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

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

ANALIZA LA FRECUENCIA CON QUÉ APARECEN EN UNA PÁGINA, WEBSITE, LA FRECUENCIA EN UN CONJUNTO MÁS GRANDE DE PÁGINAS.

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

TRE BEST CITY BIKE

10 BGP.
100 PALA/A
VS
90 K BGP
1M PALABRAS
↓
BAJA IMPORTANCIA

8 BGP.
100 PALA/A
VS
44 BGP
1M 1M PALABRAS
↓
MEDIA I.M.P.

5 BGP.
100 PALA/A
VS
95 BG
1M 1M
↓
ALTA

FÓRMULAS

$$TF = \frac{\#\ \text{TÉRMINO } t \text{ APARECE EN } d}{\#\ \text{TÉRMINOS EN } d}$$

→ ES EFECTIVO EN ÍNDICE

- BUSCADORES DE TEXTO
- FILTRADO DE INFORMACIÓN
- ANÁLISIS DE SIMILARIDAD DE DOCS.
- EXTRAER PALABRAS CLAVE

→ BIBLIOTECAS

- SCIKIT-LEARN
- NLTK
- SPACY

PROBLEMA DE LOS N-GRAM

→ MODELO DE LENGUAJE ES UNA ASIGNACIÓN DE PROBABILIDAD P(W) A CADA SECUENCIA DEL LENGUAJE $W = w_1 w_2 w_3 \ldots w_n$

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

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

w_i SON PALABRAS QUE CONTIENEN EL TEXTO W.

$$P(w_n \mid w_1, w_2, \ldots w_{n-1}) = \frac{\text{CONTEO}}{\text{CONTEO}}$$

~~SIN~~ EL CONTEO NOS DARÍ CERO PROBABILIDAD. LIBRO CERO

ESTO ES PORQUE NO USAMOS NINGÚN TIPO DE SUAVIZADO

→ $$P(w_n \mid w_1, w_2 \ldots w_{n-1}) = \frac{\text{CONTEO} + 1}{\text{CONTEO} + V}$$

V: TAMAÑO DEL VOCABULARIO.

ESTO SUAVIZADO HACE QUE NO SEA 0.

→ EVITA PROBABILIDADES QUE SON 0.

PUEDE LLEVAR A SESGADAD

HAY OTRAS TÉCNICAS QUE SON MÁS EFECTIVAS PERO FUNCIONALES.

→ SI HAY UNA PALABRA DEL TEST SET QUE NO ESTÉ EN EL VOCABULARIO SE LE LLAMA OOV (OUT-OF-VOCABULARY) COMO MENCIONADO ANTES.

SI HAY UNA OOV WORD LA PROBA. IRÁ A 0. Y NO PODRÁ HACER PREDICCIONES EL MODELO.

→ PARA PALIARLAS:

- → TOKEN <UNK>
- → SUAVIZADOS AVANZADOS (GOOD-TURING)
- → MODELADO MORFOLÓGICO
- → EMBEDDING

→ SE COLOCA ESTE TOKEN A LA PALABRA OOV Y SE LE ASIGNA PROB.