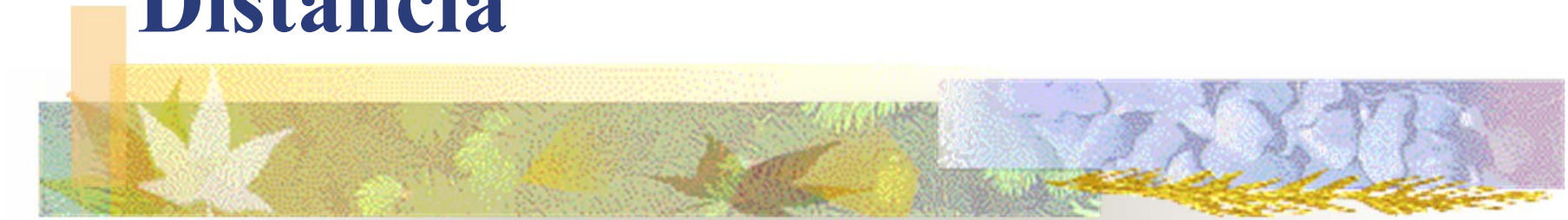
Algoritmos de Similaridad y Distancia



UCR – ECCI

CI-2414 Recuperación de Información

Prof. Kryscia Daviana Ramírez Benavides

Búsquedas en RI

- Dos opciones principales:
  - Búsquedas estructuradas.
  - Basadas en palabras clave: clasificación más sencilla.
    - Por palabra única
    - Por contexto
- *Modelo de Similaridad:* Soporte para realizar búsquedas basadas en patrones erróneos de digitación, deletreo, escaneo, etc.
- *Funciones de Distancia:* Establecen metodología y nivel de error permitido, unas miden si la hilera es aceptable o no y otras dan la similaridad.

Conceptos Generales

- *Noción de Distancia:* Comparar hileras y ver que diferencia hay entre ellas.
- *Modelos de Similaridad:* Calcular la cercanía en que están las hileras. Ayuda a búsquedas sobre patrones con errores.
- *Algoritmos de Distancia:* Funciones que permiten ver que tan diferentes son las hileras.

Propiedades

- Las funciones de distancia deben de cumplir con las siguientes condiciones:
  - $d(i,i) = 0$
  - $d(i, j) \geq 0$
  - $d(i, j) = d(j,i)$
  - $d(i, j) \leq d(i,h) + d(h, j)$

Distancia de Hamming

- Se le aplica a hileras del mismo tamaño:
  - Hileras de igual tamaño, vectores con la misma dimensión.
  - Hileras de diferente tamaño se rellena la hilera más pequeña hacia la derecha con vacíos.
- Podemos definir la distancia entre ellas como el número de caracteres en las mismas posiciones que son distintos, número de entradas diferentes entre los vectores.
- **Entre más cerca de cero más parecidas son las hileras**.

