Manejo de Vectores de texto en R

Sesión 3

Natalie Julian - www.nataliejulian.com

Estadística UC y Data Scientist en Zippedi Inc.

Vectores de caracteres

Podemos crear también vectores de texto. El texto **siempre** debe escribirse entre comillas:

```
c("Natalie", "Sergio", "Vanesa")
c("Papas fritas", "Mayonesa", "Doritos", "Cheetos")
c("Cuaderno", "Lapiz", "Regla", "Post it", "Pegamento")
```

Vista

```
39
40  #Vectores de caracteres
41
42  c("Natalie", "Sergio", "Vanesa")
43  c("Papas fritas", "Mayonesa", "Doritos", "Cheetos")
44  c("Cuaderno", "Lapiz", "Regla", "Post it", "Pegamento")
45
```

45:1 (Top Level)

Console Terminal Jobs

~/

```
> c("Natalie", "Sergio", "Vanesa")
[1] "Natalie" "Sergio"  "Vanesa"
> c("Papas fritas", "Mayonesa", "Doritos", "Cheetos")
[1] "Papas fritas" "Mayonesa"     "Doritos"      "Cheetos"
> c("Cuaderno", "Lapiz", "Regla", "Post it", "Pegamento")
[1] "Cuaderno"  "Lapiz"     "Regla"     "Post it"   "Pegamento"
```

Funciones aplicables a vectores de texto

Función	Descripción
`length()`	Largo del vector
`nchar()`	Cantidad de caracteres de los elementos del vector
`substr(vector, start=, stop=)`	Extrae determinados caracteres de cada elemento del vector
`paste(vector1, vector2, sep=" ")`	Concatena dos vectores
`sort()`	Ordena alfabéticamente el vector

Ejemplo

Se posee información sobre el índice de masa corporal de algunxs pacientes en un centro médico. La columna información contiene el RUT de(l/la) paciente sin dígito verificador y rango etario, la columna índice contiene el IMC.

Información	Índice
12466824Anciano	19,5
19566573Joven	25
18622134Adulto	27
17823471Adulto	25
20172423Infante	23
19784132Joven	23
17234124Adulto	35

a) Escriba la información en dos vectores: `Info` e `IMC`. ¿Cuántos pacientes se estudiaron?
b) Extraiga el rango etario de los pacientes en un vector llamado `etario`.
c) Obtenga mínimo, máximo, media y mediana de los índices de masa corporal. Comente.
d) ¿Cuál es el paciente que posee el mayor índice de masa corporal? ¿en qué rango etario está dicho paciente?

a) ¿Cuántos pacientes se estudiaron?

El vector `Info` posee concatenada información numérica y texto, por lo que es necesario escribirla con comillas:

```
Info<-c("12466824Anciano","19566573Joven","18622134Adulto",
"17823471Adulto", "20172423Infante","19784132Joven","17234124Adulto")
```

El vector `IMC` contiene sólo información numérica, basta con escribir los números:

```
IMC<-c(19.5,25,27,25,23,23,35)
```

Para saber cuántos pacientes se estudiaron, basta con obtener el largo de cualquiera de los dos vectores:

```
length(Info)
[1] 7
```

Es decir, en la tabla de datos se tiene información de 7 pacientes.

b) Extraiga el rango etario de los pacientes en un vector llamado `etario`.

Necesitamos extraer el rango etario. Podemos utilizar la función `substr`, pero necesitamos indicarle el `start` (número del caracter a partir del que extrae) y `end` (número del caracter hasta que debe extraer).

Sabemos que los primeros caracteres corresponden a los 8 dígitos del RUT, por lo tanto, necesitamos extraer desde el caracter 9, luego `start=9` y además, sabemos que la función `nchar()` entrega la cantidad total de caracteres, por lo tanto, basta con utilizar:

```
etario<-substr(Info, start=9, stop=nchar(Info))
```

c) Obtenga mínimo, máximo, media y mediana de los índices de masa corporal. Comente.

