | | |
| 詞彙$_{10}$ = 還不錯 | $tf_{10,1} = \frac{1}{21} = 0.0476$ | | | $tf_{7,4} = \frac{1}{18} = 0.0555$ |
| 詞彙$_{11}$ = 個人 | $tf_{11,1} = \frac{3}{21} = 0.1428$ | | | |
| 詞彙$_{12}$ = 享用 | $tf_{12,1} = \frac{1}{21} = 0.0476$ | | | |

$$idf_t = log\left(\frac{\text{文章總數}}{\text{詞彙}_t\text{出現過的文章篇數}}\right) \tag{4}$$

| **詞彙** | $idf_t = log\left(\frac{D}{d_t}\right)$ |
|---|---|
| 詞彙$_1$ = 滷 | $idf_1 = log\left(\frac{4}{1}\right) = 0.60205999$ |
| 詞彙$_2$ = 肉 | $idf_2 = log\left(\frac{4}{4}\right) = 0$ |
| 詞彙$_3$ = 飯 | $idf_3 = log\left(\frac{4}{4}\right) = 0$ |
| 詞彙$_4$ = 不 | $idf_4 = log\left(\frac{4}{4}\right) = 0$ |
| 詞彙$_5$ = 符合 | $idf_5 = log\left(\frac{4}{2}\right) = 0.30103$ |
| 詞彙$_6$ = 期待 | $idf_6 = log\left(\frac{4}{1}\right) = 0.60205999$ |
| 詞彙$_7$ = 偏 | $idf_7 = log\left(\frac{4}{3}\right) = 0.12493874$ |
| 詞彙$_8$ = 乾 | $idf_8 = log\left(\frac{4}{3}\right) = 0.12493874$ |
| 詞彙$_9$ = 其他 | $idf_9 = log\left(\frac{4}{3}\right) = 0.12493874$ |
| 詞彙$_{10}$ = 還不錯 | $idf_{10} = log\left(\frac{4}{3}\right) = 0.12493874$ |
| 詞彙$_{11}$ = 個人 | $idf_{11} = log\left(\frac{4}{2}\right) = 0.30103$ |
| 詞彙$_{12}$ = 享用 | $idf_{12} = log\left(\frac{4}{2}\right) = 0.30103$ |