

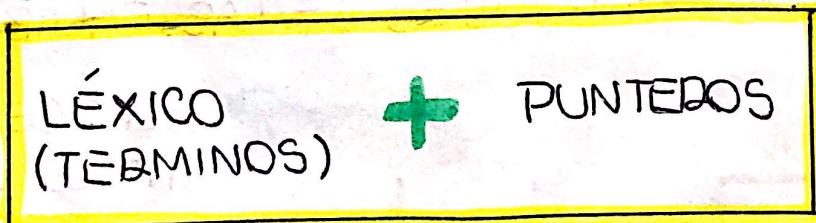
INFORMATION RETRIEVAL

INFORMATION RETRIEVAL ES LA RAMA DE LA COMPUTACIÓN QUE SE ENCARGA DE RECUPERAR INFORMACIÓN ALMACENADA EN COMPUTADORAS, MEDIANTE CONSULTAS. PARA ESTO EXISTE UNA ESTRUCTURA DE INDICE QUE NOS PERMITE RESOLVER LAS CONSULTAS, LLAMADO INDICES INVERTIDOS.

INDICES INVERTIDOS

COMO YA DIJIMOS IR ES UNA ESTRUCTURA QUE REFERENCIA NUESTROS CONJUNTOS DE DOCUMENTOS. CONTIENE TODOS LOS TÉRMINOS QUE NOS RESULTAN INTERESANTES. DADO UN TÉRMINO NOS PERMITE BUSCAR QUE DOC'S. LO CONTIENEN. ESTE INDICE TIENE QUE ESTAR ORDENADO PARA PODER HACER UNA BUSQUEDA BINARIA Y ADEMÁS TIENE QUE TENER LONGITUD FIJA.

ESTRUCTURA DEL INDICE



COMO DIJIMOS, NECESITAMOS GENERAR UNA ESTRUCTURA DE TAMAÑO FIJO. PARA PODER BUSCAR.

A CONTINUACIÓN VAMOS A VER LAS DISTINTAS MANERAS DE ALMACENAR LOS LÉXICOS Y PUNTEROS.

ALMACENAMIENTO

LÉXICO

CONCATENADO
FRONT CODING

PUNTEROS

DISTANCIAS
CODIGOUNARIO
CODIGO GAMMA
CODIGO DELTA

ALMACENAMIENTO LEXICO CONCATENADO

PARA DESARROLLAR EL EJEMPLO VAMOS A USAR LOS SIGUIENTES DOCUMENTOS

DOC 1 - PAPA PAPEL PEPA

DOC 2 - PAPEL PAPELON

DOC 3 - PAPPKA PIMENTON

DOC 4 - PAPA PIMENTA PIMENTON

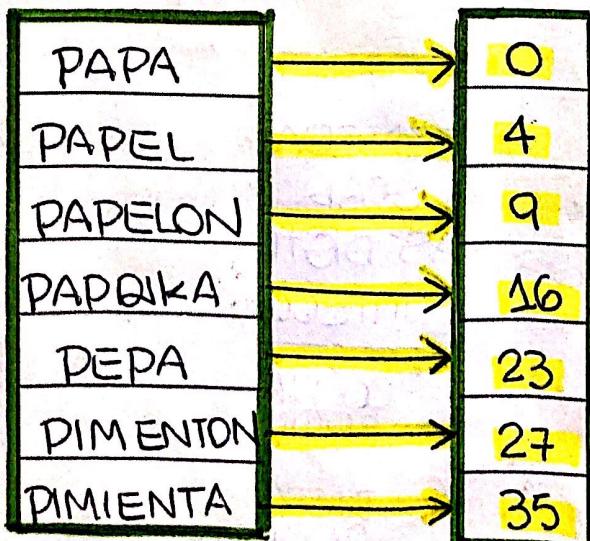
•1 ORDENAMOS LAS PALABRAS ALFABETICAMENTE

PAPA
PAPEL
PAPELON
PAPPKA
PEPA
PIMENTON
PIMENTA

•2 VAMOS A CONCATENAR TODAS LAS PALABRAS (TERMINOS)

PAPAPAPELPAPELONPAPPKAPEPAPIMENTONPIM...
0 4 9 16 23 27 35

→ EN NUESTRO INDICE NOS GUARDAMOS EL INDICE CON EL QUE COMIENZA LA PALABRA LO PODEMOS OBTENER SUMANDO LAS LONGITUDES DE LAS PALABRAS.



ASÍ CREAMOS NUESTRO INDICE LEXICO CON TAMAÑO FIJO

FRONT CODING

ESTE MECANISMO SE UTILIZA CUANDO ES NECESARIO COMPROBIR MAS EL LEXICO. NOS AyUDA A DEDUCIR EL ALMACENAMIENTO.

FRONT CODING TIENE EN CUENTA LA PARTICULARIDAD QUE LA LISTA ESTA ORDENADA ALFABETICAMENTE Y SEGUDAMENTE LOS TERMINOS VECINOS SUELEN COMPARTIR LOS PRIMEROS CARACTERES.

• 1 ORDENAMOS ALFABETICAMENTE

PAPA
PAPEL
PAPELON
PAPRIKA
PEPA
PIMENTON
PIMENTA

CHAR =	CHAR ≠	POS
0	4	0
3	2	4
5	2	6
3	4	8
1	3	12
1	7	15
3	5	22

• 2 • 3 • 4

• 2 EN ESTA COLUMNA INDICAMOS CUANTOS CARACTERES TIENE EN COMUN LA PALABRA CON LA DE ARRIBA. (LA PRIMERA SIEMPRE EMPIEZA EN 0)

PAPA - PAPEL - PAPELON
COINCIDEN 3 COINCIDEN 5

• 3 EN ESTA COLUMNA INDICAMOS CUANTOS CARACTERES DISTINTOS TIENE LA PALABRA (LOS DESTANTES DE LONG(TOTAL) - LONG (CHAR =))

PAPA - PAPEL - PAPELON
4 3 2 ≠ 5 2 ≠

• 4 CONCATENAMOS LAS PALABRAS PERO ESTA VEZ SOLO LO QUE ES DISTINTO. EN LA COLUMNA COLOCAMOS LOS COMIENZOS DE LAS PALABRAS DISTINTAS.

