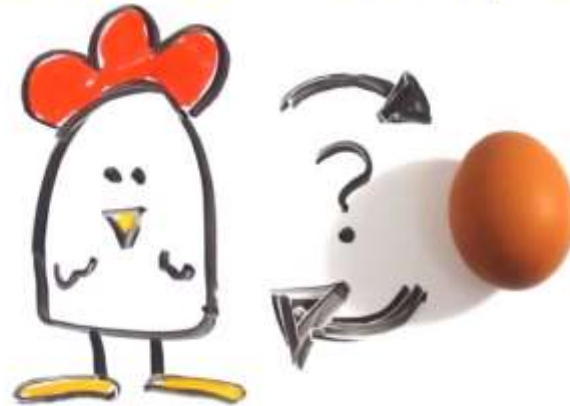


Estructura de Datos y Algoritmos

ITBA 2024-Q2

¿EL HUEVO O LA GALLINA?



¿Estructura de Datos?

O

¿Algoritmos?



Algoritmos para textos

Múltiples motivaciones...

Algoritmos para textos

- Política: este discurso ya lo escuché... Está repitiendo lo que dijo otro político.

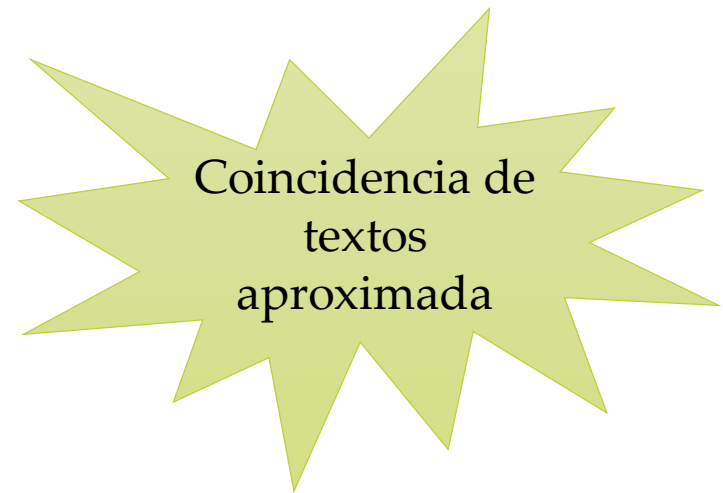


Coincidencia de
textos completa o
parcial
(exact string
matching)

- Alumnos/Autores: esta respuesta coincide con la de este otro alumno => se copió!!
- Plagio en música

Algoritmos para textos

- Política: este discurso se parece a uno que escuché...



- Biología: ADN
- Typos

Algunas definiciones

Alfabeto Σ : conjunto de símbolos o caracteres.

Dado un alfabeto Σ y $k_{\geq 0} \in \mathbb{N}$, un string S es un elemento $\in \Sigma^k$

Para $S \in \Sigma^k$, se dice que $|S|$ es k , y denota su longitud. Si $k=0$, S se dice que es el string vacío, se lo denota con λ

- ¿Por qué se lo denota con λ ?

Para evitar los problemas que tienen los compiladores!!

Un **meta-símbolo** no debería ser al mismo tiempo **parte del alfabeto** Σ (regla básica).

Los lenguajes de programación violan estas reglas y ahí surgen los problemas.

Ejemplo: En Java, el caracter **comilla doble** delimita el comienzo y fin del string. No son parte de las operaciones. El string "EDA" tiene 3 símbolos, no 5. Hasta ahí parece sencillo: quitemos los 2 caracteres externos.

Pero, si el compilador encontrara **“hola”que** daría error, porque no sabe dónde termina el string:

¿Es **“hola”** y lo que sigue está mal?

¿Debería ser **“hola”que** ?

