TF-IDF

陸裕豪

November 17, 2020

文件矩陣

假設現在有 D 篇文件 (document)，而所有文件中總共使用了 T 個詞彙 (term)，我們就可以將文章轉換成以下類型的矩陣

- 某一個給定的詞語在該文件中出現的頻率，第 t 個詞出現在第 d 篇文件的頻率記做 $tf_{t,d}$。

Table: 文件矩陣

詞彙	文件 1	文件 2	$\cdots$	文件 D
詞 1	10	1	$\cdots$	2
詞 2	15	6	$\cdots$	97
詞 3	28	54	$\cdots$	77
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
詞 5	14	77	$\cdots$	2

文件矩陣

第一欄第一列的「10」代表的是「 文件 1」中出現了 10 個「 詞 1」。這樣就可以用 $[10, 15, 28, \cdots, 14]$ 這個向量來代表「 文件 1」，同理「 文件 D」也可以用 $[2, 97, 77, \cdots, 2]$ 來表示。

- 問題 1：
  每篇文章的總字數不一樣，例如詞 2 在文件 2 中出現 6 次，在文件 D 中卻只出現 97 次，這樣是否代表詞 2 對文件 D 比較重要，對文件 2 比較不重要呢？
- 問題 2：
  時常重複出現的慣用詞彙對一個文件的影響很大。比如說，上圖中的詞 3 在每個文件中都出現好多次，可能是'the' 之類的常用詞，如此一來文件 D 的向量就會被'the' 這個字所主導，但'the' 這個詞其實沒什麼特別的意義。

TF-IDF

TF-IDF 演算法包含了兩個部分

- 詞頻 (term frequency,TF)
- 逆向文件頻率 (inverse document frequency, IDF)

詞頻 (term frequency,TF)

TF 是處理每一個「文件」中所有「詞」的問題。

- 某一個給定的詞語在該文件中出現的頻率，第 t 個詞出現在第 d 篇文件的頻率記做 $tf_{t,d}$。
- $$tf_{t,d} = \frac{n_{t,d}}{\sum_{k=1}^{T} n_{k,d}}$$
- 例如: 文件 1 總共有 100 個詞，而第 1 個詞在文件 1 出現的次數是 10 次，因此 $tf_{1,1} = 10/100$。
- 第一個問題得到修正：以頻率看待文字的重要性，而非次數，使文章與文章之間更具可比較性。

逆向文件頻率 (inverse document frequency, IDF)

IDF 是處理每一個「詞」在所有「文件」中的問題。

- 假設「詞 t」在總共在 d_t 篇文章中出現過，「詞 t」的 IDF 定義為
- $$idf_{t,d} = \log \frac{D}{d_t}$$
- 假設 D 是' 所有文件的總數'(但也可能為' 所有詞的總數')，由公式可以得知' 詞' 在越多文件中出現代表，相對應的 idf 會比較小，也就是這個' 詞' 可能不重要或沒太大意義，比如說'is、with、the' 這類型的詞。

TF-IDF

TF-IDF 就是透過 TF 和 IDF 算每一個「詞」對每一篇「文件」的分數 (score)，定義為

$$\text{score}_{t,d} = tf_{t,d} \times idf_t$$

	文件 1	文件 2	⋯	文件d	⋯	文件D
詞 1	$tf_{1,1} \times idf_1$	$tf_{1,2} \times idf_1$	$\cdots$	$tf_{1,d} \times idf_1$	$\cdots$	$tf_{1,D} \times idf_1$
詞 3	$tf_{2,1} \times idf_2$	$tf_{2,2} \times idf_2$	$\cdots$	$tf_{2,d} \times idf_2$	$\cdots$	$tf_{2,D} \times idf_2$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\cdots$	$\vdots$
詞t	$tf_{t,1} \times idf_t$	$tf_{t,2} \times idf_t$	$\cdots$	$tf_{t,d} \times idf_t$	$\cdots$	$tf_{t,D} \times idf_t$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
詞T	$tf_{T,1} \times idf_T$	$tf_{T,2} \times idf_T$	$\cdots$	$tf_{T,d} \times idf_T$	$\vdots$	$tf_{T,D} \times idf_T$