inexact: un addin de RStudio para supervisar la unión *fuzzy* de bases de datos

https://github.com/arcruz0/inexact

"inexact" es un paquete y addin de RStudio que permite supervisar el proceso de *fuzzy join*, la unión automatizada de bases de datos con discrepancias en los valores de sus columnas comunes. La presente postulación se divide en dos partes: primero se describe el problema a resolver y luego se presenta "inexact".

Veamos los siguientes datos de ejemplo, en los que las observaciones corresponden a países (codificados sus nombres según estándares ligeramente distintos):

```
(datos_c <- data.frame(
  pais  = c("Argentina", "Chile", "Uruguay", "Bolivia", "Brasil"),
  var_a = 1:5, stringsAsFactors = F
))
       pais var_a
1 Argentina     1
2     Chile     2
3   Uruguay     3
4   Bolivia     4
5    Brasil     5
```

```
(datos_d <- data.frame(
  pais  = c("Argentina", "Chile", "Uruguay", "Bolivia (Plurinational State of)", "Brazil"),
  var_b = 11:15, stringsAsFactors = F
))
                              pais var_b
1                        Argentina    11
2                            Chile    12
3                          Uruguay    13
4 Bolivia (Plurinational State of)    14
5                           Brazil    15
```

Una de las operaciones con datos más comunes es la unión izquierda: en este caso, querríamos añadir la columna "var_b" en "datos_c", para tener un solo *data frame* con la información de ambas variables. Sin embargo, aquí la columna en común, "pais", es inexacta entre las bases para las últimas dos observaciones, por lo que la unión no es completa:

```
dplyr::left_join(datos_c, datos_d, by = "pais")
       pais var_a var_b
1 Argentina     1    11
2     Chile     2    12
3   Uruguay     3    13
4   Bolivia     4    NA
5    Brasil     5    NA
```

Es posible solucionar este problema gracias a los algoritmos de pareo aproximado de texto (*approximate string matching)*, que en R están implementados en el paquete "stringdist" (van der Loo et al., 2018). Dichos algoritmos calculan distancias entre textos de acuerdo a criterios como las letras en común, la longitud, etc. Volviendo al ejemplo, podemos cruzar los valores de la columna “pais” para ambas bases y construir una matriz con las distancias correspondientes (de acuerdo al algoritmo “osa” de alineamiento optimo de textos):

```
matriz <- stringdist::stringdistmatrix(datos_c$pais, datos_d$pais, method = "osa")
rownames(matriz) <- datos_c$pais; colnames(matriz) <- datos_d$pais
matriz
          Argentina Chile Uruguay Bolivia (Plurinational State of) Brazil
Argentina         0     8       8                                27      7
Chile             8     0       7                                29      5
Uruguay           8     7       0                                30      6
Bolivia           7     6       7                                25      5
Brasil            7     5       6                                27      1
```

A partir de esta matriz se puede realizar un ejercicio corregido de unión de bases, implementado en el paquete “fuzzyjoin” de R (Robinson, 2018). Se comienza buscando, en cada fila de la matriz, el valor mínimo de distancia. Para los países del Cono Sur el pareo es exacto, por lo que sus valores en la diagonal corresponden a 0. Para "Brasil", cuya diferencia con "Brazil" es de sola una letra, la distancia asignada es de 1, lo que denota un pareo razonable. De esta forma, una unión ingenua realizada a través de “fuzzyjoin” —`fuzzyjoin::stringdist_left_join(datos_c, datos_d, method = “osa”, max_dist = Inf)`— funcionará a la hora de añadir en la nueva base el valor de “var_b” para Brasil. El argumento `max_dist = Inf` hará a la función completar la unión para cada fila, en cada caso eligiendo el valor mínimo de distancia en la matriz antes descrita.

Sin embargo, esta aproximación no es satisfactoria para el caso de Bolivia. El valor mínimo de distancia en su fila (5) está asignado para "Brazil", cuando sabemos que el pareo correcto es "Bolivia (Plurinational State of)", el valor máximo de distancia (25). En casos como este es que la supervisión humana puede complementar adecuadamente al enfoque automatizado.

Es posible hacer dicha supervisión utilizando los paquetes tratados hasta aquí, generando *data frames* intermedios que contengan más pareos que los de distancia mínima, para luego hacer filtros según el criterio del usuario(a) a la hora de inspeccionar los pareos visualmente. Sin embargo, este proceso es tedioso e ineficiente en términos de tiempo, en especial para bases de datos que no son pequeñas. "inexact" busca simplificarlo a través de un addin de RStudio, que provee una interfaz gráfica (GUI) especialmente pensada para supervisar la pertinencia de pareos aproximados.

La Figura 1 muestra su ventana inicial, con las opciones específicas para la unión del ejemplo. La Figura 2.1 muestra el corazón de "inexact": el usuario(a) puede supervisar todos los pareos imperfectos, que se muestran desde el más conflictivo al menos conflictivo (decrecientemente según la distancia mínima detectada). Luego es posible editar casos específicos (Figura 2.2), siendo la opción por defecto siempre el pareo con menor distancia. La utilización de una GUI no afecta la reproducibilidad del procedimiento, pues su resultado final es código, como muestra la Figura 3: esta es una buena práctica apreciable en otros paquetes que implementan addins de RStudio, como “questionr” (Barnier, Briatte & Larmarange, 2018).

