Term-frequency * Inverse-document-frequency

Tuesday, August 8, 2023 7:53 PM

$$\begin{array}{ccc} \text{TF-IDF} \rightarrow & \text{TF} & \times\ \text{IDF} \\ & \downarrow & \downarrow \\ & F() & \times\ F() \end{array} \rightarrow \text{Cantidad} \rightarrow \text{Matriz} \rightarrow \begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix} \rightarrow \begin{array}{c} \text{SVM} \\ \text{NB} \\ \text{MLP} \end{array}$$

(documento-término)
(document-term)

Term frequency : Cantidad de ocurrencias de una palabra w_i en el documento d_i

document frequency : Cantidad de documentos que contienen a la palabra w_i

CountVectorizer() → tFidF()

	and	document	first	is	one	second	the	third	this
D1 This is the first document.	0	$1 = \log(\frac{4}{3})$	1	1	0	0	1	0	1
D2 This document is the second document.	0	1	0	1	0	1	1	0	1
D3 And this is the third one.	1	0	0	1	1	0	1	1	1
D4 Is this the first document?	0	1	1	1	0	0	1	0	1
dF_i	1	3	2	4	1	1	4	1	4
idF_i	$\log(\frac{4}{1})$	$\log(\frac{4}{3})$	$\log(\frac{4}{2})$	$\log(\frac{4}{4})$	. - . . .				
$tF \times idF$									

(bag-of-words)

$w_i = \underline{\text{and}}$

$tF = \text{count}(t_i, d_i)$

$idF = \log\left(\frac{N}{dF_i}\right)$ N: Cantidad de documentos en el corpus $\boxed{N=4}$

para "and" → $tF(\text{"and"}, D1) = 0$

$idF\left(\frac{4}{dF_{and}}\right) = \log\left(\frac{4}{1}\right)$

$0 \times \log\left(\frac{4}{1}\right) = 0$

Matrices con muchos ceros → sparse matrix
∨ sin → dense matrix

TFiDF vs TF (Bag-of-words)

ponderaciones
↓
palabras:
menos repetidas: ↑

enfoque por contador

mas repetidas: ↓

$$\text{idf}(t) = \log \frac{1+n}{1+\text{df}(t)} + 1, \quad \text{(sklearn)}$$

$$\log\left(\frac{1+4}{1+3}\right) + 1 = \log\left(\frac{5}{4}\right) + 1$$