Distancia Hamming – Ejemplo #1

h1 = **GRANDE** ***h2*** = **GRANOS**

G	R	A	N	D	E
G	R	A	N	O	S

Distancia Hamming – Ejemplo #1

G	R	A	N	D	E
G	R	A	N	O	S
✓					

Distancia Hamming – Ejemplo #1

G	R	A	N	D	E
G	R	A	N	O	S
✓	✓				

Distancia Hamming – Ejemplo #1

G	R	A	N	D	E
G	R	A	N	O	S
✓	✓	✓			

Distancia Hamming – Ejemplo #1

G	R	A	N	D	E
G	R	A	N	O	S
✓	✓	✓	✓		

Distancia Hamming – Ejemplo #1

G	R	A	N	D	E
G	R	A	N	O	S
✓	✓	✓	✓	✗	

Distancia Hamming – Ejemplo #1

G	R	A	N	D	E
G	R	A	N	O	S
✓	✓	✓	✓	✗	✗

Nº de ✗ = *HD(h1,h2)* = 2

Distancia Hamming – Ejemplo #2

h1 = **DISTANCIA** ***h2*** = **DISTROSION**

D	I	S	T	A	N	C	I	A	
D	I	S	T	R	O	S	I	O	N

Distancia Hamming – Ejemplo #2

D	I	S	T	A	N	C	I	A	
D	I	S	T	R	O	S	I	O	N
✓									

Distancia Hamming – Ejemplo #2

D	I	S	T	A	N	C	I	A	
D	I	S	T	R	O	S	I	O	N
✓	✓								

Distancia Hamming – Ejemplo #2

D	I	S	T	A	N	C	I	A	
D	I	S	T	R	O	S	I	O	N
✓	✓	✓							

Distancia Hamming – Ejemplo #2

D	I	S	T	A	N	C	I	A	
D	I	S	T	R	O	S	I	O	N
✓	✓	✓	✓						

Distancia Hamming – Ejemplo #2

D	I	S	T	A	N	C	I	A	
D	I	S	T	R	O	S	I	O	N
✓	✓	✓	✓	✗					

Distancia Hamming – Ejemplo #2

D	I	S	T	A	N	C	I	A	
D	I	S	T	R	O	S	I	O	N
✓	✓	✓	✓	✗	✗				

Distancia Hamming – Ejemplo #2

D	I	S	T	A	N	C	I	A	
D	I	S	T	R	O	S	I	O	N
✓	✓	✓	✓	✗	✗	✗			

Distancia Hamming – Ejemplo #2

D	I	S	T	A	N	C	I	A	
D	I	S	T	R	O	S	I	O	N
✓	✓	✓	✓	✗	✗	✗	✓		

Distancia Hamming – Ejemplo #2

D	I	S	T	A	N	C	I	A	
D	I	S	T	R	O	S	I	O	N
✓	✓	✓	✓	✗	✗	✗	✓	✗	

Distancia Hamming – Ejemplo #2

D	I	S	T	A	N	C	I	A	
D	I	S	T	R	O	S	I	O	N
✓	✓	✓	✓	✗	✗	✗	✓	✗	✗

Nº de ✗ = *HD(h1,h2)* = 5

Distancia Hamming – Resumen Ejemplos

- Ejemplo 1:
  - *h1* = grande
  - *h2* = granos
  - *HD(h1,h2)* = 2
- Ejemplo 2:
  - *h1* = distancia
  - *h2* = distorsión
  - *HD(h1,h2)* = 5

Distancia de Levenstein

- Llamada también distancia de edición.
- Medida similaridad entre dos hileras:
  - Hilera fuente (*f*)
  - Hilera meta (*m*).
- La similaridad entre dos palabras se calcula como el mínimo número de operaciones que transforma a una de las palabras en la otra (transformar *f* en *m*).
- Las tres operaciones son:
  - destrucción()
  - inserción()
  - substitución()

Distancia de Levenshtein (cont.)

- A más distancia Levenshtein ~ Mayor diferencia en las hileras.
- **Entre más cerca de cero más parecidas son las hileras.**
- Algoritmo:
  - El tamaño de la hilera *f* es *x*, y el tamaño de la hilera *m* es *y*. Si $x = 0$, retornar *y*; y si $y = 0$, retornar *x*.
  - Construir una matriz con $y + 1$ filas y $x + 1$ columnas. Inicializar la primer fila de la matriz con la secuencia 0, 1, 2, …, *x*; y la primer columna de la matriz con la secuencia 0, 1, 2, …, *y*.
  - Colocar cada carácter de la hilera *f* en su correspondiente celda *i* (*i* va de 1 a *x*).
  - Colocar cada carácter de la hilera *m* en su correspondiente celda *j* (*j* va de 1 a *y*).

Distancia de Levenshtein (cont.)

- Algoritmo (cont.):
  - Si $f(i)$ es igual a $m(j)$ el costo de la celda es 0.
  - Si $f(i)$ es diferente de $m(j)$ el costo de la celda es 1.
  - El valor de la celda $d(i,j)$ es el mínimo de:
    - Valor de la celda $(i-1, j) + 1$.
    - Valor de la celda $(i, j-1) + 1$.
    - Valor de la celda $(i-1, j-1) + \text{costo}$.
  - La distancia es la celda $d(x,y)$.