Para obtener mínimo, máximo, media y mediana, utilizamos las funciones:

```
min(IMC) #Minimo
[1] 19.5
max(IMC) #Maximo
[1] 35
mean(IMC) #Media
[1] 25.35714
median(IMC) #Mediana
[1] 25
```

d) ¿Cuál es el paciente que posee el mayor índice de masa corporal? ¿en qué rango etario está dicho paciente?

```
which.max(IMC) ; etario[which.max(IMC)]
[1] 7
[1] "Adulto"
```

El último paciente de los registrados posee el mayor índice de masa corporal registrado, y pertenece al grupo etario de Adulto.

Conexión con conjuntos

Ejemplo

Dos personas fueron al supermercado y a través de la boleta se posee el conjunto de productos que compraron:

$$A = \{Palta, Leche, Pan, Queso, Arroz, Harina, Chocolates, Shampoo, Fideos, Salsa\ de\ tomate, Servilletas\}$$

$$B = \{Toalla, Shampoo, Pan, Cepillo\ dental, Queso, Jugo, Leche, Servilletas\}$$

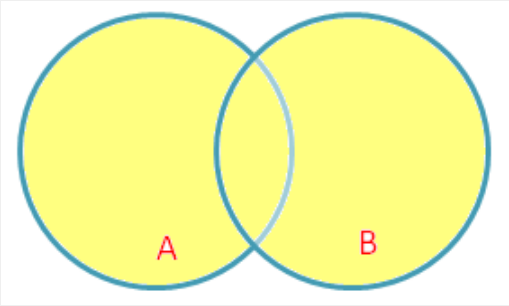
Los elementos de un vector forman un conjunto

Podemos definir estos conjuntos como vectores en R:

```
A<-c("Palta", "Leche", "Pan", "Queso", "Arroz", "Harina", "Chocolates", "Shampoo", "Fideos", "Salsa de tomate", "Servilletas")
B<-c("Toalla", "Shampoo", "Pan", "Cepillo dental", "Queso", "Jugo", "Leche", "Servilletas")
```

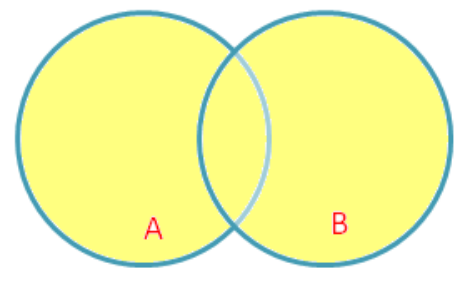
La unión de dos conjuntos: Todos los elementos

¿Cómo obtenemos todos los productos que se compraron? ¿Es decir, cómo obtenemos $A \cup B$?



La unión de conjuntos: Todos los elementos

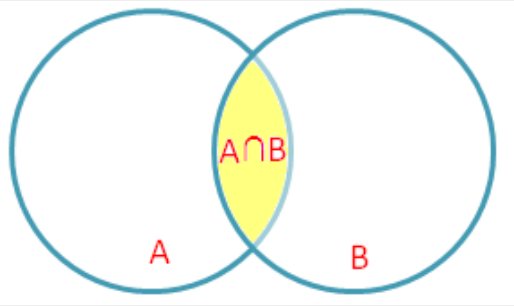
¿Cómo obtenemos todos los productos que se compraron? ¿Es decir, cómo obtenemos $A \cup B$?




```
union(A, B)
 [1] "Palta"           "Leche"           "Pan"             "Queso"
 [5] "Arroz"           "Harina"          "Chocolates"      "Shampoo"
 [9] "Fideos"          "Salsa de tomate" "Servilletas"     "Toalla"
[13] "Cepillo dental"  "Jugo"
```

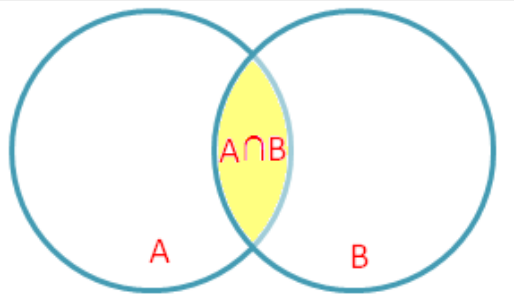
La intersección de dos conjuntos: Los elementos en común

¿Cómo obtenemos los productos que se compraron **en ambas boletas**? ¿Es decir, cómo obtenemos $A \cap B$?