PAPAELONLRIKAEPAMIENTONIENTA
0 4 6 8 12 15 22

FRONT CODING PARCIAL

REALIZA LO MISMO QUE ANTES, PERO AHORA CADA N TERMINOS SE DEINICIA.

PAPA
PAPEL
PAPELON
PAPRIKA
PEPA
PIMENTON
PIMIENTA

CHAR =	CHAR ≠	POS
0	4	0
3	2	4
5	2	6
0	7	8
1	3	15
1	7	18
0	8	25

USAMOS COMO EJEMPLO N = 3

CUANDO DEINICIO LA PALABRA SE PONE COMPLETA

- PAPATELONPAPPRIKAEPAIMENTONPIMIENTA
0 4 6 8 15 18 25

ALMACENAMIENTO PUNTEROS DISTANCIAS

PAPA
PAPEL
PAPELON
PAPPRIKA
PEPA
PIMENTON
PIMIENTA

DOCS	DIST	POS
1,4	1,3	0
1,2	1,1	2
2	2	4
3	3	5
1	1	6
4	4	7
3,4	3,1	8

- 1 OBtenemos ALFABETICAMENTE COMO SIEMPRE

- 2 ESTA COLUMNA CONTIENE LOS DOCS EN LOS QUE APARECE EL TERMINO

- 3 ESTA COLUMNA CONTIENE LA DISTANCIA ENTRE LOS DOCUMENTOS

- 4 ESTA COLUMNA TENEMOS EL COMIENZO DE CADA PALABRA DISTINTA INDICANDO LOS DOCS

13 11 23 14 31
0 2 4 5 6 7 8

14 → 13
POQUE EL DOC 1 ESTA A DISTANCIA 3 DEL DOC 4.

123 → 111

CODIGOS UNARIOS

MANTIENE LA MISMA IDEA QUE LAS DISTANCIAS, PERO CUANDO CONCENAMOS LAS DISTANCIAS LAS VAMOS A CODIFICAR CON CODIGO UNARIO

UNARIO DE $N = N - 1$ CEDOS Y 1 UNO

PAPA
PAPEL
PAPELON
PAPPILKA
PEPA
PIMENTON
PIMENTA

DOCS	DIST	POS
1,4	1,3	0
1,2	1,1	4
2	2	6
3	3	8
1	1	11
4	4	12
3,4	3,1	16

→ MANTIENE LOS MISMOS PASOS QUE EL ANTERIOR PERO PHODA LA CODIFICACIÓN ES UNARIA.

1 3 1 1 2 3 1 4 3 1

ESTE CONCATENADO LO PASAMOS A UNARIO

- $U(1) = 1 - 1$ CEDOS Y UN 1 = 1
- $U(3) = 3 - 1$ CEDOS Y UN 1 = 001
- $U(2) = 01$ Y ASÍ CON TODOS ...

1 001 1 1 01 001 1 0001 0011
0 4 6 8 11 12 16

- ESTOS OCUPAN N BITS.
- ES MUY EFICIENTE CUANDO LAS DISTANCIAS SON MUY PEQUEÑAS.

CODIGOS GAMMA

NUEVAMENTE SOLO CAMBIA LA FORMA DE CODIFICAR LAS DISTANCIAS.

$$\gamma(N) \begin{cases} \lceil \log_2(N) \rceil + 1 \text{ EN UNARIO} \\ N - 2^{\lceil \log_2(N) \rceil} \text{ EN BINARIO} \\ \text{DE } \lceil \log_2(N) \rceil \text{ BITS} \end{cases}$$

ENTONCES SI TENIAMOS 1311231431
LO CODIFICAMOS EN CODIGO GAMMA.

$$\bullet \gamma(1) \begin{cases} \lceil \log_2(1) \rceil + 1 = 0 + 1 = 1 \text{ ESTO EN UNARIO} = 1 \\ 1 - 2^0 = 0 \text{ ESTO EN BINARIO} = 0 \\ \text{DE } 1 \text{ BIT} \rightarrow \text{COMO SON CEDO BITS} \\ \text{NO CUENTA.} \end{cases}) \text{ UNO RESULT}$$

$$\bullet \gamma(3) \begin{cases} \lceil \log_2(3) \rceil + 1 = 1 + 1 = 2 \text{ ESTO EN UNARIO} = 01 \\ 3 - 2^1 = 1 \text{ ESTO ES EN BINARIO DE } 1 \text{ BIT} = 1 \end{cases})$$

ENTONCES NOS QUEDA DE LA FORMA:

1 01111010011100100 01110111
0 4 6 9 12 13 18

0
4
6
9
12
13
18

TRUCO

$\gamma(x)$ 1- ESCRIBO x EN BINARIO

2- LONG(x) = N
AGREGO N-1 CEDOS ADELANTE

$\gamma(5)$ 101 TIENE LONG=3
 $\gamma(5) = 000101$

CODIGOS DELTA

OTRA VEZ SOLO CAMBIA LA FORMA DE CODIFICAR LAS DISTANCIAS.

$$\delta(N) \left\{ \begin{array}{l} \lfloor \log_2(N) \rfloor + 1 \text{ EN GAMMA} \\ N - 2^{\lfloor \log_2(N) \rfloor} \text{ EN BINARIO.} \\ \text{DE } \lfloor \log_2(N) \rfloor \text{ BITS} \end{array} \right.$$

ENTONCE SI TENIAMOS 1311231431.
LO CODIFICAMOS EN CODIGO DELTA.

- $\delta(1)$ $\left\{ \begin{array}{l} \lfloor \log_2(1) \rfloor + 1 = 0 + 1 = 1 \text{ ESTO LO ESCRIBO EN GAMMA 1} \\ 1 - 2^0 = 0 \text{ EN BINARIO DE 1 BIT} \end{array} \right.$

- $\delta(3)$ $\left\{ \begin{array}{l} \lfloor \log_2(3) \rfloor + 1 = 1 + 1 = 2 \text{ ESTO LO ESCRIBO EN GAMMA 010} \\ 3 - 2^1 = 1 \text{ EN BINARIO DE 1 BIT 1} \end{array} \right. \text{ LOS UNO } 0101$

ENTONCES NOS QUEDA DE LA FORMA:

$\frac{1}{1}$	$\frac{3}{0101}$	$\frac{1}{1}$	$\frac{1}{1}$	$\frac{2}{0100}$	$\frac{3}{0101}$	$\frac{1}{1}$	$\frac{4}{01100}$	$\frac{3}{01011}$	$\frac{1}{1}$
0	5	7	11	15	16	21			

0
5
7
11
15
16
21

FINALMENTE ELIGIENDO EL METODO ADECUADO DE ALMACENAMIENTO PARA EL LEXICO Y PUNTEROS, EL INDICE INVERTIDO QUEDA.