Distancia de Levenshtein – Ejemplo

h1 = **CENA**

tam(*h1*) = 4

h2 = **COMA**

tam(*h1*) = 4

	0	1	2	3	4
0					
1					
2					
3					
4					

Distancia de Levenshtein – Ejemplo

		0	1	2	3	4
			C	O	M	A
1	C					
2	E					
3	N					
4	A					

Distancia de Levenshtein – Ejemplo

		0	1	2	3	4
			C	O	M	A
0						
1	C	0	1	1	1	1
2	E	1	1	1	1	1
3	N	1	1	1	1	1
4	A	1	1	1	0	

Distancia de Levenshtein – Ejemplo

		0	1	2	3	4
			C	O	M	A
0						
1	C					
2	E					
3	N					
4	A					

Distancia de Levenshtein – Ejemplo

		0	1	2	3	4
0			C	O	M	A
1	C	0	1	2	3	
2	E	1	1	2	3	
3	N	1	1	1	1	
4	A	1	1	1	0	

Distancia de Levenshtein – Ejemplo

	0	1	2	3	4
0		C	O	M	A
1	C	0	1	2	3
2	E	1	1	2	3
3	N	2	2	2	3
4	A	1	1	1	0

Distancia de Levenshtein – Ejemplo

		0	1	2	3	4
			C	O	M	A
0						
1	C		0	1	2	3
2	E		1	1	2	3
3	N		2	2	2	3
4	A		3	3	3	2

Distancia de Levenshtein – Ejemplo

		0	1	2	3	4
			C	O	M	A
0						
1	C		0	1	2	3
2	E		1	1	2	3
3	N		2	2	2	3
4	A		3	3	3	**2**

LD(h1,h2)

Distancia de Levenshtein – Resumen Ejemplo

- Intuitivamente:
  - Se tienen las hileras: *f* = coma y *m* = coma. *LD*(*f*,*m*) = 0, porque no hay que realizar transformaciones.
  - Se tienen las hileras: *f* = coma y *m* = cena. *LD*(*f*,*m*) = 2, porque con dos sustituciones se transforma *f* en *m*.
- Con el algoritmo:

	0	**1**	**2**	**3**	**4**
0		**c**	**o**	**m**	**a**
1	**c**	0	1	2	3
2	**e**	1	1	2	3
3	**n**	2	2	2	3
4	**a**	3	3	3	**2**

- http://www.cut-the-knot.org/do_you_know/Strings.shtml

Distancias con N-gramas

- La medida de similaridad puede establecerse mediante la fórmula definida como:

$$d = \frac{N * C}{A_1 + A_2 + \ldots + A_N}$$

  la cual se conoce como el coeficiente de Dice.

- Donde:
  - N = Número de gramas utilizados.
  - C = Número de digramas que comparten las hileras.
  - A_x = Número de digramas de una hilera.

Distancias con N-gramas (cont.)

- **Entre más cerca de cero más diferentes son las hileras**.
- Los valores están en el intervalo [0,1].
- Dos hileras: colaboración y colaborador, y haciendo N = 2:
  - Colaborador: co ol la ab bo or ra ad do or
    bigramas únicos: co ol la ab bo or ra ad do
  - Colaboración: co ol la ab bo or ra ac ci ió ón
    bigramas únicos: co ol la ab bo or ra ac ci ió ón
- La hilera *colaborador* tiene 10 bigramas, de los cuáles 9 son únicos; y la hilera *colaboración* tiene 11 bigramas, todos únicos. Comparten 7 bigramas: co ol la ab bo or ra.
- Aplicando la fórmula: $D(s_1, s_2) = \frac{2*7}{9+11} = 0.7$

Distancia con Bigramas

- Tomar cada par de caracteres juntos, contando la cantidad total de pares.
- Por palabra se cuentan solo los bigramas únicos.
- Ejemplo: *casaca*
  - Bigramas = ca as sa ac ca = 5
  - Bigramas únicos = ca as sa ac = 4
- Se cuentan bigramas comunes entre palabras (para medir distancia).
- $D(A,B)$ = 2 * bigramas comunes / (bigramas únicos A + bigramas únicos B).
- Si $D(A,B) = 1$ => palabras iguales.

