

Guía 5: LSH

1. Tenemos la siguiente tabla representando el valor de 6 minhashes para tres documentos:

	D1	D2	D3
MH1	1	0	1
MH2	3	1	3
MH3	1	2	1
MH4	2	2	3
MH5	0	0	2
MH6	0	0	2

Se usa $b=2$ y $r=3$. Se decide usar 7 buckets para cada banda.

Encontrar una única función de hashing perteneciente a una flia universal $LSH(r1,r2,r3)$ de forma tal que en la primer banda solo D1 y D3 sean candidatos a ser similares pero en la segunda banda los tres documentos sean candidatos a ser similares.

2. Si se tiene la siguiente familia de funciones LSH $(0.15,0.85,0.85,0.15)$ indique de qué forma quedaría amplificada usando $r=3$ y $b=4$. Finalmente interprete el resultado de la familia amplificada indicando qué cantidad de falsos positivos o falsos negativos se producirían.
3. Usando LSH en una construcción de 5 ANDs y 3 Ors se observa que algunos pares de documentos que deberían ser candidatos a similares no lo son. ¿Qué cambiaría?
4. Se quiere aplicar LSH a un conjunto de documentos para encontrar los pares de documentos mas similares. Queremos que si $J(D1,D2) \geq 0.7$ entonces la probabilidad de que D1 y D2 sean candidatos sea ≥ 0.9 y queremos que si $J(D1,D2) \leq 0.5$ entonces la probabilidad de que sean candidatos sea ≤ 0.3 . Indique cuántas funciones minhash usaría y que combinación de AND y OR usaría para lograr lo pedido.
5. Se quiere construir una función LSH usando Jaccard que detecte aquellos documentos cuya semejanza esté entre 0.8 y 1.0. Vamos a pedir que si dos documentos tienen semejanza 0.9 o mayor la probabilidad de detectarlos sea 0.95 y que si dos documentos tienen semejanza 0.8 o menor la probabilidad de detectarlos sea inferior a 0.2. Construir la función LSH pedida usando la menor cantidad de funciones de hashing posible, indicar r y b . Reflexione sobre lo que pasó en este ejercicio.
6. Se tienen los siguientes puntos en el plano: (2,3) (3,4) (24,30) (21,32). Sean el siguiente vector al azar: (5,3). Indique cuál debería el valor de w para que al aplicar LSH para la distancia euclidean los puntos 1 y 2 sean semejantes pero los puntos 3 y 4 no.

7. Usamos LSH para encontrar documentos parecidos a un documento consulta. Desafortunadamente nuestra función LSH no está encontrando varios documentos que son similares a los buscados y esto es importante para nuestro problema. Indique diferentes formas de solucionar este problema.
8. Usando la distancia de Jaccard y 36 minhashes se quiere comparar el efecto de usar 6 construcciones OR y luego 6 AND contra usar primero 6 construcciones AND y luego 6 OR. ¿En qué casos tendremos mas falsos positivos y en que casos mas falsos negativos? Si fijamos $d_1=0.2$ y $d_2=0.5$ ¿cuál es la probabilidad de que dos documentos sean candidatos en cada caso?
9. Usamos LSH para la distancia de Jaccard para comparar frases breves usando 4-shingles con 6 funciones de hashing que agrupamos en 3 construcciones OR de 2 construcciones AND cada una. Queremos obtener los strings que sean al menos 80% semejantes a "use the force". Describa detalladamente todos los pasos necesarios para encontrar las frases que cumplan con lo pedido.
10. Dados los vectores: $x=[1,3,-1,2]$; $y=[-1,-2,-1,-1]$, $z=[2,4,-1,3]$ y los hiperplanos aleatorios: $r_1 = [+1,-1,+1]$ $r_2=[+1,+1,-1]$ $r_3=[-1,-1,-1]$, $r_4=[+1,+1,+1]$. Queremos usar 3 (tres) hiperplanos para aproximar el coseno entre los vectores usando LSH. ¿Cuáles son los 3 hiperplanos que hay que elegir entre los 4 propuestos? Justifique adecuadamente
11. Sean los siguientes vectores en 5 dimensiones: $v_1 = [4 \ 4 \ -5 \ -2 \ 3]$; $v_2 = [-3 \ -2 \ -4 \ 5 \ 0]$; $v_3=[3 \ 2 \ -1 \ -2 \ 1]$. Y sean los siguientes 6 hiperplanos aleatorios: $r_1=[1 \ 1 \ 1 \ 1 \ -1]$; $r_2=[-1 \ 1 \ 1 \ -1 \ -1]$; $r_3=[1 \ -1 \ -1 \ -1 \ -1]$; $r_4=[1 \ -1 \ -1 \ -1 \ 1]$; $r_5=[1 \ -1 \ -1 \ -1 \ 1]$; $r_6=[-1 \ 1 \ 1 \ -1 \ 1]$. Se pide comparar las alternativas $r=3$, $b=2$ vs $r=2$, $b=3$ indicando en cada caso que colisiones se producirían.
12. Si la probabilidad de que dos vectores colisionen usando un único hiperplano es mayor a 0.95.
 - a) ¿Cuál es el ángulo máximo entre los vectores?
 - b) De un ejemplo de un hiperplano para el cual dos vectores que están a la distancia indicada en el punto anterior no colisionen.