	TERM	DOCS	
0	PAPA	1,3	0
4	PAPEL	1,1	4
9	PAPELON	2	6
16	PAPRIKA	3	9
23	DEPA	1	12
27	DIMENTON	4	13
35	PIMIENTA	3,1	18

CONCATENADO

CODIGO GAMA

RESUMIENDO LA CONSTRUCCIÓN

- 1) RECOBREMOS LOS DOCUMENTOS A INDEXAR PARA OBTENER LOS TERMINOS
- 2) ABMAMOS UN LISTADO TERMINO DOCUMENTO
- 3) ORDENAMOS LA LISTA ALFABETICAMENTE POR TERMINO.
- 4) GENERAMOS EL LISTADO DE TERMINOS CONCATENADOS Y APLICAMOS CONCATENADO O' FRONT CODING
- 5) PASAMOS LA REFERENCIAS DE DOCUMENTOS A DISTANCIA
- 6) CODIFICAMOS LAS DISTANCIAS, APLICANDO CODIGO UNARIO GAMMA O DELTA.
- 7) GENERAMOS EL LISTADO DE DISTANCIAS CONCATENADAS.
- 8) GENERAMOS EL INDICE PRINCIPAL, CON UN PUNTERO AL TERMINO EN LA LISTA DE TERMINOS CONCATENADOS Y UN PUNTERO A LOS PUNTEROS DE DOCS, CODIFICADOS.

CONSULTAS

COMO DIJIMOS ANTES LOS INDICES INVERTIDOS SON UNA ESTRUCTURA QUE REFERENCIA NUESTROS CONJUNTO DE DOCS. UNA VEZ REALIZADO SE VA A UTILIZAR PARA HACER DISTINTOS TIPOS DE CONSULTAS UTILIZANDO BUSQUEDA BINARIA.

BUSQUEDA BINARIA

DADO UN VECTOR ORDENADO CON VALORES ÚNICOS PERMITE BUSCAR UN VALOR DENTRO DE EL EN O().

COMENZAMOS PARTIENDO EL VECTOR A LA MITAD Y COMPARANDO SI NUESTRA CONSULTA ES MAYOR O MENOR AL VALOR EN EL MEDIO. DEPENDIENDO DE SI ES MAYOR O MENOR NOS QUEDAMOS CON EL NUEVO VECTOR (IZQ O DER) Y VOLVEMOS A REPETIR.

PUNTUALES

VAMOS A QUEZER BUSCAR UN TERMINO EN EL INDICE INVERTIDO QUE CREAMOS.

DADO LOS SIGUIENTES DOCUMENTOS :

- 1 ROJO VERDE AMARILLO
- 2 VERDE VERDE AZUL
- 3 AZUL AMARILLO VERDE
- 4 AMARILLO ROJO

TERM	DOCS
AMARILLO	1,2,1
AZUL	2,1
ROJO	1,3
VERDE	1,1,1



TERM	DOCS
O	0
8	5
12	9
16	13

- UTILIZAMOS CONCATENADO y CODIGOS GAMMA.

Q = "AZUL"

1) CALCULAMOS LA LONG DEL INDICE DE TERMINOS
LONG(TERM) = 4

2) VAMOS A LA MITAD DEL INDICE (2), EN CASO DE SER UN FLOAT NOS QUEDAMOS CON LA PARTE ENTERA.

3) ACCEDEMOS A LO QUE HAY EN LA POS 2 Y COMPARAMOS

- AMARILLO AZUL ROJO VERDE

8 12 16

TENEMOS EL 12



ROJO > AZUL? SI

ENTONCES NOS QUEDAMOS CON NUESTRO NUEVO VECTOR Y APPLICAMOS DE NUEVO.

1) CALCULO LONG DEL NUEVO VECTOR LONG(TERM) = 2

2) DIVIDO A LA MITAD, OBTENEMOS POS 1

3) ACCEDEMOS A LA POS 1

TENEMOS EL 8

- AMARILLO AZUL ROJO VERDE

8 12 16



AZUL > AZUL? SON IGUALES!

TENEMOS LO QUE BUSCABAMOS

4) COMO LO QUE BUSCABAMOS ESTABA EN LA POS 1 DEL INDICE DE TERMINOS, AHORA VAMOS A LA POS 1 DEL INDICE DE DOCS, PARA VER EN QUE DOCUMENTOS ESTA AZUL.

- EN LA POS 1, TENEMOS 5 ENTONCES

1010101011011111
0 5 9 13

↑ 0101 010 = 2 }
1 = 1 }
RECORDEMOS QUE }
ESTA EN DISTANCIAS }

AZUL ESTA EN EL DOC 2, DOC 3,

BOOLEANAS

LAS CONSULTAS PUNTUALES NOS PERMITEN RESOLVER CONSULTAS BOOLEANAS.

$Q = \text{'AMARILLO \& AZUL'}$

1) DESOLVEMOS LAS CONSULTAS **PUNTUALES** PARA CADA TERMINO

- AMARILLO ESTA EN LOS DOCS 1,3,4
- AZUL ESTA EN LOS DOCS 2,3

2) REALIZAMOS LA OPERACIÓN BOOLEANA ENTRE LOS DOCUMENTOS

DOCS 1,3,4 **AND** DOC 2,3

DOC 3

PROXIMIDAD

LAS CONSULTAS DE PROXIMIDAD QUEDEMOS QUE LOS TERMINOS DE NUESTRA CONSULTA APAREZCAN EN LOS DOCS CERCAOS UNOS A OTROS. EN PARTICULAR UNA **FRASE** ES UNA CONSULTA CON PROXIMIDAD 1, ES DECIR QUE CADA TERMINO DEBE APARECER A CONTINUACIÓN DE OTRO.

