# Lenguajes de programación - T08: Implementación de la distancia Damerau-Levenshtein en el lenguaje R

Jorge Aurelio Morales Manrique
C.C. 1010075711
jomorales@unal.edu.co

Universidad Nacional de Colombia
Abril 13 de 2021

## 1. Marco Teórico

- **Lenguaje de programación.** Un lenguaje de programación es un lenguaje formal (o artificial, es decir, un lenguaje con reglas gramaticales bien definidas) que le proporciona a una persona, en este caso el programador, la capacidad de escribir (o programar) una serie de instrucciones o secuencias de órdenes en forma de algoritmos con el fin de controlar el comportamiento físico o lógico de un sistema informático, de manera que se puedan obtener diversas clases de datos o ejecutar determinadas tareas. A todo este conjunto de órdenes escritas mediante un lenguaje de programación se le denomina programa informático.

- **Palabras reservadas en un lenguaje de programación.** En los lenguajes informáticos, una palabra reservada es una palabra que tiene un significado gramatical especial para ese lenguaje y no puede ser utilizada como un identificador de objetos en códigos del mismo, como pueden ser las variables. Por ejemplo, en SQL, un usuario no puede ser llamado "group", porque la palabra group es usada para indicar que un identificador se refiere a un grupo, no a un usuario. Al tratarse de una palabra clave su uso queda restringido. Ocasionalmente la especificación de un lenguaje de programación puede tener palabras reservadas que están previstas para un posible uso en futuras versiones. En Java const y goto son palabras reservadas — no tienen significado en Java, pero tampoco pueden ser usadas como identificadores. Al reservar los términos pueden ser implementados en futuras versiones de Java, si se desea, sin que el código fuente más antiguo escrito en Java deje de funcionar.

- **Distancia Damerau-Levenshtein.** En la teoría de la información y en la ciencia de computadores, se llama distancia de Damerau-Levenshtein o distancia de edición al número mínimo de operaciones requeridas para transformar una cadena de caracteres en otra. Se entiende por operación, bien una inserción, eliminación, sustitución o transposición de dos caracteres. Lo que la distingue de la distancia de Levenshtein es que esta última cuenta como una sola operación de edición a cualquiera de las tres primeras, pero cuenta la transposición como dos operaciones de edición. A continuación se muestra en pseudocódigo de un algoritmo que implementa lo antes mencionado.

```
algorithm OSA-distance is
    input: strings a[1..length(a)], b[1..length(b)]
    output: distance, integer

    let d[0..length(a), 0..length(b)] be a 2-d array of integers, dimensions length(a)+1, length(b)+1
    // note that d is zero-indexed, while a and b are one-indexed.

    for i := 0 to length(a) inclusive do
        d[i, 0] := i
    for j := 0 to length(b) inclusive do
        d[0, j] := j

    for i := 1 to length(a) inclusive do
        for j := 1 to length(b) inclusive do
            if a[i] = b[j] then
                cost := 0
            else
                cost := 1
            d[i, j] := minimum(d[i-1, j] + 1,     // deletion
                               d[i, j-1] + 1,     // insertion
                               d[i-1, j-1] + cost)  // substitution
            if i > 1 and j > 1 and a[i] = b[j-1] and a[i-1] = b[j] then
                d[i, j] := minimum(d[i, j],
                                   d[i-2, j-2] + 1)  // transposition
    return d[length(a), length(b)]
```

Figura 1: Algoritmo Distancia DL.

# Referencias

[1] Aaby, Anthony. *Introduction to Programming Languages*.

[2] ISO/IEC. *International Organization for Standardization*.

[3] Damerau, Fred J. *A technique for computer detection and correction of spelling errors*. Communications of the ACM.