La intersección de conjuntos: Los elementos en común

¿Cómo obtenemos los productos que se compraron **en ambas boletas**? ¿Es decir, cómo obtenemos $A \cap B$?

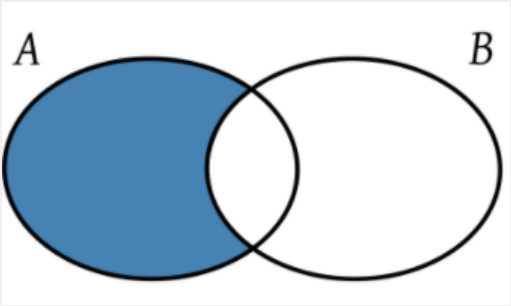



```
intersect(A, B)
[1] "Leche"       "Pan"         "Queso"       "Shampoo"     "Servilletas"
```

Si la intersección es vacía, dos conjuntos se dicen disjuntos.

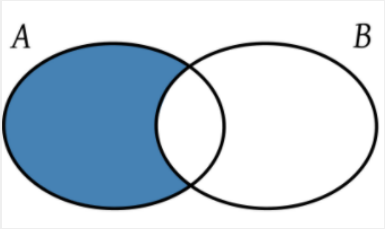
Diferencia de conjuntos: Quitar los elementos de un conjunto a otro

¿Cómo obtenemos los productos que se compraron en la primera boleta pero no en la segunda? ¿Es decir, cómo obtenemos $A - B$?



Diferencia de conjuntos: Quitar los elementos de un conjunto a otro

¿Cómo obtenemos los productos que se compraron en la primera boleta y no en la segunda? ¿Es decir, cómo obtenemos $A - B$?

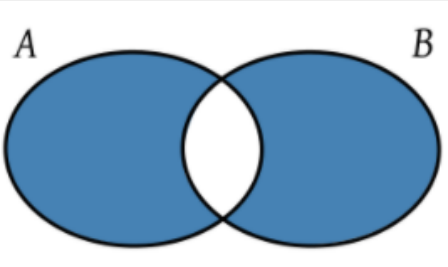



```
setdiff(A, B) #A-B
[1] "Palta"           "Arroz"           "Harina"          "Chocolates"
[5] "Fideos"          "Salsa de tomate"

setdiff(B, A)  #B-A
[1] "Toalla"         "Cepillo dental" "Jugo"
```

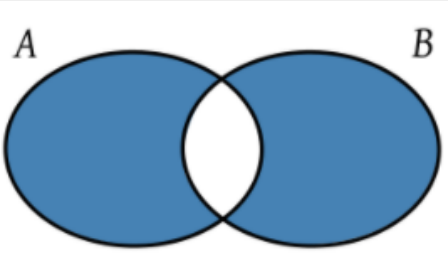
Diferencia simétrica

¿Cómo obtener la diferencia simétrica? Es decir, los productos que sólo se compraron **en una y sólo una boleta**?



Diferencia simétrica

¿Cómo obtener la diferencia simétrica? Es decir, los productos que sólo se compraron **en una y sólo una boleta**?




```
#Forma 1
setdiff(union(A, B), intersect(A, B))
[1] "Palta"           "Arroz"           "Harina"          "Chocolates"
[5] "Fideos"          "Salsa de tomate" "Toalla"          "Cepillo dental"
[9] "Jugo"

#Forma 2
union(setdiff(A, B), setdiff(B, A))
[1] "Palta"           "Arroz"           "Harina"          "Chocolates"
[5] "Fideos"          "Salsa de tomate" "Toalla"          "Cepillo dental"
[9] "Jugo"
```

Funciones útiles para conjuntos

Función	Descripción
`union(vector1, vector2)`	Entrega los elementos en ambos vectores
`intersect(vector1, vector2)`	Entrega los elementos en común en ambos vectores
`setdiff(vector1, vector2)`	Le quita los elementos del vector2 que están en el vector1
`setequal(vector1, vector2)`	Indica si los vectores poseen o no los mismos elementos
`is.element(vector1, vector2)`	Indica si el vector1 está contenido en el vector2
`unique(vector1)`	Entrega los elementos del vector1