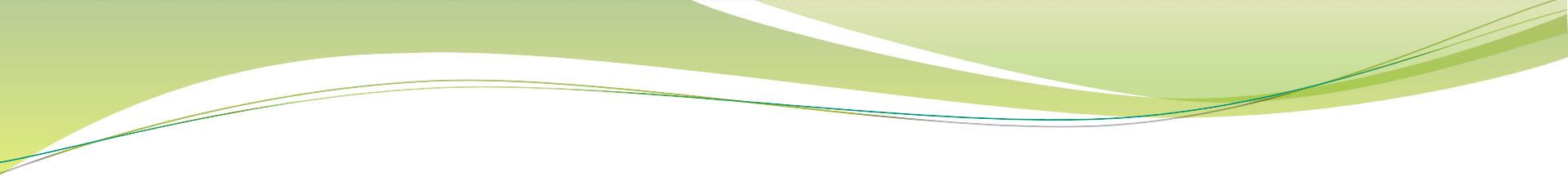
Para evitar ambigüedades obliga a escapar al símbolo comillas dobles cuando participa del string. Se lo escapa con la barra invertida: **\"**

Pero yo lo veo como un doble caracter, pero representa uno solo!!!!

Ejemplo: **“hola\"que** estaría queriendo representar al string **hola”que** de 8 caracteres y no de 9 caracteres.

Pero la ambigüedad no está solucionada. Otra vez, el símbolo barra invertida es **meta-símbolo** y **parte del alfabeto Σ** .

Ejemplo: **“\\hola\"que** estaría queriendo representar al string **\\hola”que** de 9 caracteres no de 11 caracteres.



Si no nos presenta confusión, podemos hablar del ""
como el string vacío...

Algunas definiciones

Conjunto de todos los strings posibles sobre cierto alfabeto

Dado un alfabeto Σ , $\Sigma^* = \bigcup \Sigma^k$ con $k_{\geq 0} \in \mathbb{N}$

Algunas definiciones

Concatenación de Strings

Dado un alfabeto Σ , y $u \in \Sigma^*$, $w \in \Sigma^*$.

Se llama **concatenación** al string definido como uw (un elemento a continuación del otro, sin símbolos extra entre ellos).

Prefijos, Sufijos y Substrings

Dados un alfabeto Σ y los strings $x \in \Sigma^*$, $w \in \Sigma^*$, $z \in \Sigma^*$. Sea $p = xwz$.

Se dice que x es un **prefijo** de p . Se dice que w es un **substring** de p . Se dice que z es un **sufijo** de p .

Bordes

Dados un alfabeto Σ y los strings $x \in \Sigma^*$, $w \in \Sigma^*$, $z \in \Sigma^*$.
Si $p = wx = zw$ donde $|x| = |z|$, se dice que w es un **border** de p .

Ejemplos

$\Sigma = \{0, 1, 2, 3, 4, 5\}$

Sea $s = "01230"$

¿Cuáles son los prefijos de s ?

Rta: $""$, $"0"$, $"01"$, $"012"$, $"0123"$, s

¿Cuáles son los sufijos de s ?

Rta: $""$, $"0"$, $"30"$, $"230"$, $"1230"$, s

¿Cuáles son los borders de s ?

Rta: $""$, s , $"0"$. Como mínimo hay 2 borders: $""$ y s

¿Cuáles son los substrings de s ?

Rta: $""$, $"0"$, $"01"$, $"012"$, $"0123"$, s , $"30"$, $"230"$, $"1230"$, $"1"$, $"12"$, $"123"$, $"2"$, $"23"$, $"3"$, $"30"$, $"0"$

Data Quality - Matching

El tópico de “búsqueda aproximada” fue estudiado ampliamente.