Distancia con Bigramas – Ejemplo #1

- *h1* = escuela = es sc cu ue el la = 6
- *h2* = escuela = es sc cu ue el la = 6
- Bigramas únicos de *A* = 6
- Bigramas únicos de *B* = 6
- Bigramas comunes = 6
- 2 * 6 / (6 + 6) = 1
- *BiD*(*h1*,*h2*) = 1

Distancia con Bigramas – Ejemplo #2

- *h1* = escuela = es sc cu ue el la = 6
- *h2* = comidas = co om mi id da as = 6
- Bigramas únicos de *A* = 6
- Bigramas únicos de *B* = 6
- Bigramas comunes = 0
- 2 * 0 / (6 + 6) = 0
- *BiD*(*h1*,*h2*) = 0

Distancia Euclidiana

- Muy usada por su simplicidad
- Cálculo a partir de los valores numéricos de cada posición en un vector. Distancia entre vectores, se da peso a los caracteres.
- Sean *u* y *v* hileras de caracteres, entonces:

$$d(u,v) = \sqrt{\sum_{k=1}^{f} (u_k - v_k)^2} = \sqrt{|u_1 - v_1|^2 + \ldots + |u_f - v_f|^2}$$

- Con *f* = tamaño de la hilera más grande.
- **Entre más cerca de cero más parecidas son las hileras**.

Distancia Euclidiana – Ejemplo #1

- *h1* = escuela
- *h2* = escuelas
- Valor ASCII de los caracteres:
  - e = 101, s = 115, c = 99, u = 117, l = 108, a = 97, espacio en blanco = 32.

$$ED(h1,h2) = \sqrt{|101-101|^2 + |115-115|^2 + |99-99|^2 + |117-117|^2 + |101-101|^2 + |108-108|^2 + |97-97|^2 + |32-115|^2}$$

$$ED(h1,h2) = \sqrt{0+0+0+0+0+0+6889} = \sqrt{6889} = 83$$

- $ED(h1,h2) = 83$

Distancia Euclidiana – Ejemplo #2

- *h1* = escuela
- *h2* = comidas
- Valor ASCII de los caracteres:
  - e = 101, s = 115, c = 99, u = 117, l = 108, a = 97, o = 111, m = 109, i = 105, d = 100.

$$ED(h1,h2) = \sqrt{|101-99|^2 + |115-111|^2 + |99-109|^2 + |117-105|^2 + |101-100|^2 + |108-97|^2 + |97-115|^2}$$

$$ED(h1,h2) = \sqrt{4+16+100+144+1+121+324} = \sqrt{710} = 26.65$$

- $ED(h1,h2) = 26.65$

Distancia Ultramétrica

- Al igual que la de n-gramas define distancias no enteras, y es para prefijos.
- Sean *u* y *v* dos palabras: $u = u_1 ... u_{|u|}$ y $v = v_1 ... v_{|v|}$ defínase como la frontera inferior del conjunto $\{k \mid u_k \neq v_k\}$.
- La distancia se define como:

$$T^* x T^* \rightarrow \Re$$

$$(u,v) \rightarrow \frac{1}{k_{(u,v)}} \quad \text{si} \quad u \neq v$$

$$(u,u) \rightarrow 0$$

- **Entre más cerca de cero más parecidas son las hileras**.

Distancia Ultramétrica – Ejemplo #1

h1 = **ESCUELA** ***h2*** = **ESCUELAS**

E	S	C	U	E	L	A	
E	S	C	U	E	L	A	S

Distancia Ultramétrica – Ejemplo #1

E		S	C	U	E	L	A	
E		S	C	U	E	L	A	S

Distancia Ultramétrica – Ejemplo #1

E	S		C	U	E	L	A	
E	S		C	U	E	L	A	S

Distancia Ultramétrica – Ejemplo #1

E	S	C		U	E	L	A	
E	S	C		U	E	L	A	S

Distancia Ultramétrica – Ejemplo #1

E	S	C	U		E	L	A	
E	S	C	U		E	L	A	S

Distancia Ultramétrica – Ejemplo #1

E	S	C	U	E		L	A	
E	S	C	U	E		L	A	S

Distancia Ultramétrica – Ejemplo #1

E	S	C	U	E	L		A	
E	S	C	U	E	L		A	S

Distancia Ultramétrica – Ejemplo #1

E	S	C	U	E	L	A		
E	S	C	U	E	L	A	S	

Posición de la primera X = *UD(h1,h2)* = 1/8

Distancia Ultramétrica – Ejemplo #2

h1 **= ESCUELA** ***h2*** **= COMIDAS**

E	S	C	U	E	L	A
C	O	M	I	D	A	S

Posición de la primera X = *UD(h1,h2)* = 1/1 = 1

Distancia Ultramétrica – Resumen Ejemplos

- Ejemplo 1:
  - *h1* = escuela
  - *h2* = escuelas

    En el octavo carácter es diferente
  - *UD*(*h1*,*h2*) = 1/8 = 0.125
- Ejemplo 2:
  - *h1* = escuela
  - *h2* = comidas