PARA PODER DESOLVER ESTAS CONSULTAS, VAMOS A AGREGAR EN NUESTRO INDICE DE PUNTEROS, INFO SOBRE LAS POSICIONES DE CADA TERMINO EN CADA DOC

- DOC 1 = ROJO AZUL VERDE AZUL
- DOC 2 = VERDE AZUL AMARILLO
- DOC 3 = BLANCO VERDE BLANCO AZUL

$Q = \text{"VERDE AZUL"} \text{ FRASE}$

$Q = \text{"VERDE AZUL"} \sim 2 \text{ PROXIMIDAD}$
VERDE Y AZUL, VERDE O AZUL, ... CON DIST=2

TERM	DOCS
AMARILLO	2
AZUL	1,2,3
BLANCO	3
ROJO	1
VERDE	1,2,3

EL ALMACENAMIENTO LÉXICO LO VAMOS A HACER COMO SIEMPRE, LO QUE VA A CAMBIAR ES EL ALMACENAMIENTO DE PUNTEDOS. AHORA VAMOS A TENER QUE GUARDAR.

- DOCUMENTO

- FRECUENCIA DEL TÉRMINO EN DOC

- POSICIONES

→ LAS POSICIONES Y DOCS SE EXPRESAN EN DISTANCIA

AMARILLO: 2 1 2

0

1,2

AZUL: 1 2 1 3 2 1 1 3 1 3

3

1,2

BLANCO: 3 2 0 2 1 ESTAN A DIST 1 LOS DOCS

13

1

ROJO: 1 1 0

17

VERDE 1 1 2 2 1 0 3 1 1

20

DOCS
0
3
13
17
20

TODO ESTO LO CONCATENAMOS Y OBTENEMOS LOS NUEVOS ÍNDICES PARA EL ALMACENA DE PUNTEDOS

ESTO DEBE ESTAR EN CUALQUIER CODIFICACIÓN

PODEMOS DECIR QUÉ AMARILLO ESTÁ EN EL DOC 2 CON 1 SOLA APARICIÓN EN LA POS 2.

AZUL APARECE EN EL DOC 1 CON 2 APARICIONES EN LAS POS 1 Y POS 3. TMB ESTÁ EN EL DOC 2 CON UNA APARICIÓN EN LA POS 1 Y POR ÚLTIMO EN EL DOC 3 1 AP. EN POS 3.

ENTONCES ESTO ALMACENAMOS EN PUNTEDOS, Y DESOLVEMOS LA CONSULTA:

Q = "VERDE AZUL" (CONSULTA DE FASE)

1) DESOLVEMOS CONSULTA PUNTUAL DE CADA TÉRMINO

VERDE = 1 1 2 1 1 0 1 1 1

AZUL = 1 2 1 2 1 1 1 1 3

2) VEMOS QUE DOCUMENTOS TIENEN EN COMÚN, EN ESTE CASO LOS 3: DOC 1, DOC 2, DOC 3.

3) REVISAMOS LAS POS (PROXIMIDAD PARA CADA DOC)

DOC 1 = 1 1 2 → VERDE POS 2

1 2 1 2 → AZUL POS 1 POS 3) MATCH POS 2

DOC 3 = 1 1 1

1 1 3

NO HAY MATCH

DOC 2 = 1 1 0 → VERDE POS 0

1 1 1 → AZUL POS 1) MATCH POS 1

DOC 1 Y DOC 2

SIGNATURE FILES

EN LOS CASOS EN LOS CUALES UN INDICE INVERTIDO NO PUEDE USARSE POR NO HABER ESPACIO SUFFICIENTE PODEMOS RECUARDB A UN INDICE QUE TIENE UN COSTO MENOR EN ESPACIO QUE SON LOS SIG. FILES. ACA VAMOS A USAR FUNCIONES DE HASHING PARA OBTENER UNA REPRESENTACION DE CADA DOCUMENTO LA REPRESENTACION DE CADA DOC SUBGE DE LA REPRESENTACION MEDIANTE HASHING DE CADA TERMINO, A ESTO LO LLAMAMOS **SIGNATURE**.

SIGNATURE DE TERMINO

VAMOS A USAR EL DOC = "HOLA MUNDO".

- DECIDIMOS LOS VALORES DE K Y B.

- K** CANTIDAD DE FUNCIONES DE HASHING
- B** CANTIDAD DEL ESPACIO DE DIRECCIONES

USAMOS $K=2$, $B=8$

"HOLA" LE APLICAMOS LAS DOS FUNCIONES HASH

$$\left. \begin{array}{l} H_1(\text{HOLA}) = 3 \\ H_2(\text{HOLA}) = 0 \end{array} \right\} \text{ENCENDEMOS ESTOS BITS EN EL VECTOR}$$

$\begin{smallmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{smallmatrix}$
10010000

SIGNATURE DE 'HOLA'

"MUNDO"

$$\left. \begin{array}{l} H_1(\text{MUNDO}) = 0 \\ H_2(\text{MUNDO}) = 6 \end{array} \right\} \text{ENCENDEMOS ESTOS BITS EN EL VECTOR}$$

$\begin{smallmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{smallmatrix}$
10000010

SIGNATURE DE 'MUNDO'

SIGNATURE DE DOCUMENTO

EL SIGNATURE DEL DOCUMENTO (DOC = "HOLA MUNDO") ES EL **OR** DE LOS SIGNATURES DE TODOS SUS TERMS.

$$\left. \begin{array}{l} 10010000 \\ 10000010 \end{array} \right\} \begin{smallmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{smallmatrix} 10010010$$

SIGNATURE DEL DOCUMENTO

CONSULTA CON SIGNATURE FILE

$Q = "T"$ A NUESTRA CONSULTA LE APLICAMOS LAS K FUNCIONES DE HASHING Y OBTENEMOS SU SIGNATURE

AHORA TENEMOS QUE RECORRER TODOS LOS DOCS PARA LOS CUALES HAY UN 1 EN LAS POSICIONES PARA LAS CUALES HAY UN 1 EN NUESTRO $S(Q)$.