Data Quality - Matching

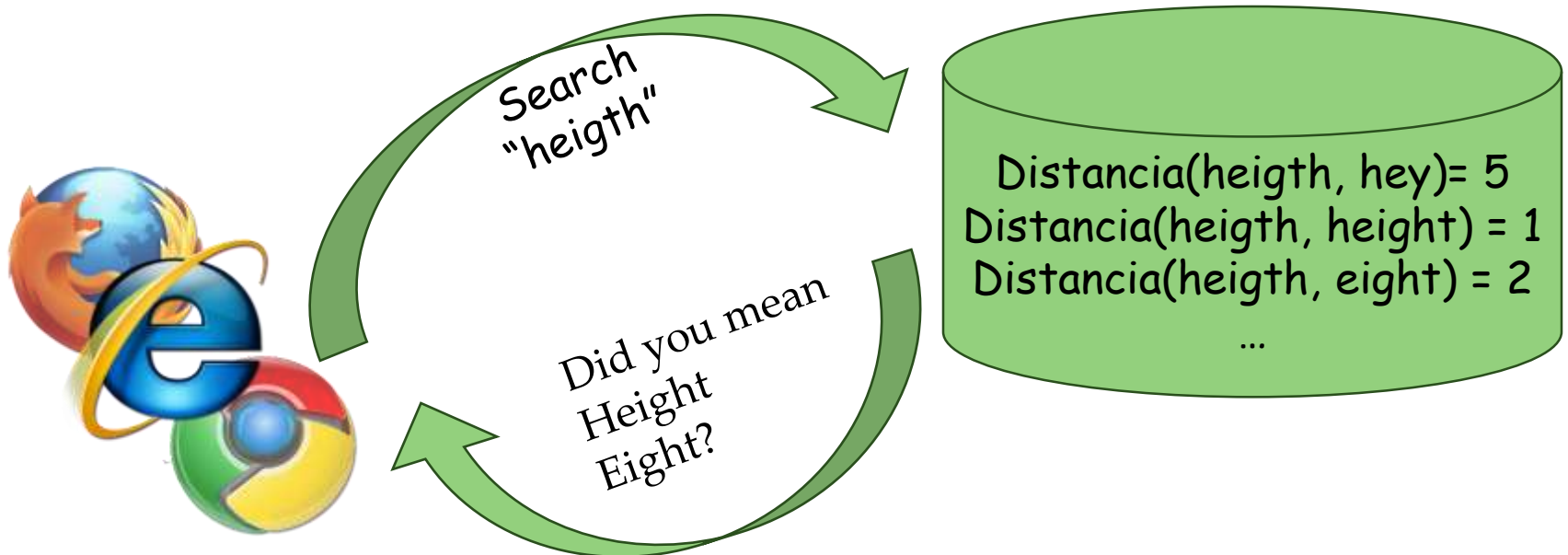


Los buscadores usan alguna estrategia, en el caso de que la búsqueda lanzada no sea reconocida en el “corpus” que poseen sobre búsquedas y documentos indizados.

Data Quality - Matching

Las estrategias pueden ser muy variadas (combinaciones de una o más de estas):