    En el primer carácter es diferente
  - *UD*(*h1*,*h2*) = 1/1 = 1

Distancias Generadas por Secuencias Positivas

- Introducen un costo sobre las diferencias entre dos hileras.
- Da peso a cada posición.
- Sean $\delta_{x,y}$, que toma valores de 1 si $x = y$ y 0 en el otro caso y $(a_i)_{i \in N}$, una secuencia de reales estrictamente positivos. Se define como $d[a_i]$:

$$T^* x T^* \rightarrow R$$

$$(u, v) \rightarrow \sum_i a_i \delta_{u_i v_i}$$

- **Entre más cerca de cero más diferentes son las hileras**.

Distancias Generadas por Secuencias Positivas – Ejemplo #1

h1 = **ESCUELA** ***h2*** = **ESCUELAS**

E	S	C	U	E	L	A	
E	S	C	U	E	L	A	S

Distancias Generadas por Secuencias Positivas – Ejemplo #1

E	S	C	U	E	L	A	
E	S	C	U	E	L	A	S
1							

Distancias Generadas por Secuencias Positivas – Ejemplo #1

E	S	C	U	E	L	A	
E	S	C	U	E	L	A	S
1	1						

Distancias Generadas por Secuencias Positivas – Ejemplo #1

E	S	C	U	E	L	A	
E	S	C	U	E	L	A	S
1	1	1					

Distancias Generadas por Secuencias Positivas – Ejemplo #1

E	S	C	U	E	L	A	
E	S	C	U	E	L	A	S
1	1	1	1				

Distancias Generadas por Secuencias Positivas – Ejemplo #1

E	S	C	U	E	L	A	
E	S	C	U	E	L	A	S
1	1	1	1	1			

Distancias Generadas por Secuencias Positivas – Ejemplo #1

E	S	C	U	E	L	A	
E	S	C	U	E	L	A	S
1	1	1	1	1	1		

Distancias Generadas por Secuencias Positivas – Ejemplo #1

E	S	C	U	E	L	A	
E	S	C	U	E	L	A	S
1	1	1	1	1	1	1	

Distancias Generadas por Secuencias Positivas – Ejemplo #1

Secuencia a_i = [1,1,1,1,1,1,1,1]

E	S	C	U	E	L	A	
E	S	C	U	E	L	A	S
1	1	1	1	1	1	1	0

SPD(h1,h2) = 1*1 + 1*1 + 1*1 + 1*1 + 1*1 + 1*1 + 1*1 + 1*0 = 7

Distancias Generadas por Secuencias Positivas – Ejemplo #1

h1 **= ESCUELA** ***h2*** **= COMIDAS**

E	S	C	U	E	L	A
C	O	M	I	D	A	S

Distancias Generadas por Secuencias Positivas – Ejemplo #1

Secuencia a_i = [1,1,1,1,1,1,1,1]

E	S	C	U	E	L	A
C	O	M	I	D	A	S
0	0	0	0	0	0	0

SPD(h1,h2) = 1*0 + 1*0 + 1*0 + 1*0 + 1*0 + 1*0 + 1*0 + 1*0 = 0

Distancias Generadas por Secuencias Positivas – Resumen Ejemplos

- Ejemplo 1:
  - *h1* = escuela
  - *h2* = escuelas

    En el octavo carácter es diferente, son iguales en 7 caracteres.
  - a_i = [1,1,1,1,1,1,1,1]
  - *SPD*(*h1*, *h2*) = 1*1 + 1*1 + 1*1 + 1*1 + 1*1 + 1*1 + 1*1 + 1*0 = 7

Distancias Generadas por Secuencias Positivas – Resumen Ejemplos

- Ejemplo 2:
  - *h1* = escuela
  - *h2* = comidas

    En el primer carácter es diferente, son iguales en 0 caracteres.
  - a_i = [1,1,1,1,1,1,1,1]
  - *SPD*(*h1*, *h2*) = 1*0 + 1*0 + 1*0 + 1*0 + 1*0 + 1*0 + 1*0 + 1*0 = 0

Otros Usos

- Estrategia de control de redes de interconexión.
- Reconocimiento de patrones.
- Aprendizaje automático de controladores basados en reglas.
- La biología computacional para búsqueda sobre secuencias de ADN y el reconocimiento de cadenas completas de proteínas.

