

# LSH

LSH ES UNA SOLUCIÓN MUY EFICIENTE AL PROBLEMA DE LOS VECINOS MAS CERCANOS. EN PARTICULAR BUSCAMOS HALLAR DADO UN DETERMINADO PUNTO, CUALES SON LOS VECINOS MAS CERCANOS CON UNA DETERMINADA DISTANCIA.

EL CONCEPTO DETRÁS DE LSH ES MUY SIMPLE: APLICAR UNA FUNCIÓN DE HASHING A LOS PUNTOS DE FORMA TAL QUE LOS PUNTOS QUE SON PARECIDOS (CERCANOS) COLISIONEN. DE ESTA FORMA, CADA VEZ QUE BUSQUEMOS ALGO ('QUERY') OBTENDREMOS UN BUCKET CON TODOS LOS ELEMENTOS QUE COLISIONAN.

$$H(D_1, D_2, P_1, P_2)$$

ESTA FUNCIÓN REPRESENTA UNA LSH.

- $D_1$ : DISTANCIA MAX. PARA CONSIDERAR 2 OBJETOS SIMILARES.
- $D_2$ : DISTANCIA MÍN. PARA CONSIDERAR 2 OBJETOS MUY DISTINTOS.
- $P_1$ : PROBABILIDAD QUE 2 OBJETOS SIMILARES COLISIONEN EN EL MISMO BUCKET.
- $P_2$ : PROBABILIDAD QUE 2 OBJETOS DISTINTOS COLISIONEN EN EL MISMO BUCKET.

## CARACTERIZACIÓN

- FUNCIÓN DE HASHING PUEDE PROCESAR CUALQUIER PUNTO EN  $\mathbb{R}^d$  DIM. Y DAR NOS COMO RESULTADO EL NÚMERO DE BUCKET.
- ELEMENTOS SIMILARES CAIGAN EN UN MISMO BUCKET Y ELEMENTOS DISTINTOS NO CAIGAN EN UN MISMO BUCKET.
- COMPLEJIDAD  $O(1)$



PARA LA FUNCIÓN DE LSH, FIJAMOS  $D_2 = C * D_1$  PARA UNA CIERTA CONSTANTE  $C$ , Y QUEREMOS QUE  $P_1 > P_2$ . CUANTO MAYOR ES LA DIFERENCIA ENTRE  $P_1$  Y  $P_2$ , MEJOR SERÁ LA FUNCIÓN DE LSH.

$$f = \frac{\log 1/P_1}{\log 1/P_2}$$

CON ESTA FÓRMULA PODREMOS COMPARAR FUNCIONES DE LSH Y MAXIMIZAR SU VALOR.

## MINHASHES

LLAMAMOS A UNA FUNCIÓN DE HASHING, 'MINHASH' SI CUMPLE QUE LA PROBABILIDAD DE QUE DOS CLAVES COLISIONEN DEPENDA DE LA DISTANCIA ENTRE LAS CLAVES

$$P[H(x) == H(y)] = f(\|x - y\|)$$

## CONDICIONES

- COMO MENCIONAMOS ANTES, QUE LA PROBABILIDAD DEPENDA DE LA DISTANCIA.
- LA FUNCIÓN DEBE SER MONOTONA Y CONTINUA. A MEDIDA QUE LAS CLAVES ESTÉN MAS CERCA, LA PROB DE COLISIONAR SEA MAYOR
- DEBE SER EFICIENTE
- DEBE SER POSIBLE PARAMETRIZARLA, DE FORMA TAL DE PODER OBTENER MUCHOS MH. DIFERENTES PARA LA MISMA CLAVE. PODEMOS DECIR QUE EL MH DEFINE UNA FAMILIA
- EN GENERAL DESEAMOS QUE LA FUNCIÓN SEA DECREciente EN FUNCION A LA DISTANCIA. A MEDIDA QUE LA DISTANCIA ES MAYOR, LA PROBABILIDAD DE QUE DOS ELEMENTOS COLISIONEN SEA MENOR.

POR LO TANTO PODEMOS ESCRIBIR UNA FAMILIA GENÉTICA PARA CUALQUIER MH DE LA FORMA.

$$H(D_1, D_2, P(D_1), P(D_2))$$

$$\left. \begin{array}{l} \text{SIENDO } P(D_1) = 1 - D_1 \\ P(D_2) = 1 - D_2 \end{array} \right\}$$

SI NORMALIZAMOS LAS DISTANCIAS ENTRE 0 y 1.

## AMPLIFICACIÓN DE FAMILIAS

**FALSO POSITIVO:** LLAMAMOS POSITIVOS A LOS CASOS QUE COLISIONAN  
ENTONCES, UN FALSO POSITIVO ES CUANDO HAY UNA COLISIÓN ENTRE ELEMENTOS NO SIMILARES.