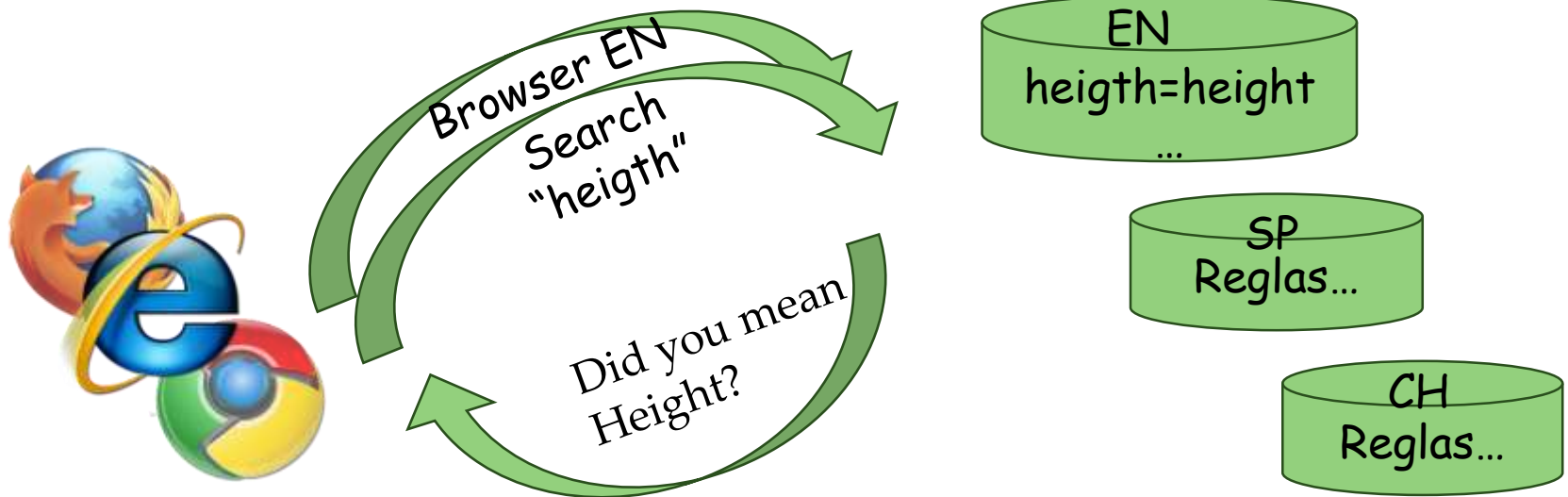
- Buscar las palabras y si las palabras no están en el corpus de los documentos indizados, encontrar las que mayor similitud posean y sugerirlas.



Data Quality - Matching

Las estrategias pueden ser muy variadas (combinaciones de una o más de estas):

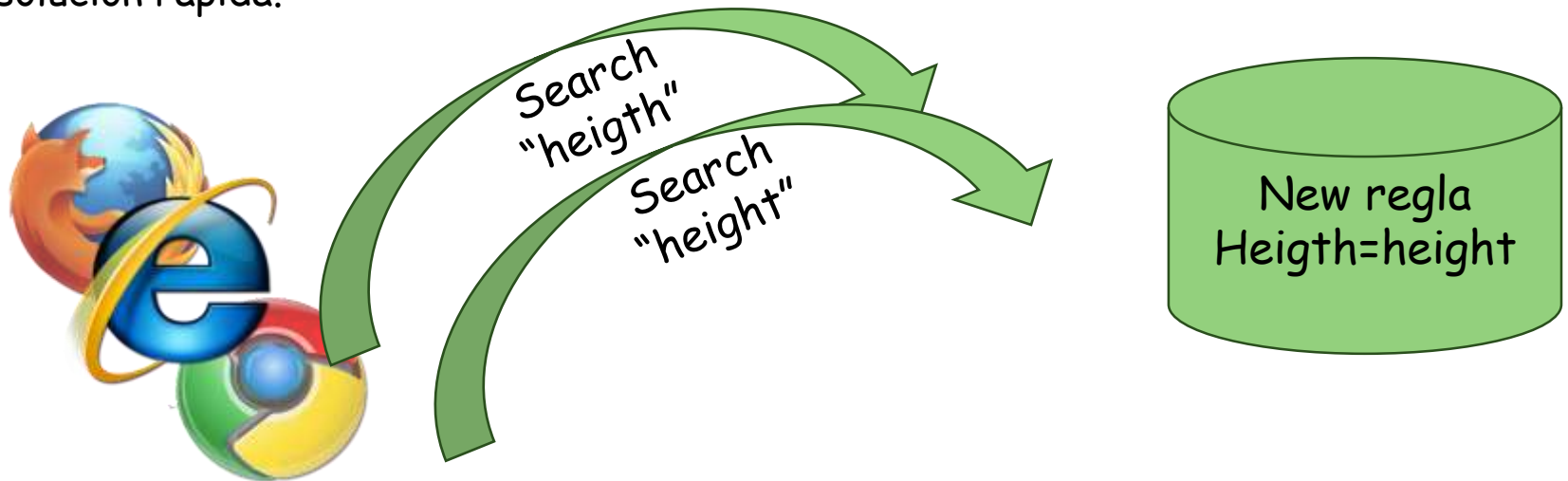
- Tomar el idioma que tiene configurado el browser para saber en qué corpus buscar las palabras usadas. Hay reglas conocidas por el idioma en cuestión. Ejemplo, en inglés HT con TH: height vs width, length.