ESTOS DOCS SON CANDIDATOS, PORQUE PUEDE HABER **FALSOS POSITIVOS** YA QUE EN ALGUN DOCUMENTO LOS BITS PUEDEN HABER SIDO ENCENDIDOS POR OTROS TERMINOS (COLISION) DIFERENTE AL QUE BUSCAMOS.

→ PODEMOS MEJORAR LA EFICIENCIA SI EN VEZ DE ALMACENAR DOCUMENTO \times BITS LA ALMACENAMOS TRANSPISTA BITS \times DOCUMENTO

BITSLICES

ENTONCES SI TENEMOS $S(T) = 10010001$ PARA QUE UN DOC SEA CANDIDATO TIENE QUE TENER LOS BITS 0, 3 Y 7 PRENDIDOS, ENTONCES ACCEDEMOS A LAS FILAS 0, 3, 7 Y VEMOS QUE DOCS SON LOS QUE TIENEN LOS 3 BITS. EL USO DE BITSLICE IMPLICA ACCEDER UNICAMENTE A TANTOS SLICES COMO 1s TENGA EL SIGNATURE PRENDIDO.

CONSULTA BOOLEANA

$Q = "PLANE AND (CRASH OR ACCIDENT)"$ ANTES DE DEVOLVER EL RESULTADO DE Q , HAY QUE VERIFICAR SI

$S(\text{PLANE}) = 0100$
 $S(\text{CRASH}) = 1000$
 $S(\text{ACCIDENT}) = 1001$

REALMENTE ESTAN

ENTONCES BUSCAMOS DOCS QUE TENGAN PRENDIDO EL 2BIT SI OS Y QUE ADEMÁS TENGA EL 1BIT Y EL ULTIMO ENCENDIDO O NO

→ PARA QUE SIGNATURE FILE FUNCIONE CORRECTAMENTE LA CANTIDAD DE 1s Y 0s EN EL MISMO DEBE SER PADEJA BUSCAR UNA F. HASHING QUE EQUILIBRE

MUCHOS 1s **FALSOS POSITIVOS** MUCHOS 0s **DESPERDICIO DE ESPACIO**

BITMAPS

UN BITMAP ES UNA FORMA DE INDICE MUY PARTICULAR. TE NEMOS QUE TENER UNA ENTRADA POR TERMINO PERO LO QUE SEGURODA PARA CADA TERMINO ES UN VECTOR DE N BITS (SIENDO N LA CANTIDAD DE DOCUMENTOS) DONDE CADA BIT ES 0 O 1 SEGUN SE ENCUENTRE O NO EL TERMINO EN EL DOC.

DOCS

- PEDRO Y PABLO
- PEDRO CORRE
- PABLO RESPIRA
- PEDRO CORRE Y RESPIRA
- PEDRO CORRE PEDRO

EL BITMAP CORRESPONDIENTE ES :

TERM	D1	D2	D3	D4	D5
CORRE	0	1	0	1	1
PABLO	1	0	1	0	0
PEDRO	1	1	0	1	1
RESPIRA	0	0	1	1	0
Y	1	0	0	1	0

LOS BITMAPS SON MUY COMODOS PARA DESOLVER CONSULTAS. EN LAS CONSULTAS BOOLEANAS BASTA CON APLICAR ORs Y ANDs ENTRE LOS TERMINOS PARA OBTENER LOS DOCS BUSCADOS.