**FALSO NEGATIVO:** LLAMAMOS NEGATIVOS A LOS CASOS QUE NO COLISI..  
ENTONCES, UN FALSO NEGATIVO ES CUANDO NO HAY COLISIÓN ENTRE DOS ELEMENTOS SIMILARES.

### EJEMPLO

PARA OBJETOS A DIST = 0,2 NUESTRA PROB DE COLISIÓN = 0,8. ES DECIR, QUE A PARTIR DE HASHEAR UN QUERY PODEMOS RECUPERAR UN 80% DE LOS OBJETOS QUE ESTAN A DISTANCIA 0,2 O <. EL 20% DE OBJETOS QUE SON CERCANOS PERO NO RECUPERAMOS SON. **F.N.** (20% → 1 - P1)

POK OTRO LADO, SI NUESTROS OBJETOS ESTAN A DIST = 0,6 O >, ENTONCES LA PROBABILIDAD DE RECUPERABLOS ES 0,4. SI CONSIDERAMOS QUE OBJETOS A DIST 0,6 O >, NO SON CERCANOS, IGUAL VAMOS A RECUPERAR UN 40% DE ELLOS. ESTOS OBJETOS QUE NO SON CERCANOS PERO COLISIONAN SON **F.P.** A ESTOS OBJETOS DEBEMOS EXAMINABLOS Y DESCARTABLOS (40% → P2)

NOTAR QUE SE CUMPLE QUE )

$$P1 = 1 - D_1 ; P2 = 1 - D_2$$



# REDUCCIÓN DE FALSOS POSITIVOS

ES POSIBLE REDUCIR SU CANTIDAD USANDO MÁS DE UNA FUNCIÓN DE MH POR TABLA.<sup>↑R</sup> ES DECIR, VAMOS A PEDIR QUE

LOS DOS ELEMENTOS COLISIONEN EN MÁS DE UN MINHASH. POR EJEMPLO SI TENEMOS  $R=3$ , SIENDO Q LA CANT DE MH POR TABLA, LE APLICAMOS A  $\times 3$  MH DISTINTOS  $H_1(x) = 21 \quad H_2(x) = 30 \quad H_3(x) = 5$  (21, 30, 5 SON DISTINTOS BUCKETS) SI RECUPERAMOS LOS ELEMENTOS DENTRO DE CADA BUCKET Y HACEMOS UNA (INTERSECCIÓN) OBTENEMOS LOS MÁS CERCAOS A NUESTRO QUERY. LA PROBABILIDAD DE COLISIONAR EN LOS R BUCKETS =  $P^R$ . **💡** QUE ESTO AUMENTA LA CANTIDAD DE FALSOS NEGATIVOS.

# REDUCCIÓN DE FALSOS NEGATIVOS

LA REDUCCIÓN DE ESTOS IMPLICA PODER RECUPERAR MÁS ELEMENTOS.

LA IDEA CONSISTE EN USAR MÁS DE UNA TABLA DE HASH.<sup>↑B</sup> SI USAMOS B TABLAS PODEMOS CONSIDERAR OBJETOS CANDIDATOS A AQUELLOS QUE SON CANDIDATO EN (ALGUNA) TABLA.

LA PROBABILIDAD DE QUE NO COLISIONEN EN NINGUNO ES  $(1 - P)^B$ , ENTONCES DE QUE COLISIONEN  $1 - (1 - P)^B$

**💡** ESTO AUMENTA LA CANTIDAD DE FALSOS POSITIVOS.

# COMBINANDO AMBAS COSAS

PODEMOS LOGRAR UN NÚMERO TAN BAJO COMO DESEMOS DE FALSOS POSITIVOS Y NEGATIVOS USANDO AMBOS MÉTODOS COMBINADOS, ES DECIR B TABLAS DE HASH Y R MH POR TABLA. CON UNA CANTIDAD TOTAL DE MH DE  $B * R$ .

ENTONCES LA PROBABILIDAD DE COLISIONAR:

$$P_{R,B} = 1 - (1 - P^R)^B$$

# ENCONTRANDO LOS VALORES DE BY R

EL OBJETIVO ES ENCONTRAR VALORES DE BY R TALES QUE PARA D1 TENGA UN VALOR DE P1 (DESEADO) Y PARA D2 TENGAMOS UN VALOR DE P2 (DESEADO) MEDIANTE **GRID - SEARCH** VAMOS A BUSCAR ESTOS VALORES.