Data Quality - Matching

Las estrategias pueden ser muy variadas (combinaciones de una o más de estas):

- Sabiendo que los usuarios buscan palabras y cuando fue un error de typo/ortografía/etc., no cliquean nada del resultado e intentan realizar inmediatamente la búsqueda arreglada, almacenan esas búsquedas erróneas con la que arrojó resultados navegados. Es decir, hay un **MATCHING** entre errores viejos y soluciones que los mismos usuarios hicieron. Si esos errores son frecuentes, tendrán solución rápida.



Data Quality - Matching

No sólo Google/Yahoo/Bing intentan mejorar y arreglar las búsquedas erróneas.

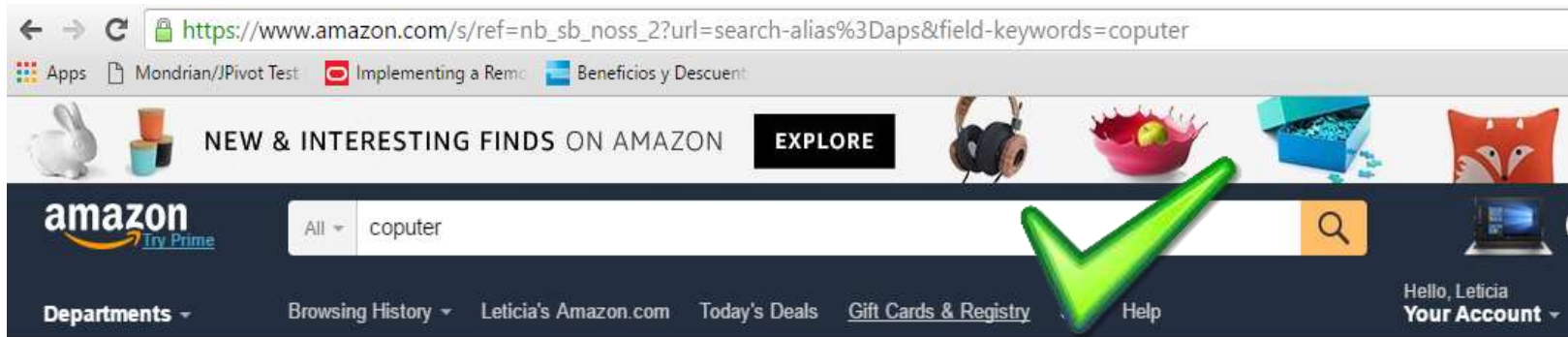
Los sitios que permiten realizar compras por Internet, también.

Si un producto no se encuentra (en la categoría esperada), entonces el usuario intenta usar el botón de búsqueda. Pero si esta no da coincidencia => el usuario abandona el sitio y va a otro (en no más de 2 intentos).

Sería fácil encontrar en este sitio? <http://qualtex.blogspot.com/>

Data Quality - Matching

Mi notebook está teniendo un problema con la letra M. No me toma la tecla, salvo que la presione fuertemente.



Show results for

Computers & Accessories >

- Laptop Computers
- Desktop Computers
- Tower Computers
- Computer Cases
- Traditional Laptop Computers
- + See more

+ See All 35 Departments

Showing results for **computer**. Search instead for "coputer".

Showing most relevant results. See all results for **computer**.