$Q = \text{PEDRO AND (CORRE OR RESPIRA)}$

$11011 \text{ AND } (01011 \text{ OR } 00110)$

$11011 \text{ AND } 01111$

$Q = 01011 \rightarrow$ BUSCAMOS QUE FILA CUMPLE CON ESTE QUERY Y LOS BITS ENCENDIDOS SON LOS DOCS

↓
DOC2, DOC4, DOCS

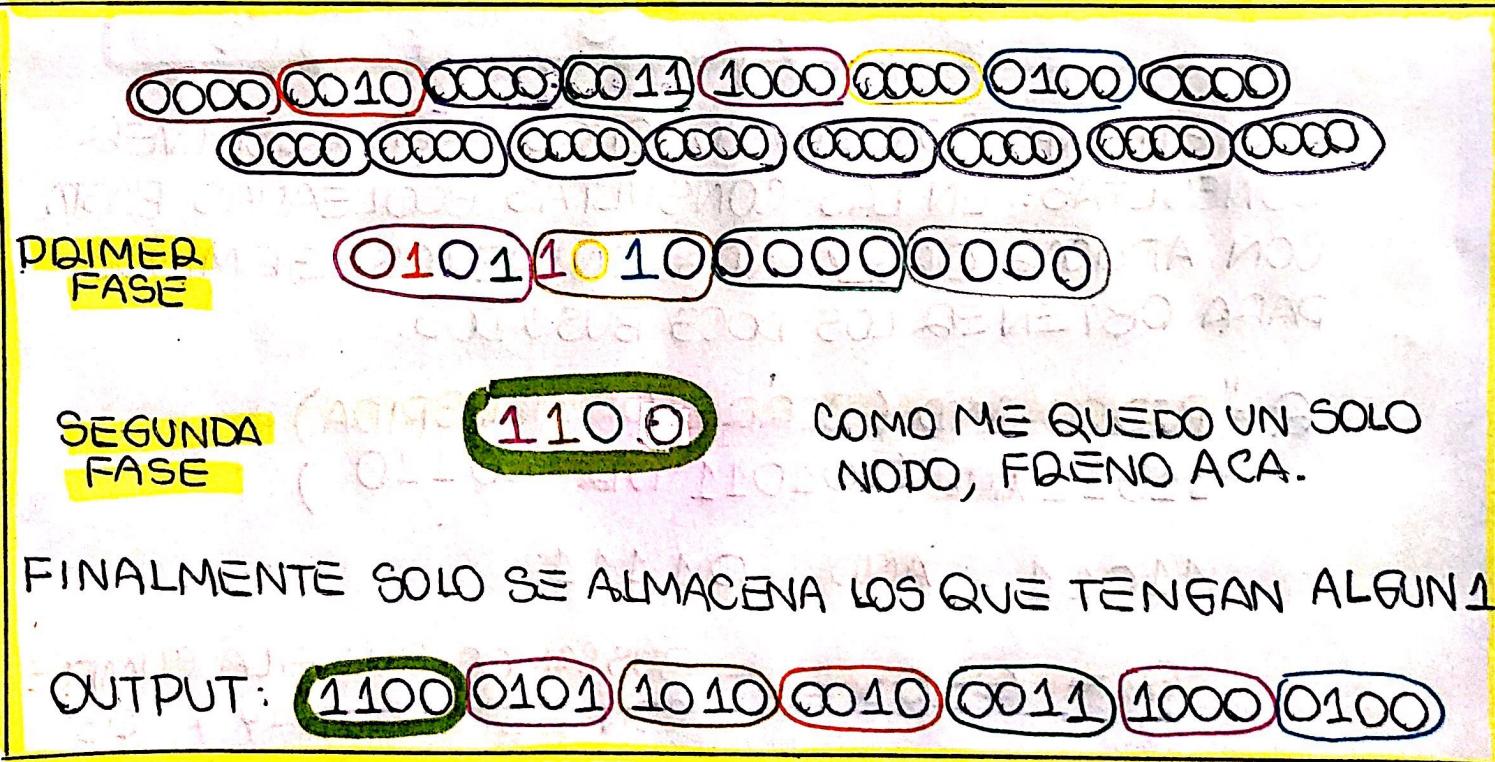
COMPRESIÓN DE BITMAPS

A PESAR DE QUE SON GENIALES PARA DESOLVER Q'S OCUPAN DEMASIADO ESPACIO, SI VAMOS A UTILIZARLOS HAY QUE COMPRIMIRLOS.

CADA ENTRADA EN UN INDICE DEL TIPO BITMAP TIENE LA PARTICULAR CARACTÉRISTICA DE ESTAR FORMADA POR ALGUNOS UNOS Y MUCHOS CEROS, PARA COMPRIMIRLOS VAMOS A USAR UN **ARBOL DE DERIVACIÓN BINARIA**. ESTE METODO SIGUE LOS SIGUIENTES PASOS :

- 1) TOMA N BITS Y GENERA UN BIT HIJO.
- 2) EL BIT HIJO ES 1 SI ALGUNO DE LOS N BITS DEL PADRE ES 1, SINO ES 0.
- 3) SE REPITE HASTA TENER UN SOLO NODO
- 4) SE ALMACENAN TODOS LOS NIVELES DEL ARBOL QUE TENGAN AL MENOS UN 1.

POR EJEMPLO TENEMOS UN VECTOR DE 64 BITS Y ELEGIMOS $N = 4$:



SI EN CUALQUIER MOMENTO LA CANTIDAD DE BITS DE UN NIVEL NO ES MULTIPLO DE N, EL ÚLTIMO NODO DEL NIVEL SUPERIOR NO TENDRA N BITS. COMPLETO CON CEROS

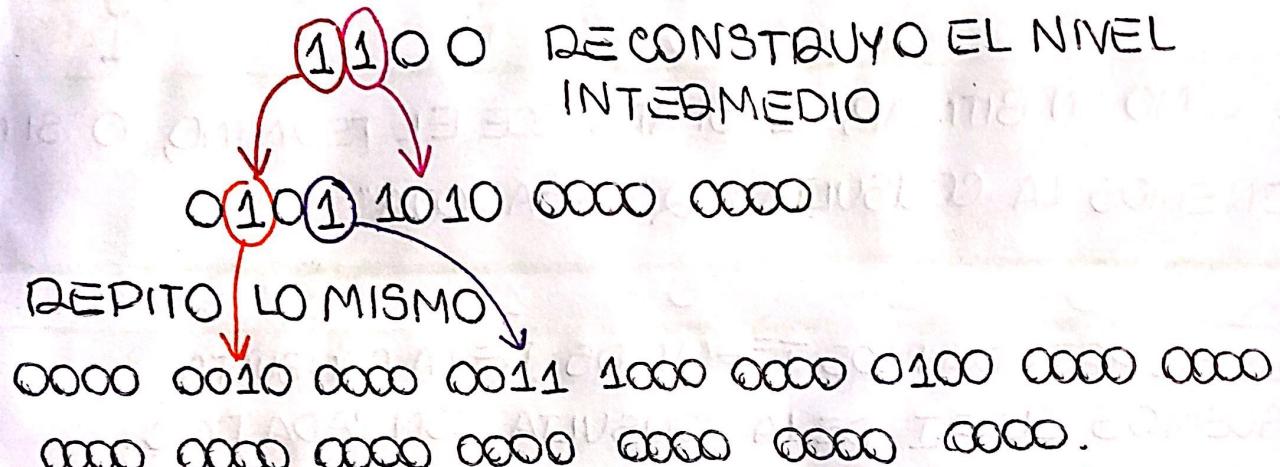
DECOMPRESSION DE BITMAPS⁹

DADA DESCOMPRIMIDA EL OUTPUT SE PROCEDA:

- UN BIT EN 1 INDICA QUE HAY UN NODO
 - UN BIT EN 0 INDICA QUE EL NODO ERA DE OS.

OUTPUT: 1100 0101 1010 0010 0011 1000 0100

COMIENZO POR PRIMER NODO.



CONSULTAS RANQUEADAS

LAS CONSULTAS BANQUEADAS SE TRATAN DE DADA UNA LISTA DE LOS DOCUMENTOS MAS RELEVANTES PARA LA CONSULTA ORDENADOS POR RELEVANCIA.

UNA PRIMERA APROXIMACIÓN ES CONTAR CUANTOS TERMINOS DE LA CONSULTA APARECEN EN CADA DOCUMENTO Y USAR ESE PUNTAJE PARA RANKEARLO.

ESTO ES EQUIVALENTE A REALIZAR EL PRODUCTO INTERNO ENTRE EL VECTOR CONSULTA Y EL VECTOR DE CADA DOCUMENTO USANDO EL MODELO BOW.

LA DESVENTAJA ES QUE NO TIENE EN CUENTA LA FRECUENCIA DEL TERMINO EN EL DOC, NI TAN POCO QUE TERMINO ES MAS PROBABLE O NO.