ESTE MÉTODO CONSISTE IR AUMENTANDO EL VALOR DE R, HASTA CONSEGUIR LA P2 DESEADA.

$$P_{R,B_2} = 1 - \left( 1 - (1 - D_2)^R \right)^B \rightarrow \text{SE MANTIENE}$$

CUANDO LLEGAMOS AL P2 DESEADO, VAMOS A VER QUE P1 NO ES EL VALOR DESEADO ENTONCES, VAMOS AUMENTANDO B HASTA LLEGAR AL P1 DESEADO.

CUANDO LO ENCONTRAMOS, P2 SE MOVIÓ ENTONCES VOLVEMOS A REPETIR HASTA CUMPLIR AMBAS P.

$$P_{R,B_1} = 1 - \left( 1 - (1 - D_1)^R \right)^B \rightarrow \text{VA AUMENTANDO}$$

SE MANTIENE CON EL VALOR DEL PASO ANTERIOR

## ¿COMO OPERA LSH?

VAMOS A MOSTRAR COMO OPERA EL LSH CON UN EJEMPLO A PEQUEÑA ESCALA PARA ENTENDER EL FUNCIONAMIENTO.

EN ESTE EJEMPLO VAMOS A UTILIZAR B=2, R=2 (2 BANDAS Ó TABLAS Y 2 MH PARA CADA TABLA, ES DECIR 4 MH EN TOTAL)

ADEMÁS VAMOS A TENER 4 ELEMENTOS.



VAMOS A SUPONER QUE LOS MH, DEVUELVEN NÚMEROS ENTRE 0 Y 3. ESTO SIGNIFICA QUE LAS 2 TABLAS VAN A TENER 4 BUCKETS

	MH1	MH2	MH3	MH4
A	1	1	0	2
B	0	0	1	0
C	1	3	1	0
D	0	0	1	0

COMO  $B=2$  Y  $R=2$  SIGNIFICA QUE TIENE 2 TABLAS CON 2 MH CADA UNA.

	MH1	MH2
A	1	1
B	0	0
C	1	3
D	0	0

	MH3	MH4
A	0	2
B	1	0
C	1	0
D	1	0

LOS VALORES EN ESTAS TABLAS SON POSICIONES / INDICES. NOS UBICAN CADA ELEMENTO HASHEADO EN ELLA.

TABLA 1

B,D	A,C		C
0	1	2	3

TABLA 2

A, <sup>D</sup> <sub>B,C</sub>	BCD	A	
0	1	2	3

AHORA SE COMPLETABON PONIENDO EL ELEMENTO EN DONDE INDICA, SI AMBOS MH MANDAN AL MISMO ELEMENTO AL MISMO LUGAR, SE PONE UNA VEZ

AHORA TENEMOS UN QUERY, QUE AL APLICARLE LOS MH DEVUELVE  $0 \quad 0 \quad | \quad 0 \quad 1$   
 $MH_1 \quad MH_2 \quad | \quad MH_3 \quad MH_4$

- BUSCAMOS EN T1 LO QUE HAY EN 0:

$$Q_1 = \{B, D\}$$

- BUSCAMOS EN T2 LO QUE HAY EN 0:

$$Q_2 = \{A, B, C, D\} \cap \{B, C, D\} = \{B, C, D\}$$

EL RESULTADO SERIA  $Q_1 \cup Q_2 = \{D, B\} \cup \{B, C, D\} = \{B, C, D\}$

# DISTANCIA DE JACCARD

$$SJ(D_1, D_2) = \frac{D_1 \cap D_2}{D_1 \cup D_2}$$

PARA HACER LA INTERSECCIÓN Y LA UNIÓN VAMOS A UTILIZAR LOS **SHINGLES** DE LOS DOCS.

$$0 < SJ < 1$$

$$DJ(D_1, D_2) = 1 - SJ$$

## CONSTRUCCIÓN DE MH

EN ESTA SECCIÓN DESARROLLAREMOS UNA FUNCIÓN MIN HASH PARA DOCUMENTOS DE TEXTO BASADA EN LA DISTANCIA DE JACCARD



1 → EMPEZAMOS PROCESANDO CADA DOCUMENTO OBTENIENDO TODOS LOS **N-GRAMAS** DE N CARACTERES.

A CADA ELEMENTO ÚNICO DEL N-GRAMA LO LLAMAREMOS **SHINGLES**.

UN EJEMPLO DOC 1 → "HOLA", OBTENGAMOS LOS BIGRAMAS ( $N=2$ ) → { \$H, HO, OL, LA, A\$ }, COMO NINGUNO SE REPITE SON LOS SHINGLES.

2 → VAMOS A CREAR UNA TABLA DONDE LAS COLUMNAS SON LOS DOCUMENTOS Y LAS FILAS SON LOS SHINGLES.

EN CADA POSICIÓN UN '1' IMPLICA QUE ESE SHINGLE SE ENCUENTRA EN EL DOCUMENTO, CASO CONTRARIO, APARECE UN '0'.

