資料擷取

TF-IDF

MA690107 郭柏均

- TF-IDF 是 term frequency–inverse document frequency 的英文縮寫
- 是一種用於資訊檢索與文字挖掘的常用加權技術
- 用以評估一字詞對於一個檔案集或一個語料庫中的其中一份檔案的重要程度
- 字詞的重要性隨著它在檔案中出現的次數成正比增加，但同時會隨著它在語料庫中出現的頻率成反比下降

Term 代表的是關鍵字，此統計方法能夠以關鍵字來尋找與此關鍵字有高度相關的內文，而此方法也常為搜尋引擎所使用

TF

詞頻 (term frequency，TF) 指的是某一個給定的詞語在該檔案中出現的頻率，用來防止它偏向長的檔案 (同一個詞語在長檔案裡可能會比短檔案有更高的詞數，而不管該詞語重要與否)

$$TF(T_i, D_j) = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

- T_i：某一字詞
- D_j：該字詞所在的檔案
- $n_{i,j}$：是該字詞在檔案 D_j 中的出現頻率
- 分母則是在檔案 D_j 中所有字詞的出現次數之和

為何不只利用 $n_{i,j}$ 計算就好？

$$TF(T_i, D_j) = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

因為每篇文件的長短不盡相同，因此必須利用分母來予以正規化，才能讓每篇文件的計算結果有相同的意義

IDF

逆向檔案頻率 (inverse document frequency，IDF) 是評估某一個詞在所有文件中所扮演的重要性

$$IDF(T_i) = \log \frac{|D|}{|DF(T_j)|}$$

- $|D|$：訓練集中的檔案總數
- $|DF(T_j)|$：包含詞語 T_i 的檔案數目，如果詞語不在資料中，就導致分母為零，因此一般情況下會使用 $1 + |DF(T_j)|$

TF-IDF

綜合 TF 及 IDF ，便可同時將詞在特定單一檔案中的影響力，以及某個詞在所有檔案中的影響力一起納入評估

$$TF - IDF(i,j) = TF(i,j) \times IDF_i$$

- $TF(i,j)$：T_i 在檔案 D_j 中出現的頻率越高，就代表 T_i 在檔案 D_j 中的重要性越高
- $TF - IDF(i,j)$：在所有檔案中，T_i 在幾個檔案中出現

為何不只利用 TF 計算就好？

$$TF - IDF(i,j) = TF(i,j) \times IDF_i$$

- TF 只用來判斷該字詞有多普遍，但不管該字詞是否重要
- 但 IDF 是來評估該字詞在所有文件中所扮演的重要性
- 因此不同的字詞會依不同檔案而有不同重要性

實作案例

共有 100 篇文章，
從其中兩篇文章中篩選出兩個重要詞：
T_1 和 T_2，每篇文章都共有 200 字詞
下面已 D_1 和 D_2 為例

- T_1 在 D_1 中出現 70 次
- T_2 在 D_1 中出現 30 次
- T_1 在 D_2 中出現 40 次
- T_2 在 D_2 中出現 60 次

TF

$$
\begin{aligned}
TF(T_1, D_1) &= \frac{70}{200} = 0.35 \\
TF(T_2, D_1) &= \frac{30}{200} = 0.15 \\
TF(T_1, D_2) &= \frac{40}{200} = 0.2 \\
TF(T_2, D_2) &= \frac{60}{200} = 0.3
\end{aligned}
$$

TF 值愈高，其單詞出現的頻率越高

- T_1 對 D_1 比較重要，T_2 對 D_2 比較重要
- 搜尋 T_1，那 D_1 會出現在較前面的位置
- 搜尋 T_2，則 D_2 會出現在較前面的位置

IDF

- 共有 100 篇文章
- T_1 在 10 篇文章中出現，T_2 在 50 篇文章中出現

$$IDF(T_i) = \log \frac{|D|}{|DF(T_j)|}$$

$$IDF(T_1) = \log \frac{100}{10} = 1$$
$$IDF(T_2) = \log \frac{100}{50} = 0.3$$

IDF 值越高，代表該字詞在所有文章中越重要；
IDF 值越低，代表該字詞越常見，重要性較低

因此 T_1 比 T_2 還要重要

TD-IDF

以某一特定文件內的高單詞頻率，乘上該單詞在檔案總數中的低文件頻率，便可產生 TF-IDF 權重值

$$TF - IDF(T_i, D_j) = TF(T_i, D_j) \times IDF(T_i)$$

$$TF - IDF(T_1, D_1) = 0.35 \times 1 = 0.35$$
$$TF - IDF(T_1, D_2) = 0.15 \times 1 = 0.15$$
$$TF - IDF(T_2, D_1) = 0.2 \times 0.3 = 0.06$$
$$TF - IDF(T_2, D_2) = 0.3 \times 0.3 = 0.09$$

TF-IDF 用於過濾掉常見的詞語，保留重要的詞語，因此 T_2 重要性不高

END