PARA PODER SOLUCIONAR ESTOS PROBLEMAS, VAMOS A UTILIZAR. **TF-IDF**

- DOC 1 = LA CASA ROSA
- DOC 2 = LA ROSA ROJA
- DOC 3 = LA MANZANA ROJA Y LA CASA AMARILLA

BOW

TENEMOS LA CONSULTA $Q = \text{"CASA ROJA"}$

	LA	CASA	ROSA	ROJA	MANZANA	Y	AMARILLA
D1	1	1	1	0	0	0	0
D2	1	0	1	1	0	0	0
D3	1	1	0	1	1	1	1

ES COMO UN BITMAP, 1 SI APARECE EL TERMINO, 0 SI NO.

TENEMOS LA CONSULTA: $Q = \text{"CASA ROJA"}$

$Q = 0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0$

SOLO SE PRENDEN LOS TERMINOS DE LA CONSULTA

HACEMOS EL PI DE LA CONSULTA CON CADA DOC.

- $Q \times D_1 = 0 + 1 + 0 + 0 + 0 + 0 + 0 = 1$
- $Q \times D_2 = 0 + 0 + 0 + 1 + 0 + 0 + 0 = 1$
- $Q \times D_3 = 0 + 1 + 0 + 1 + 0 + 0 + 0 = 2$

DANQUEAMOS
2
3
1

TF-IDF

PARA EL CASO ANTERIOR AHORA VAMOS A TENER EN CUENTA:

TF: CUENTA LA CANTIDAD DE VECES QUE APARECE CADA TERMINO DE LA CONSULTA EN CADA DOCUMENTO

IDF: LE DA A CADA TERMINO UN PESO INVERSAMENTE PROPORCIONAL A SU FRECUENCIA.

VA A SUCEDER QUE

$$\begin{array}{r} +TF \\ -TF \end{array} \quad \begin{array}{r} -IDF \\ +IDF \end{array}$$

$$IDF(T_i) = \log \left(\frac{N+1}{FT_i} \right)$$

N ES LA CANTIDAD DE DOCUMENTOS TOTALES

FT_i CANTIDAD DE DOCUMENTOS EN LOS QUE APARECE EL TERMINO

LOS TERM. QUE APARECEN EN MUCHOS DOCS SON MENOS IMPORTANTE QUE LOS TERM. QUE SOLO APARECEN EN POCOS DOCS.

DE ESTA FORMA PARA CADA TERMINO DE LA CONSULTA Y CADA DOCUMENTO MULTIPLICAMOS EL VALOR DE TF POR IDF Y ESO NOS DA EL PESO DEL TERMINO PARA ESE DOCUMENTO.

EL PUNTAJE DEL DOCUMENTO ES:

$$R(Q, D_j) = \sum FT_{ij} \log\left(\frac{N+1}{FT_i}\right)$$

DONDE FT_{ij} ES LA FRECUENCIA DEL TERMINO i EN EL DOCUMENTO j (CUANTAS VECES APARECE EN EL DOC)

EJEMPLO

- DOC1 = LA CASA DOSA
- DOC2 = LA ROSA ROJA MUY ROJA BIEN ROJA
- DOC3 = LA CASA ES ROJA.

1 CALCULAMOS TF Y IDF PARA CADA TERMINO

TF

FRECUENCIA DEL TERMINO
EN CADA DOCUMENTO

IDF

$\log\left(\frac{N+1}{FT_i}\right)$

	D1	D2	D3
LA	1	1	1
CASA	1	0	1
DOSAS	1	1	0
ROJA	0	3	1
ES	0	0	1
MUY	0	1	0
BIEN	0	1	0

0,12
0,30
0,30
0,30
0,60
0,60
0,60

- 'LA' = $\log\left(\frac{4}{3}\right) = 0,12$
- 'CASA' = $\log\left(\frac{4}{2}\right) = 0,30$
- 'BIEN' = $\log\left(\frac{4}{1}\right) = 0,60$

PODEMOS OBSERVAR
QUE MENOS APARICIÓN
MAS PESO TIENE.

AHORA VAMOS A RESOLVER
UNA CONSULTA.

Q = "CASA ROJA"

PARA DESOLVER LA CONSULTA APLICAMOS $R(Q, D_i)$

- $R(Q, D_i) = TF \times IDF$ PARA CADA TERMINO DE LA CONSULTA QUE ESTE EN EL DOC

- $R(Q, D_1) = 1 \times 0,3 = 0,3$

- $R(Q, D_2) = 3 \times 0,3 = 0,9$

- $R(Q, D_3) = 1 \times 0,3 + 1 \times 0,3 = 0,6$

} ③
① RANKING
②

POQUE EN EL
DOC3 APARECEN AMBOS TERMINOS.

RESOLVIENDO ESTA CONSULTA PODEMOS VER QUE EL DOC MAS IMPORTANTE SALE EN SEGUNDO LUGAR. ESTO SE DEBE A QUE TF TIENE DEMASIADO PESO EN LOS PUNTAJES DE CONSULTA. PARA ESTO PODRIAMOS REDUCIRLO CON DISTINTOS METODOS.

BM-25

UNA APROXIMACION MUY UTILIZADA PARA REDUCIR EL IMPACTO DE TF.

$$TF_{ij} = \frac{(k+1)FT_{ij}}{FT_{ij} + k}$$

DONDE k ES UN PARAMETRO QUE DEPENDE DE LA FREC. LOS MAS UTILIZADOS SON:

$k=2$ ó $k=1.2$

ENTONCES LA NUEVA FORMULA PARA DESOLVER LA CONSULTA

$$R(Q, D_i) = \sum \frac{(k+1)FT_{ij}}{FT_{ij} + k} \cdot \log \left(\frac{N+1}{FT_i} \right)$$

SI SEGUIMOS EL EJEMPLO ANTERIOR LOS VALORES DE IDF SE MANTIENEN IGUAL, LO QUE CAMBIA ES COMO CALCULO TF.

0,12
0,30
0,30
0,30
0,60
0,60
0,60

IDF

UTILIZAMOS $k = 2$

FT_{ij} FREC DEL TER.
EN CADA DOC.