3 → REORDENAMOS LAS FILAS AL AZAR, Y EL MH DE CADA DOCUMENTO ES LA POSICIÓN EN LA QUE APARECE EL PRIMER 1.

→ SI VOLVEMOS A PERMUTAR LAS FILAS AL AZAR REPITIENDO EL PASO 3, OBTENEMOS UN NUEVO MH EN LA FAMILIA DE FUNCIONES.

VAMOS A REALIZAR UN EJEMPLO CON LOS SIGUIENTES DOCS.

	d1	d2	d3	d4
s1	1	0	1	0
s2	1	0	0	1
s3	0	1	0	1
s4	0	1	0	1
s5	0	1	0	1
s6	1	0	1	0
s7	1	0	1	0

1 ESTE PASO YA FUE REALIZADO CON EL FIN DE AGILIZAR EL EJEMPLO.

2 ACA TENEMOS CREADA LA TABLA CON 1 Y 0

3 REORDENAMOS LAS FILAS AL AZAR Y GENERAMOS EL PRIMER MH, CON LA POS DEL PRIMER 0.

	d1	d2	d3	d4
s5	0	1	0	1
s6	1	0	1	0
s1	1	0	1	0
s2	1	0	0	1
s7	1	0	1	0
s4	0	1	0	1
s3	0	1	0	1

	MH1	MH2	MH3
d1	1		
d2	0		
d3	1		
d4	0		

EL PRIMER MH GENERADO (SON LAS POS DEL PRIMER 0)

4 VOLVEMOS A REPETIR EL PASO ANTERIOR 2 VECES MAS, PARA OBTENER 2 MH MAS Y PODER FINALMENTE APLICARLE LSH

	d1	d2	d3	d4
s3	0	1	0	1
s2	1	0	0	1
s4	0	1	0	1
s1	1	0	1	0
s7	1	0	1	0
s5	0	1	0	1
s6	1	0	1	0

	MH1	MH2	MH3
d1	1	1	
d2	0	0	
d3	1	3	
d4	0	0	

A SIMPLE VISTA PODEMOS VER QUE D1 Y D3 COINCIDEN 2/3

	MH1	MH2	MH3
d1	1	1	0
d2	0	0	1
d3	1	3	0
d4	0	0	1

CON ESTA TABLA PODEMOS APLICAR LSH PARA VER DOCS SIMILARES.

COMO SON LOS INDEXOS ANTES.

	d1	d2	d3	d4
s1	1	0	1	0
s5	0	1	0	1
s2	1	0	0	1
s7	1	0	1	0
s6	1	0	1	0
s4	0	1	0	1
s3	0	1	0	1