El código que permite a “inexact” funcionar tras bambalinas proviene de “fuzzyjoin” (Robinson, 2018), por lo que solo actúa como una máscara en lo que refiere al trabajo computacional de la unión. Tras aplicar el código generado en el paso 3 se consigue la unión requerida, tras la aplicación del proceso automatizado y la supervisión humana:

```
       pais var_a var_b
1 Argentina     1    11
2     Chile     2    12
3   Uruguay     3    13
4   Bolivia     4    14
5    Brasil     5    15
```

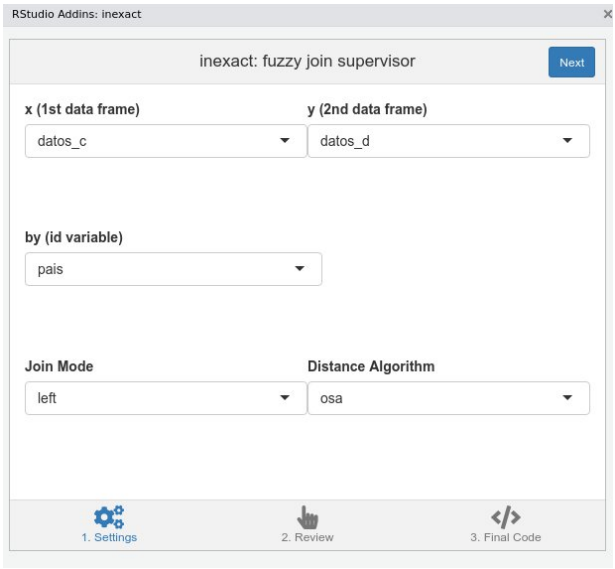

Figura 1. Panel inicial de “inexact”

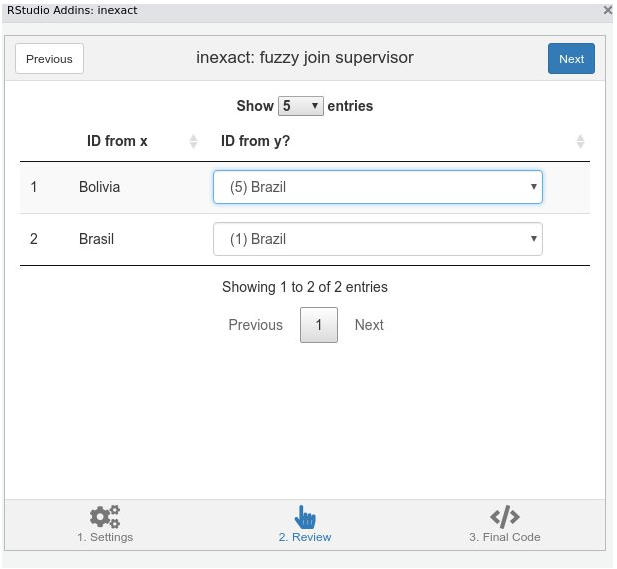

Figura 2.1. Panel de supervisión de “inexact”

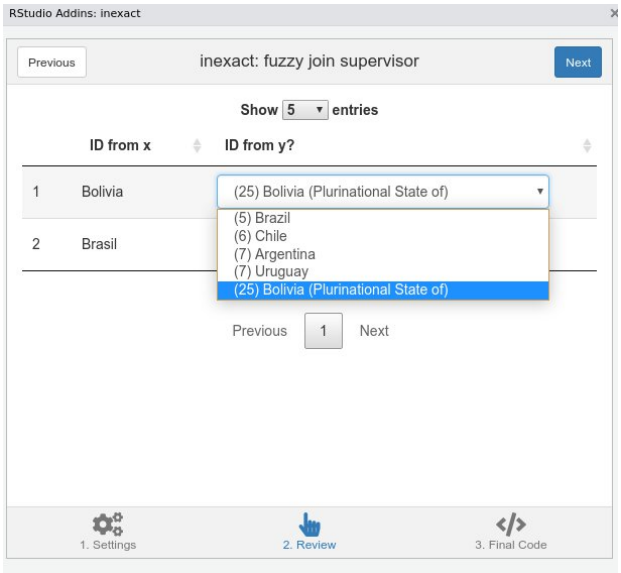

Figura 2.2. Panel de supervisión de “Inexact”

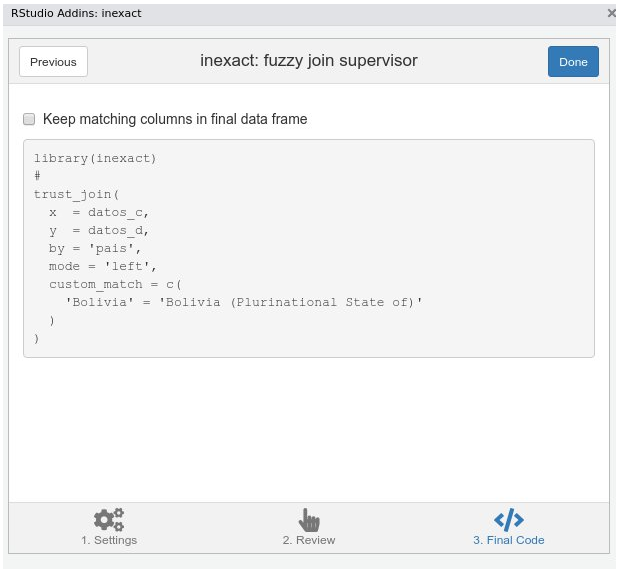

Figura 3. Panel de código de “inexact”