Dell OptiPlex Desktop Complete Computer Package with Windows 7 Home 32-Bit - Keyboard, Mouse, 19" LCD Monitor (brands vary)

by Dell

\$89.00 used (15 offers)

★★★★☆ 284

Electronics: See all 9,818,883 items

Data Quality - Matching

Fíjense la diferencia con este sitio. En este caso el problema fue con la tecla L:



Data Quality - Matching

O en un error de ortografía:

The screenshot shows the Disco website interface. At the top, the URL is https://www.disco.com.ar/Comprar/Home.aspx?#_atCategory=false&_atGrilla=true&_query=gayetita. Below the URL bar, there are tabs for 'Mondrian/JPIVOT Test', 'Implementing a Remo', and 'Beneficios y Descuent'. The main header features the Disco logo, a phone number '0810-777-8888', and a search bar containing 'gayetita'. Below the search bar, there is a navigation bar with icons for various categories: Ofertas, Almacén, Bebidas, Frutas y Verduras, Carnes y Pescados, Quesos y Fiambres, Lácteos, Congelados, Perfumería, and Limpieza. A large red 'X' is overlaid on the 'Congelados' category icon. Below the navigation bar, a message reads 'Resultados de búsqueda para: "gayetita"'. At the bottom, a message states 'No se encontraron resultados.'.

Data Quality - Matching

Mínimas reglas que deberían aplicarse:

- ❖ Sacar blancos del comienzo y final (trim).

Pero no es suficiente. Si la palabra es compuesta habría que sacar blancos internos.

Ej: ' yogurt bebible '

- ❖ Buscar pasando todo a mayúscula o minúscula. Ej: YogUrt= YOGURT

- ❖ Si se conocen abreviaturas, usarlas. Ej: BA por Buenos Aires

- ❖ Los símbolos de puntuación, eliminarlos. Ej: Bs. As. por Bs As

- ❖ Si se conocen sinónimos, usarlos. Ej: *computadora* por *ordenador*, *teléfono celular* por *teléfono móvil*. Inclusive entre diferentes idiomas.

Data Quality - Matching

Reglas específicas que deberían aplicarse para los tipos de datos:

- ❖ Fechas y sus formatos

Ej: 12/10/2016 = 12 Oct 2016 = 2016-10-12

- ❖ La hora y sus formatos

Ej: 15:30 = 3:30 PM

- ❖ Números

Ej: 12.300.140 = 12300140

- ❖ Números Decimales

Ej: 12,1 = 12.1

- ❖ String correspondientes a nombre y apellidos

Ej: John Peter Doe = John P. Doe = J. D. Doe = Doe, John Peter = Doe, John P. = Doe, J. P.



Algoritmos

String Matching

SOUNDEX

Es un algoritmo fonético, es decir codifica a una palabra según "suena". Intenta solucionar problemas de pronunciación.

Fue creado para el alfabeto inglés (o sea, codifica las 26 letras del mismo). Existen otras adaptaciones como Soundex_FR para idioma francés.

Originalmente fue propuesto a comienzos del siglo XX por Rusell y Odell. Se lo utilizó para el censo de USA en los 30'.

Aquellas palabras que "suenan igual", aunque no se escriban igual, deben ser codificadas de la misma manera.

Existen varias versiones y adaptaciones. Estudiaremos la siguiente, que codifica un string IN en otro OUT.