17	4551-1202	Irma M C de J R d
4775-0618	4541-7760	Juan C F de M
4854-3691	4571-4299	Jorge M Ur
4866-1738	4587-0002	Jorge B A
4813-1892	4593-3785	Juan
4832-7841	4603-0128	Juan P
4556-1380	4636-3785	Juan P
4521-8309	4653-3785	Juan P
4824-9674	4672-1152	Juan P
4703-2815	4673-1152	Juan P
4812-5878	4674-1152	Juan P
4804-4253	4675-1152	Juan P
4552-2698	4676-1152	Juan P
4817-0155	4677-1152	Juan P
4815-7198	4678-1152	Juan P
4925-3672	4679-1152	Juan P
4781-3070	4680-1152	Juan P
4824-0261	4681-1152	Juan P
4883-2013	4682-1152	Juan P
4864-8827	4683-1152	Juan P
4842-2232	4684-1152	Juan P
4843-1084	4685-1152	Juan P
4861-0877	4686-1152	Juan P
RIELLO E	4551-1202	Irma M C de J R d
RIELLO A	4541-7760	Juan C F de M
RIELLO AM	4571-4299	Jorge M Ur
RIELLOS JO	4587-0002	Jorge B A
RIELLO /	4593-3785	Juan
RIELLO /	4603-0128	Juan P
RIELLO /	4636-3785	Juan P
RIELLO /	4653-3785	Juan P
RIELLO /	4672-1152	Juan P
RIELLO /	4673-1152	Juan P
RIELLO /	4674-1152	Juan P
RIELLO /	4675-1152	Juan P
RIELLO /	4676-1152	Juan P
RIELLO /	4677-1152	Juan P
RIELLO /	4678-1152	Juan P
RIELLO /	4679-1152	Juan P
RIELLO /	4680-1152	Juan P
RIELLO /	4681-1152	Juan P
RIELLO /	4682-1152	Juan P
RIELLO /	4683-1152	Juan P
RIELLO /	4684-1152	Juan P
RIELLO /	4685-1152	Juan P
RIELLO /	4686-1152	Juan P
RIELLO /	4687-1152	Juan P
RIELLO /	4688-1152	Juan P
RIELLO /	4689-1152	Juan P
RIELLO /	4690-1152	Juan P
RIELLO /	4691-1152	Juan P
RIELLO /	4692-1152	Juan P
RIELLO /	4693-1152	Juan P
RIELLO /	4694-1152	Juan P
RIELLO /	4695-1152	Juan P
RIELLO /	4696-1152	Juan P
RIELLO /	4697-1152	Juan P
RIELLO /	4698-1152	Juan P
RIELLO /	4699-1152	Juan P
RIELLO /	4700-1152	Juan P
RIELLO /	4701-1152	Juan P
RIELLO /	4702-1152	Juan P
RIELLO /	4703-1152	Juan P
RIELLO /	4704-1152	Juan P
RIELLO /	4705-1152	Juan P
RIELLO /	4706-1152	Juan P
RIELLO /	4707-1152	Juan P
RIELLO /	4708-1152	Juan P
RIELLO /	4709-1152	Juan P
RIELLO /	4710-1152	Juan P
RIELLO /	4711-1152	Juan P
RIELLO /	4712-1152	Juan P
RIELLO /	4713-1152	Juan P
RIELLO /	4714-1152	Juan P
RIELLO /	4715-1152	Juan P
RIELLO /	4716-1152	Juan P
RIELLO /	4717-1152	Juan P
RIELLO /	4718-1152	Juan P
RIELLO /	4719-1152	Juan P
RIELLO /	4720-1152	Juan P
RIELLO /	4721-1152	Juan P
RIELLO /	4722-1152	Juan P
RIELLO /	4723-1152	Juan P
RIELLO /	4724-1152	Juan P
RIELLO /	4725-1152	Juan P
RIELLO /	4726-1152	Juan P
RIELLO /	4727-1152	Juan P
RIELLO /	4728-1152	Juan P
RIELLO /	4729-1152	Juan P
RIELLO /	4730-1152	Juan P
RIELLO /	4731-1152	Juan P
RIELLO /	4732-1152	Juan P
RIELLO /	4733-1152	Juan P
RIELLO /	4734-1152	Juan P
RIELLO /	4735-1152	Juan P
RIELLO /	4736-1152	Juan P
RIELLO /	4737-1152	Juan P
RIELLO /	4738-1152	Juan P
RIELLO /	4739-1152	Juan P
RIELLO /	4740-1152	Juan P
RIELLO /	4741-1152	Juan P
RIELLO /	4742-1152	Juan P
RIELLO /	4743-1152	Juan P
RIELLO /	4744-1152	Juan P
RIELLO /	4745-1152	Juan P
RIELLO /	4746-1152	Juan P
RIELLO /	4747-1152	Juan P
RIELLO /	4748-1152	Juan P
RIELLO /	4749-1152	Juan P
RIELLO /	4750-1152	Juan P
RIELLO /	4751-1152	Juan P
RIELLO /	4752-1152	Juan P
RIELLO /	4753-1152	Juan P
RIELLO /	4754-1152	Juan P
RIELLO /	4755-1152	Juan P
RIELLO /	4756-1152	Juan P
RIELLO /	4757-1152	Juan P
RIELLO /	4758-1152	Juan P
RIELLO /	4759-1152	Juan P
RIELLO /	4760-1152	Juan P
RIELLO /	4761-1152	Juan P
RIELLO /	4762-1152	Juan P
RIELLO /	4763-1152	Juan P
RIELLO /	4764-1152	Juan P
RIELLO /	4765-1152	Juan P
RIELLO /	4766-1152	Juan P
RIELLO /	4767-1152	Juan P
RIELLO /	4768-1152	Juan P
RIELLO /	4769-1152	Juan P
RIELLO /	4770-1152	Juan P
RIELLO /	4771-1152	Juan P
RIELLO /	4772-1152	Juan P
RIELLO /	4773-1152	Juan P
RIELLO /	4774-1152	Juan P
RIELLO /	4775-1152	Juan P
RIELLO /	4776-1152	Juan P
RIELLO /	4777-1152	Juan P
RIELLO /	4778-1152	Juan P
RIELLO /	4779-1152	Juan P
RIELLO /	4780-1152	Juan P
RIELLO /	4781-1152	Juan P
RIELLO /	4782-1152	Juan P
RIELLO /	4783-1152	Juan P
RIELLO /	4784-1152	Juan P
RIELLO /	4785-1152	Juan P
RIELLO /	4786-1152	Juan P
RIELLO /	4787-1152	Juan P
RIELLO /	4788-1152	Juan P
RIELLO /	4789-1152	Juan P
RIELLO /	4790-1152	Juan P
RIELLO /	4791-1152	Juan P
RIELLO /	4792-1152	Juan P
RIELLO /	4793-1152	Juan P
RIELLO /	4794-1152	Juan P
RIELLO /	4795-1152	Juan P
RIELLO /	4796-1152	Juan P
RIELLO /	4797-1152	Juan P
RIELLO /	4798-1152	Juan P
RIELLO /	4799-1152	Juan P
RIELLO /	4800-1152	Juan P
RIELLO /	4801-1152	Juan P
RIELLO /	4802-1152	Juan P
RIELLO /	4803-1152	Juan P
RIELLO /	4804-1152	Juan P
RIELLO /	4805-1152	Juan P
RIELLO /	4806-1152	Juan P
RIELLO /	4807-1152	Juan P
RIELLO /	4808-1152	Juan P
RIELLO /	4809-1152	Juan P
RIELLO /	4810-1152	Juan P
RIELLO /	4811-1152	Juan P
RIELLO /	4812-1152	Juan P
RIELLO /	4813-1152	Juan P
RIELLO /	4814-1152	Juan P
RIELLO /	4815-1152	Juan P
RIELLO /	4816-1152	Juan P
RIELLO /	4817-1152	Juan P
RIELLO /	4818-1152	Juan P
RIELLO /	4819-1152	Juan P
RIELLO /	4820-1152	Juan P
RIELLO /	4821-1152	Juan P
RIELLO /	4822-1152	Juan P
RIELLO /	4823-1152	Juan P
RIELLO /	4824-1152	Juan P
RIELLO /	4825-1152	Juan P
RIELLO /	4826-1152	Juan P
RIELLO /	4827-1152	Juan P
RIELLO /	4828-1152	Juan P
RIELLO /	4829-1152	Juan P
RIELLO /	4830-1152	Juan P
RIELLO /	4831-1152</td	

