

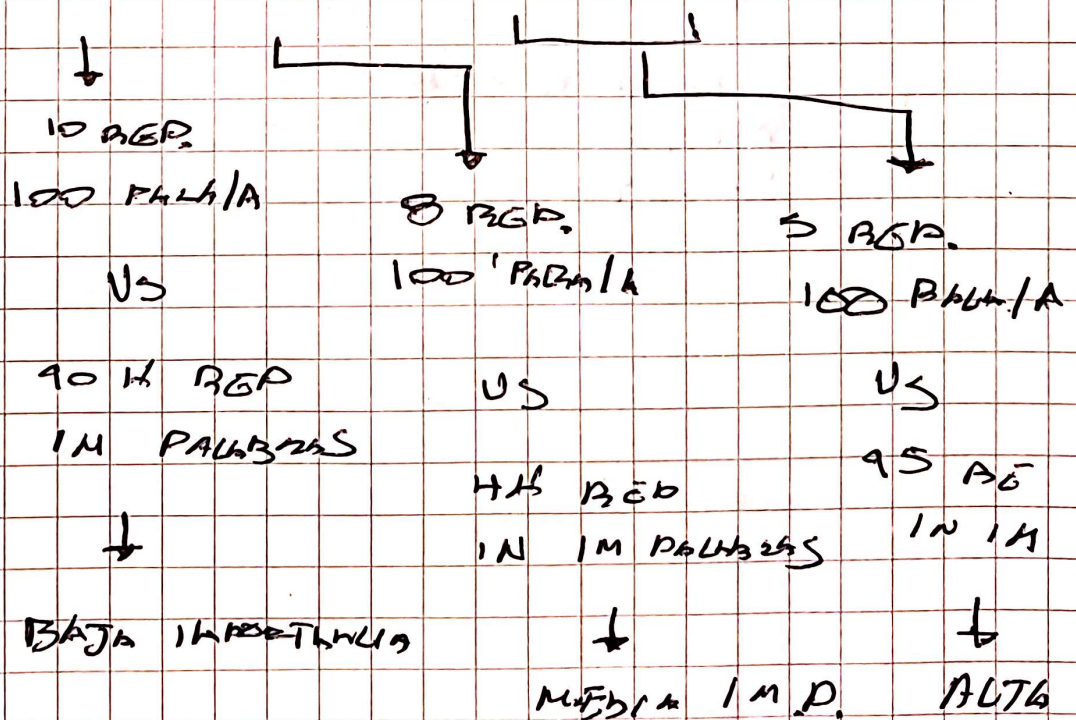
TF-IDF (TERM FREQUENCY - INVERSE DOC. FREQ.)

→ ES UN CÁLCULO ESTADÍSTICO QUE MIDE QUE TÉRMINOS SON MÁS RELEVANTES PARA UN ASUNTO ESPECÍFICO.

ANALISA LA FRECUENCIA CON QUE APARECEN EN UNA PÁGINA, VERSUS, LA FRECUENCIA EN UN CONJUNTO MÁS GRANDE DE PÁGINAS.

→ EL VALOR DEL IDF CONSIDERA TÉRMINOS QUE SON FRECUENTES EN LOS TEXTOS COMO LOS ARTÍCULOS (EL, LA, LOS, LAS) Y NO TIENEN RELEVANCIA PARA LOS DOCUMENTOS.

THE BEST CITY BIKE



Fórmulas

$$TF = \frac{\# \text{ TÉRMINO A APLICAR EN } d}{\# \text{ TÉRMINOS EN } d}$$

→ ES EFECTIVO EN:

- BUSCADORES DE TEXTO
- FILTRADO DE INFORMACIÓN
- ANÁLISIS DE SIMILITUDES DE REGS.
- EXTRAER PALABRAS CLAVE

→ BIBLIOTECAS

- SCILIT REAZN
- NLTK
- SPACY

PROBLEMA DE LOS N-GRAM

→ MODELO DE LENGUAJE ES UNA ASIGNACIÓN DE PROBABILIDADES $P(W)$ A CADA POSIBLE FRASE DEL LENGUAJE $W = w_1 w_2 w_3 \dots w_n$.

EL CONTADOR N-GRAMAS ES UNO DE ESTOS MODELOS DE LENGUAJE.

$$W = w_1 w_2 w_3 \dots w_n$$

w_i SON PALABRAS QUE CONTIENEN EL TEXTO W .

$$P(w_n | w_1, w_2, \dots, w_{n-1}) = \frac{\text{CONTED}}{\text{CONTED}}$$

~~STA~~ EL CONTED NOS DA LA CONDICIÓN. LI HAS CERO.

ESTO ES PORQUE NO USAMOS NINGÚN TIPO DE SUJIZADO

$$\rightarrow P(w_n | w_1, w_2, \dots, w_{n-1}) = \frac{\text{CONTED} + 1}{\text{CONTED} + V}$$

V = TAMAÑO DEL VOCABULARIO.

ESTÉ SUJIZADO HAZÉ QUE NO SE 0.

→ EUTR PROBABILIDADES QUE SON 0.
NO SE PUEDEN SUSPENDER
HAY OTROS 46 CHILAS QUE SON MUY
EFFECTIVAS PARA FUNCIONES.

→ SI HAY UNA PACHA DE TOST SÚ
QUE NO ESTE EN EL VOLUNTARIO SE
LE LLEVA DON (DOT-OF-VOLUNTARY
REQ) COMO MENTORADO HAYOS.

SI HAY UNA DON WOT LA PACHA.
HAY A 0. Y NO ROBERT HAYOS PACHA
LICHES EL MODELO.

→ PACHA PACHA:

- TOXEN <UNK>
 - SUBSTITUTIONS (GOOD-TORING)
 - MATHS MATHS
 - CHANGING
- SE LLEVA ESTE TOXEN A LA PACHA
DON Y SE LE ASIGNA PACHA.