TF-IDF

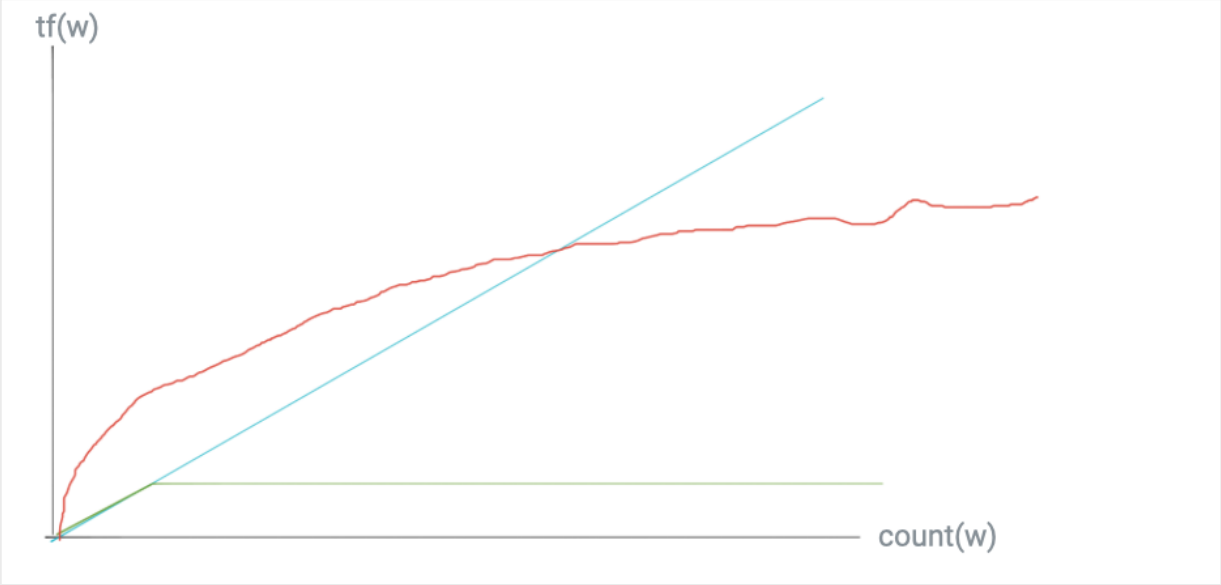
TF - IDF = term frequency–inverse document frequent

NOTA:
El que sea el valor nulo lo debemos hacer 0 porque "a pesar de no aparecer existe y debe ser contrastada"

TOMAR MUESTRAS
- Aumentando la muestra
- Distribuyendo los textos

CORPUS = Conjunto de textos compilados con un fin específico que idealmente es representativo de los fenómenos que se pretenden estudiar y balanceado con relación a la lengua.

Omitir totalmente el **GIGO** (Garbage in, garbage out)



- Tf es la frecuencia de las palabras en cada documento del CORPUS
- Count es el conteo de las palabras

Funciones sub-lineales => son todas las que van por debajo del conteo directo (Esta es lineal)

Inverse document frequency [edit]

The **inverse document frequency** is a measure of how much information the word provides, i.e., if it's common or rare across all documents. It is the logarithmically scaled inverse fraction of the documents that contain the word (obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient):

$$\operatorname{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

with

- N: total number of documents in the corpus $N = |D|$
- $|\{d \in D : t \in d\}|$: number of documents where the term t appears (i.e., $\operatorname{tf}(t, d) \neq 0$). If the term is not in the corpus, this will lead to a division-by-zero. It is therefore common to adjust the denominator to $1 + |\{d \in D : t \in d\}|$.

Variants of inverse document frequency (idf) weight

weighting scheme	idf weight ($n_t = \|\{d \in D : t \in d\}\|$)
unary	1
inverse document frequency	$\log \frac{N}{n_t} = -\log \frac{n_t}{N}$
inverse document frequency smooth	$\log\left(\frac{N}{1+n_t}\right) + 1$
inverse document frequency max	$\log\left(\frac{\max_{\{t' \in d\}} n_{t'}}{1+n_t}\right)$
probabilistic inverse document frequency	$\log \frac{N-n_t}{n_t}$

EJERCICIO

SOBRE REPRESENTACIÓN = Palabra que se repite ya no es importante

		w1	w2	w3
2	d1	0,23	0,12	0,23
3	d2	0,12	0,23	0,12
4	d3	1,23	0,23	1,23
5	d4	0,23	0,12	0,50
6	d5	0,12	0,12	8,23
7	d6	1,23	0,23	4
8	d7	0,23	0,23	1,35
9	d8	0,12	0,12	0,23
10	d9	1,23	0,12	0,23
11	d10	4,123	0,23	0,23

En este caso el TF e IDF es muy alto , es decir es muy frecuente esa palabra específicamente en ese documento

Esta representación sirve para el resumen y sobre todo para agrupar documentos muy parecidos y lo validamos mediante:

- Determinar distancias entre vectores

		w1	w2	w3
2	d1	0,23	0,12	0,23
3	d2	0,12	0,23	0,12
4	d3	1,23	0,23	1,23
5	d4	0,23	0,12	0,50
6	d5	0,12	0,12	8,23
7	d6	1,23	0,23	4
8	d7	0,23	0,23	1,35
9	d8	0,12	0,12	0,23
10	d9	1,23	0,12	0,23
11	d10	4,123	0,23	0,23

De esta forma vemos que tan habitual es una palabra en el universo además del significado de la misma

- Si las palabras tiende a aparecer en contextos similares, estas tendrás un significado similar