UNA PROPIEDAD QUE SE PUEDE DEMOSTRAR DE ESTE MH ES QUE LA PROBABILIDAD DE QUE DOS DOCUMENTOS COINCIDAN ES IGUAL A LA SEMEJANZA DE JACCARD.

$$\text{IP}[\text{MH}(D_1) = \text{MH}(D_2)] = S_j(D_1, D_2)$$

## DISTANCIA ANGULAR

ES UNA METRICA MUY POPULAR PARA CUANDO LOS DATOS QUE TENEMOS SE REPRESENTAN COMO UN VECTOR.

LA DISTANCIA QUE NOS INTERESA ENTRE LOS VECTORES ES EL ANGULO ENTRE ELLOS. EL COSENO ENTRE LOS VECTORES SIRVE COMO MEDIDA DE SEMEJANZA.

$$\cos \theta = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} \xrightarrow{\text{NORMALIZAMOS}} \cos \theta = \mathbf{x} \cdot \mathbf{y}$$

- DOS VECTORES QUE TIENEN IGUAL DIRECCIÓN TIENEN  $\text{D.A.} = 0$  Y  $\cos \theta = 1$ , MIENTRAS QUE DOS VECTORES CON DIRECCIONES OPUESTAS TIENEN  $\text{D.A.} = 180$  Y  $\cos \theta = -1$ .

PARA APROXIMAR LA D.A. USANDO LSH EXISTEN VARIOS MH POSIBLES

## CONSTRUCCIÓN DE MH CON EL METODO DE LOS HIPERPLANOS

PARA GENERAR UN MH VAMOS A CALCULAR EL PRODUCTO INTERNO ENTRE NUESTRO VECTOR Y UN VECTOR ALEATORIO. ESTE VECTOR ALEATORIO SOLO PUEDE SER UN **SKETCH**, ES DECIR SOLO PUEDE CONTENER LOS VALORES  $1$  o  $-1$ .

EL VALOR DEL MH VA A SER  $1$  SI EL PRODUCTO INT. ES POSITIVO Y  $-1$  SI EL RESULTADO ES NEGATIVO.

VAMOS A HACER UN EJEMPLO: DADO LOS VECTORES Y LOS HIPERPLANOS ALEATORIOS

$$v_1 = [4 \ 4 \ -5 \ -2 \ 3]; v_2 = [-3 \ -2 \ -4 \ 5 \ 0]; v_3 = [3 \ 2 \ -1 \ -2 \ 1]. \quad r_1 = [1 \ 1 \ 1 \ 1 \ -1]; r_2 = [-1 \ 1 \ 1 \ -1 \ -1]; r_3 = [1 \ -1 \ -1 \ -1 \ -1]; \quad r_4 = [1 \ -1 \ -1 \ -1 \ 1]; r_5 = [1 \ -1 \ -1 \ -1 \ 1]; r_6 = [-1 \ 1 \ 1 \ 1 \ 1].$$