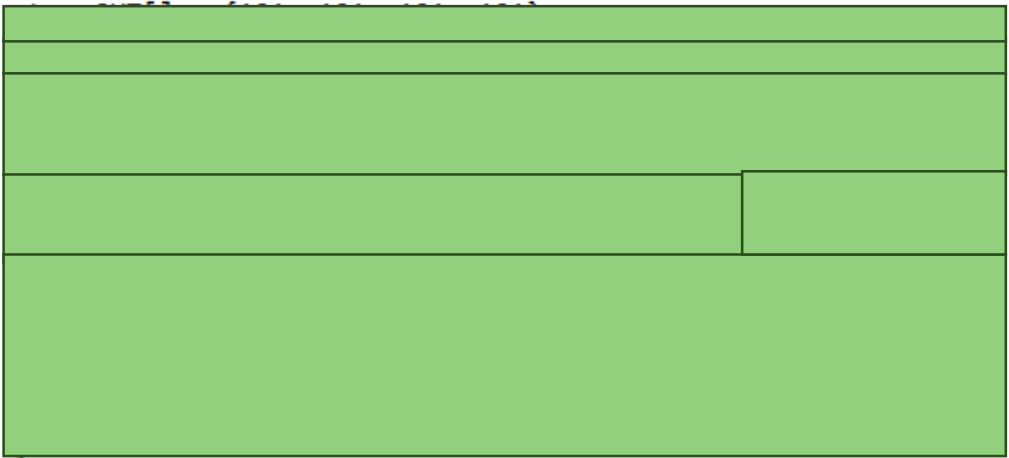
26 Letras	Pesos fonéticos
A, E, I, O, U, Y, W, H	0 -- no se codifica
B, F, P, V	1
C, G, J, K, Q, S, X, Z	2
D, T	3
L	4
M, N	5
R	6

Soundex siempre devuelve un código OUT de exactamente 4 caracteres, formados por: primero una letra y luego 3 dígitos (pesos fonéticos).

Si hace falta, para completar el código de 4 caracteres, se completan con 0s (ceros) al final.

26 Letras	Pesos fonéticos
A, E, I, O, U, Y, W, H	0 -- no se codifica
B, F, P, V	1
C, G, J, K, Q, S, X, Z	2
D, T	3
L	4
M, N	5
R	6

```
char IN[] = new String("...").toCharArray();
```



```
return new String(OUT);
```

- Paso 1 (opcional): Pasar a mayúsculas y dejar sólo las letras (dígitos, símbolos de puntuación, espacios, etc. se eliminan).
- Paso 2: Colocar OUT[0]=IN[0].
- Paso 3: Se calcula vble. **last** como el peso fonético de IN[0]
- Paso 4: Para cada letra **iter** siguiente en IN y hasta completar 3 dígitos o terminar de procesar IN, hacer
 - 3.1) calcular vble **current** con peso fonético de **iter**. Si es diferente a 0 y no coincide con **last**, appendear **current** en **OUT**.
 - 3.2) **independiente del paso anterior**, tapar **last** = **current**.
- Paso 5: si hace falta completar con '0's y devolver OUT.

Actividad



Apliquemos **Soundex** para codificar strings por su fonética.

26 Letras	Pesos fonéticos
A, E, I, O, U, Y, W, H	0 -- no se codifica
B, F, P, V	1
C, G, J, K, Q, S, X, Z	2
D, T	3
L	4
M, N	5
R	6

Soundex(' LUxuRY')=

Soundex(' LUxuRY YYAAAE')=

Actividad



Apliquemos **Soundex** para codificar strings por su fonética.

26 Letras	Pesos fonéticos
A, E, I, O, U, Y, W, H	0 -- no se codifica
B, F, P, V	1
C, G, J, K, Q, S, X, Z	2
D, T	3
L	4
M, N	5
R	6

Soundex(' SZLLOYDTIRUL')=

Cómo usar soundex?

Soundex codifica, pero ¿Cómo usarlo para comparar proximidad entre palabras?

Soundex NO es una métrica.

Hay que definir cómo obtener una métrica a partir de soundex.

Cómo usar soundex?

Definición de Similitud para Soundex

Es la proporción de caracteres coincidentes entre los encodings respecto a la longitud del encoding.

Cómo usar soundex?

Ej: Soundex ("threshold") = T624
Soundex("hold") = H430
Soundex("zresjoulding") = Z624
Soundex("phone") = P500
Soundex ("foun") = F500

Soundex ("threshold", "hold") = 0
Soundex("threshold", "zresjoulding") = $\frac{3}{4} = 0.75$
Soundex("phone", "foun") = $\frac{3}{4} = 0.75$

¿ Cuáles son los únicos valores posibles de similitud que obtenemos?
Rta 0, 0.25, 0.5, 0.75, 1