Práctica

- Sean:
  - *s1* = perro, *s2* = perros, *s3* = pescado.
- Calcular las distancias *d(s1,s2)* y *d(s1,s3)* empleando Hamming, Levenstein, N-gramas (con N = 2, o sea, Bigramas), Euclideano, Ultramétrico y Secuencias Positivas.

Práctica – Hamming

- Número de caracteres en la misma posición que no coinciden (se rellena con vacíos los espacios en blanco de la hilera más corta):
  - $d(s1, s2) = 1$.
  - $d(s1, s3) = 5$.

Práctica – Levenstein

- Tres operaciones posibles, destrucción (*des(x)*), inserción (*ins(x)*) y substitución (*subs(x)*):
  - *d(s1, s2)* = 1 (solo es necesario efectuar *des(s)* en *s2*).
  - *d(s1, s3)* = 4 (es necesario realizar operaciones sobre la hilera más corta (*s1*) y aprovechando la o: *subs(r)* por *s*, *subs(r)* por *c*, *ins(a)* y *ins(d)*).

Práctica – Bigrama

- *s1* = pe er rr ro. Bigramas únicos: pe er rr ro.
- *s2* = pe er rr ro os. Bigramas únicos: pe er rr ro os.
- *s1* tiene 4 bigramas, 4 únicos.
- *s2* tiene 5 bigramas, todos únicos.
- Comparten 4 bigramas: pe er rr ro,
- Usando el coeficiente de Dice tenemos:
  - $d(s1, s2) = (2*4)/(4+5) = 8/9 = 0{,}88\ldots$

Práctica – Bigrama (cont.)

- *s1* = pe er rr ro. Bigramas únicos: pe er rr ro.
- *s3* = pe es sc ca ad do. Bigramas únicos: pe es sc ca ad do.
- *s1* tiene 4 bigramas, todos únicos.
- *s3* tiene 6 bigramas, todos únicos.
- Comparten solo 1 bigrama: pe.
- Usando el coeficiente de Dice tenemos:
  - *d(s1, s3)* = (2*1)/(4+6) = 1/5 = 0,2.

Práctica – Euclideano

- La distancia es la raíz de la sumatoria de las diferencia entre los valores ASCII de los campos de las hileras. Esto para no considerar los negativos.
- Sean los valores ASCII:
  - a = 97, e = 101, o = 111, c = 99, d = 100, p = 112, r = 114, s = 115 y espacio en blanco = 32.
- *d(s1, s2)* = 83.

$$\sqrt{|112-112|^2+|101-101|^2+|114-114|^2+|114-114|^2+|111-111|^2+|32-115|^2}$$

- *d(s1, s3)* = 106.24

$$\sqrt{|112-112|^2+|101-101|^2+|114-115|^2+|114-99|^2+|111-97|^2+|32-100|^2+|32-111|^2}$$

Práctica – Ultramétrico

- La primera posición en donde las cadenas son distintas (iniciando en la posición 1):
  - $d(s1, s2) = 1/6 = 0{,}166\ldots$
  - $d(s1, s3) = 1/3 = 0{,}33\ldots$

Práctica – Secuencias positivas

- Se requiere de una secuencia de reales $(a_i)_{i \in N}$ con i = medida de la cadena mayor. Es aquí en donde puede jugarse con los pesos de cada posición.
- *d(s1, s2)* = 1*1 + 1*2 + 1*3 + 1*4 + 1*5 + 0*6 = 15
  para este caso sea $a = [1,2,3,4,5,6]$
- *d(s1, s3)* = 1*1 + 1*2 + 0*3 + 0*4 + 0*5 + 0*6 + 0*7 = 3
  para este caso sea $a = [1,2,3,4,5,6,7]$

Tabla Comparativa de Resultados

	HD	LD	BiD	ED	UD	SPD
d(s1, s2)	1	1	8/9 = 0,8...	83	1/6 = 0,16…	15
d(s1, s3)	5	4	1/5 = 0,2	117,25	1/3 = 0,3…	3

Cada función se interpreta distinto, a veces entre más alto sea el resultado, más alta será la distancia, en otras es todo lo contrario.

Referencias Bibliográficas

- La información fue tomada de:
  - Libro de texto del curso.