## • 1 CALCULAMOS EL PRODUCTO INTERNO DE CADA VECTOR CON CADA HIPERPLANO.

$$\begin{aligned} [4 \ 4 \ -5 \ -2 \ 3] * [1 \ 1 \ 1 \ 1 \ -1] &= -2 \Rightarrow \text{minhash } -1 \\ [4 \ 4 \ -5 \ -2 \ 3] * [-1 \ 1 \ 1 \ -1 \ -1] &= -6 \Rightarrow \text{minhash } -1 \\ [4 \ 4 \ -5 \ -2 \ 3] * [1 \ -1 \ -1 \ -1 \ -1] &= 4 \Rightarrow \text{minhash } +1 \\ [4 \ 4 \ -5 \ -2 \ 3] * [1 \ -1 \ -1 \ -1 \ 1] &= 10 \Rightarrow \text{minhash } +1 \\ [4 \ 4 \ -5 \ -2 \ 3] * [1 \ -1 \ -1 \ -1 \ 1] &= 10 \Rightarrow \text{minhash } +1 \\ [4 \ 4 \ -5 \ -2 \ 3] * [1 \ -1 \ -1 \ -1 \ 1] &= 10 \Rightarrow \text{minhash } +1 \end{aligned}$$

$$\begin{aligned} [-3 \ -2 \ -4 \ 5 \ 0] * [1 \ 1 \ 1 \ 1 \ -1] &= -4 \Rightarrow \text{minhash } -1 \\ [-3 \ -2 \ -4 \ 5 \ 0] * [-1 \ 1 \ 1 \ -1 \ -1] &= -8 \Rightarrow \text{minhash } -1 \\ [-3 \ -2 \ -4 \ 5 \ 0] * [1 \ -1 \ -1 \ -1 \ -1] &= -2 \Rightarrow \text{minhash } -1 \\ [-3 \ -2 \ -4 \ 5 \ 0] * [1 \ -1 \ -1 \ -1 \ 1] &= -2 \Rightarrow \text{minhash } -1 \\ [-3 \ -2 \ -4 \ 5 \ 0] * [1 \ -1 \ -1 \ -1 \ 1] &= -2 \Rightarrow \text{minhash } -1 \\ [-3 \ -2 \ -4 \ 5 \ 0] * [1 \ -1 \ -1 \ -1 \ 1] &= -2 \Rightarrow \text{minhash } -1 \end{aligned}$$

## v1 CON TODOS LOS HIPER.

CON TODOS ESTOS RESULTADOS GENERAMOS LA TABLA.

	v1	v2	v3
r1	-1	-1	1
r2	-1	-1	-1
r3	1	-1	1
r4	1	-1	1
r5	1	-1	1
r6	1	-1	1

ACA A SIMPLE VISTA PODEMOS VER QUE V1 Y V3 SON MUY PARECIDOS YA QUE COINCIDEN EN 5/6

MH que coinciden	Angulo
0/6	180
1/6	150
2/6	120
3/6	90
4/6	60
5/6	30
6/6	0

ENTONCES COMO TIENEN 5/6 SU DISTANCIA ANGULAR ES DE 30°.

## v2 CON TODOS LOS HIPER.

$$\begin{aligned} [3 \ 2 \ -1 \ -2 \ 1] * [1 \ 1 \ 1 \ 1 \ -1] &= 1 \Rightarrow \text{minhash } +1 \\ [3 \ 2 \ -1 \ -2 \ 1] * [-1 \ 1 \ 1 \ -1 \ -1] &= -1 \Rightarrow \text{minhash } -1 \\ [3 \ 2 \ -1 \ -2 \ 1] * [1 \ -1 \ -1 \ -1 \ -1] &= 3 \Rightarrow \text{minhash } +1 \\ [3 \ 2 \ -1 \ -2 \ 1] * [1 \ -1 \ -1 \ -1 \ 1] &= 5 \Rightarrow \text{minhash } +1 \\ [3 \ 2 \ -1 \ -2 \ 1] * [1 \ -1 \ -1 \ -1 \ 1] &= 5 \Rightarrow \text{minhash } +1 \\ [3 \ 2 \ -1 \ -2 \ 1] * [1 \ -1 \ -1 \ -1 \ 1] &= 5 \Rightarrow \text{minhash } +1 \end{aligned}$$

## v3 CON TODOS LOS HIPER.

CON ESTA TABLA YA PODEMOS APLICAR LSH, PERO AL REALIZARLO VAMOS A NOTAR UNA LEVE DIFERENCIA. USAMOS B=2 Y R=3.

	v1	v2	v3
r1	0	0	1
r2	0 1	0 0	0 5
r3	1	0	1

## BANDA 1

LOS -1 FUERON PASADOS A 0.

	v1	v2	v3
r4	1	0	1
r5	1 7	0 0	1 7
r6	1	0	1

## BANDA 2

LA DIFERENCIA ES QUE LOS 3 BITS NOS INDICAN LA POSICION EN LA TABLA EN BINARIO → TABLA 2 BITS POS.