TF_{ij} $TF_{ij} = \frac{(k+1) FT_{ij}}{FT_{ij} + k}$

	D1	D2	D3		D1	D2	D3
LA	1	1	1		1	1	1
CASA	1	0	1		1	0	1
ROSA	1	1	0		1	1	0
DOJA	0	3	1		0	1,8	1
ES	0	0	1		0	0	1
MUY	0	1	0		0	1	0
BIEN	0	1	0		0	1	0

$$\bullet TF_{3,2} = \frac{(2+1) \cdot 3}{3+2} = \frac{9}{5} = 1,8$$

$$\bullet TF_{0,1} = \frac{(2+1) \cdot 1}{1+2} = \frac{3}{3} = 1$$

PODEMOS OBSERVAR QUE CUANDO LA FREC ES 1, SE MANTIENE IGUAL.

ENTONCES PARA LA CONSULTA " CASA DOJA " QUEDA :

$$\bullet Q(Q, D_1) = 1 * 0,3 = 0,3 \quad (3)$$

BANKING

$$\bullet Q(Q, D_2) = 1,8 * 0,3 = 0,54 \quad (2)$$

$$\bullet Q(Q, D_3) = 1 * 0,3 + 1 * 0,3 = 0,6 \quad (1)$$

PODEMOS VER QUE SE SOLUCIONÓ NUESTRO BANKING PERO SIGUE OCURRIENDO QUE LOS DOCUMENTOS MÁS LARGOS TIENEN VENTAJA SOBRE LOS CORTOS . VAMOS A VER UNA SOLUCIÓN A ESTO.

NORMALIZACIÓN

OTRA APROXIMACIÓN MUY UTILIZADA PARA REDUCIR EL IMPACTO DE TF ES ESTA, VAMOS A NORMALIZAR LA LONGITUD DE LOS DOCUMENTOS PARA QUE NO QUEDA QUE LOS DOCS MÁS LARGOS TENGAN VENTAJA SOBRE LOS DOCS CORTOS .

$$\text{NORMALIZACIÓN} = 1 - B + B \cdot \frac{|D_j|}{AvL}$$

DONDE B ESTA $0 < B < 1$ Y EL VALOR MÁS UTILIZADO ES $B=0.75$. $|D_j|$ ES LA LONG DEL DOC j y AvL EL PROM.

ENTONCES TENEMOS EL NUEVO TF_{ij}

$$TF_{ij} = \frac{FT_{ij} (k+1)}{(FT_{ij} + k) \text{ NORM}}$$

Y ESTE NUEVO TF_{ij}
SE APLICA EN D(Q, D_j)

CONTINUANDO EL EJECUCIÓN ANTERIOR:

- DOC 1 = LA CASA DOSA
- DOC 2 = LA DOSA DOJA MUY DOJA BIEN DOJA
- DOC 3 = LA CASA ES DOJA

PARA LA NORMALIZACIÓN

$$\text{LONG}(D_1) = 3 \quad \left. \begin{array}{l} \text{EL PROMEDIO} \\ \text{AVDL} = \frac{3+7+4}{3} \end{array} \right\}$$

$$\text{LONG}(D_2) = 7 \quad \left. \begin{array}{l} \text{AVDL} = \frac{3+7+4}{7} \end{array} \right\}$$

$$\text{LONG}(D_3) = 4 \quad \left. \begin{array}{l} \text{AVDL} = \frac{3}{4,66} \end{array} \right\}$$

ENTONCES:

$$N_1 = 1 - 0,75 + 0,75 \frac{3}{4,66}$$

$$N_2 = 1 - 0,75 + 0,75 \frac{7}{4,66}$$

$$N_3 = 1 - 0,75 + 0,75 \frac{4}{4,66}$$

$$N_1 = 0,73 \quad N_2 = 1,38 \quad N_3 = 0,89$$

IDF

TF	D1	D2	D3
LA	1,37	0,72	1,12
CASA	1,37	0,00	1,12
DOSA	1,37	0,72	0,00
DOJA	0,00	1,30	1,12
ES	0,00	0,00	1,12
MUY	0,00	0,72	0,00
BIEN	0,00	0,72	0,00

0,12
0,30
0,30
0,30
0,60
0,60
0,60

- $T_{32} = \frac{(2+1) \cdot 1}{(3+2) \cdot 1,38} = \frac{9}{6,9} = 1,30$ \rightarrow FDEC EN EL DOC.
- $T_{03} = \frac{(2+1) \cdot 1}{(1+2) \cdot 0,89} = 1,12$
- $T_{01} = \frac{(2+1) \cdot 1}{(1+2) \cdot 0,73} = \frac{3}{2,19} = 1,37$
- $T_{62} = \frac{(2+1) \cdot 1}{(1+2) \cdot 1,38} = \frac{3}{4,14} = 0,72$

ENTONCES CUANDO DESOLVEMOS

$Q = \text{'CASA ROJA'}$

$$Q(Q, D_1) = 1,37 * 0,3 = 0,41 \quad (2)$$

$$Q(Q, D_2) = 1,30 * 0,3 = 0,39 \quad (3)$$

$$Q(Q, D_3) = 1,12 * 0,3 + 1,12 * 0,3 = 0,67 \quad (1)$$

RANKING

EVALUACIÓN DE CONSULTAS

EN ESTA SECCIÓN ANALIZAREMOS DE QUE FORMA EVALUAR EL FUNCIONAMIENTO DE UN SIST DE CONSULTAS BANQUEADAS O COMO COMPARAR DOS SISTEMAS Y DECIDIR CUAL ES MEJOR.

PODEMOS MEDIR LA PRECISIÓN Y EL RECALL SI INTÉRVENIMOS DICENDO QUE DOCUMENTOS SON RELEVANTES PARA LA CONSULTA

DOCS RELEVANTES	A	B
DOCS NO RELEVANTES	C	D
	DOCS RECUPERADOS	DOCS NO RECUPERADOS

$$\text{PRECISION} = \frac{A}{A+C}$$

$$\text{RECALL} = \frac{A}{A+B}$$

SI COMBINAMOS AMBAS CON $B = 1$:

$$F_B = \frac{(B^2 + 1) PQ}{B^2 P + Q}$$

PRECISION INTERPOLADA