$\sqrt{2}$	$\sqrt{1}$					$\sqrt{3}$
0	1	2	3	4	5	7

↓ 8

$\sqrt{2}$						$\sqrt{3}$
0	1	2	3	4	5	7

BL

BL

ENTONCES LA PROBABILIDAD DE QUE DOS VECES PROYECTADOS AL AZAR SOBRE UN HIPERPLANO TENGAN DIFERENTE SIGNO ES  $\theta/180^\circ$  YA QUE WANTO MAS CHICO ES EL ANGULO MAS PROBABLE ES QUE LAS PROYECCIONES TENGAN IGUAL SENTIDO.

LA PROBABILIDAD DE QUE LOS MH COINCIDAN PARA DOS VECTORES ES  $1 - \theta/180^\circ$   
NOS QUEDA DEFINIDA UNA NUEVA FAMILIA DE LSH

$$H(D_1, D_2, 1 - \frac{D_1}{\pi}, 1 - \frac{D_2}{\pi})$$

D<sub>1</sub> Y D<sub>2</sub> SON DIST. ANGULARES

POZ ÚLTIMO, PODEMOS DECIR QUE CADA MH NOS DEFINE UN BIT Y CALCULAMOS UN MH DE K-BITS, USANDO K Hiperplanos. EL MH NOS DEVUELVE VALORES ENTRE 0 - 2<sup>K</sup>, SIENDO ESTE EL VALOR DE CADA UNA DE LAS TABLAS.

→ A MAYOR VALOR DE K, AUMENTAN LOS FALSOS NEGATIVOS YA QUE NECESITAMOS QUE LOS PUNTOS ESTEN DEL MISMO LADO DE LOS K Hiperplanos USADOS PARA QUE EL VALOR DEL MH COINCIDA.

LA DISTANCIA ANGULAR SE PUEDE ESTIMAR EN FUNCIÓN DE LA CANTIDAD DE MH QUE COINCIDEN (COMO SE VE EN LA TABLA EN EL EJEMPLO)

## DISTANCIA EUCLIDEANA

ESPACIO EUCLIDEO →

TIENE MULTIPLES DIMENSIONES Y UN VALOR REAL UBICA CADA PUNTO EN EL MISMO EN CADA UNAS DE ELLAS. LA DISTANCIA EUCLIDEA SE BASA EN LA POSICION DE LOS PUNTOS EN ESE ESPACIO.

CUALQUIER OTRO ESPACIO ES NO-EUCLIDEO Y SUS DISTANCIAS SE BASAN EN LAS PROPIEDADES DE LOS PUNTOS, PERO NO SU POSICIÓN EN EL ESPACIO.

↳ EJ JACCARD, ANGULAR.



# CONSTRUCCIÓN DE MH CON EL METODO DE LAS PROYECCIONES

LA IDEA ES SIMILAR A LA USADA PARA LA DISTANCIA ANGULAR. VAMOS A USAR PROYECCIONES ALEATORIAS, EN ESTE CASO LOS HP ALEATORIOS SUBENDE UNA DISTRIBUCIÓN P-ESTABLE. CADA VECTOR LO VAMOS A MULTIPLICAR POR UN HP PARA OBTENER UN NÚMERO. Y ESTE NÚMERO LO VAMOS A DISCRETIZAR A UN BUCKET DIVIDIENDO POR UN PARÁMETRO  $w$ . EL RESULTADO ES EL MH PARA LA DISTANCIA EUCLÉDIANA.

$$MH_i(x) = \left\lfloor \frac{x \cdot v_i + a}{w} \right\rfloor$$

ES DIFÍCIL HALLAR UNO BUENO, DONDE  $w$  DEPENDE DE LOS DATOS Y  $a$  ES UN NÚMERO ENTRE 0 Y  $w-1$ .  $v_i$  ES VECTOR ALEATORIO DE MISMA DIM QUE  $x$ .

EN CADA PROYECCIÓN LOS PUNTOS QUE ESTAN EN UNA MISMA FRANJA SON SIMILARES.  $w$  REGULA EL ANCHO DE LAS BANDAS. POR LO TANTO DEFINIMOS UNA FAMILIA DE LSH:

$$H\left(\frac{w}{2}, 2w, p \geq \frac{1}{2}, p \leq \frac{1}{3}\right)$$